Projet CASIMIR

INRA SCIENCE & IMPACT

Outil d'aide à la conception de stratégies d'échantillonnage pour caractériser la composante biotique des agroécosystèmes

Auteurs : *M. Froger, V. Cellier et J-N. Aubertot.*

Le 20/06/2016

M. Froger[1], V. Cellier[1*], J-N. Aubertot[2].**
(1) INRA, UE115 Domaine Expérimental d'Epoisses, F-21100 Bretenière, France.
(2) INRA, Auzeville, UMR AGIR, 24 chemin de Borde-Rouge CS 52627, F-31326 Castanet Tolosan, France.

*vincent.cellier@inra.fr
**jean-noel.aubertot@toulouse.inra.fr

ÉCOPHYTO REDUIRE ET AMÉLIORER L'UTILISATION DES PHYTOS
ÉCOPHYTO DEPHY
ONEMA Office national de l'eau et des milieux aquatiques

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

ITAB Institut Technique de l'Agriculture Biologique
ACTA
RésoEst
Réseau PIC PMNET INRA cirad

Projet CASIMIR
Avant-Propos

L'objectif du projet CASIMIR est de développer des méthodes pour une caractérisation simplifiée de la composante biotique des agroécosystèmes. La mise au point de stratégies d'échantillonnage optimisées en termes de coût et de précision fait partie de ces développements méthodologique. L'outil proposé est adapté à la caractérisation de la composante biotique des agroécosystèmes. C'est avant tout un outil pédagogique permettant aux futurs utilisateurs de se poser les bonnes questions et d'avoir accès à quelques solutions pour concevoir leur stratégie d'échantillonnage selon leurs objectifs, leur contexte et leurs contraintes de travail.

Cet outil est issu du travail effectué par Eloi Navarro dans le cadre de son stage de fin d'Etude à l'INRA de Toulouse encadré par Jean-Noël Aubertot et Michel Goulard (UMR AGIR, INRA de Toulouse). Pour étayer son discours Eloi s'est appuyé sur des données obtenues *in situ* fournies par Sandrine Petit (UMR Agroécologie, INRA de Dijon) et Claire Lavigne (partenaires du projet CASIMIR, Unité PSH (UR1115), INRA d'Avigon). Enfin la version finale de cet outil émane d'un travail collaboratif de relecture avec Morgane Froger (IE du projet CASIMIR) et Vincent Cellier (Coordinateur du projet CASIMIR).

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Le Guide

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Annexes

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Table des Figures

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Liste des Tableaux

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Introduction

Comprendre et caractériser les phénomènes biologiques dans le cadre de pratiques agricoles plus écologiques est un enjeu important pour l'agriculture de demain.
L'une des préoccupations est d'obtenir une vision globale de la composante biotique des parcelles agricoles via le recueil de données exploitables scientifiquement. De plus, la volonté d'optimiser les stratégies d'échantillonnage pour qu'elles soient les moins coûteuses possibles, est une réalité.
Force est de constater que des stratégies d'échantillonnages sont souvent reprises d'une année sur l'autre (ou à partir de publications passées), sans que la précision et le coût ne soient évalués. Ainsi, la profession exprime ses besoins d'informations (basiques surtout) sur l'échantillonnage et la précision des résultats obtenus.

Le document suivant proposé par le projet CASIMIR, se veut synthétique, c'est pourquoi les différents plans d'échantillonnage et modèles sont décrits brièvement, et que seuls les aspects mobilisables pour la mise en place de stratégies d'échantillonnage sont développés. Pour obtenir des informations plus générales ou complémentaires, on pourra se référer, suivant les cas à Taylor [1984], Cressie [1993], Vaillant [1996], Madden and Hughes [1999], Madden et al. [2006] et plus rarement aux autres sources.

Architecture du document

- **Le Guide :** Il permet d'établir une stratégie d'échantillonnage et/ou d'analyser une stratégie d'échantillonnage éprouvée afin, si possible, d'en réduire les coûts d'échantillonnage ou bien, de renforcer l'effort d'échantillonnage pour atteindre les objectifs visés. L'information obtenue via l'utilisation de ce guide sera de type « diagnostic » pour caractériser les pressions biotiques et/ou les populations d'ennemis naturels.

- **Ressources théoriques sur les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes :** Cette partie propose des pistes de réflexion pour élaborer des stratégies d'échantillonnage visant à obtenir des informations de type diagnostics.

- **Exemples :** ce sont des études scientifiques menées par l'INRA ou issues de la bibliographie qui servent de support de démonstration à l'optimisation de l'échantillonnage. Trois exemples sont présentés :
  - Etude du taux de prédation de graines adventices
  - Etude de suivi des piégeages de carpocapse
  - Etude de la sévérité du Phoma du Colza

- **Annexes :** compléments d'information sur les statistiques, des éléments de script pour réaliser les analyses statistiques sous le logiciel R et des exemples de réponses possibles suite à des questions abordant l'échantillonnage dans le cadre du réseau Rés0pest.

Comment utiliser le document

L'utilisateur peut utiliser directement le « Guide » (s'il est aguerri en statistiques), néanmoins il est conseillé d'utiliser en parallèle la partie « Ressources théoriques sur les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes » comme appui au guide. Enfin, l'utilisateur trouvera des exemples et des réflexions sur l'optimisation de l'échantillonnage dans les parties « Exemples » et « Annexe ».

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

1 Objectifs du Guide

Le guide d'aide à l'élaboration de stratégies d'échantillonnage vise à obtenir une information de type « diagnostic », à travers l'évaluation des populations de bioagresseurs et/ou d'ennemis naturels, ainsi que l'évaluation de dégâts et/ou l'évaluation des régulations biologiques. Il permet d'établir une stratégie d'échantillonnage pour la caractérisation d'un stress biotique ou d'une régulation biologique (potentielle), et/ou d'analyser une stratégie d'échantillonnage éprouvée afin d'évaluer s'il est possible d'en réduire les coûts d'échantillonnage ou bien, au contraire, si l'effort d'échantillonnage doit être plus important pour atteindre les objectifs visés.

2 Présentation du Guide

Le Guide est structuré par différentes étapes spécifiques avec, dans un premier temps une réflexion préparatoire puis, des étapes successives pouvant être menées en parallèle, tout en veillant à leur cohérence.

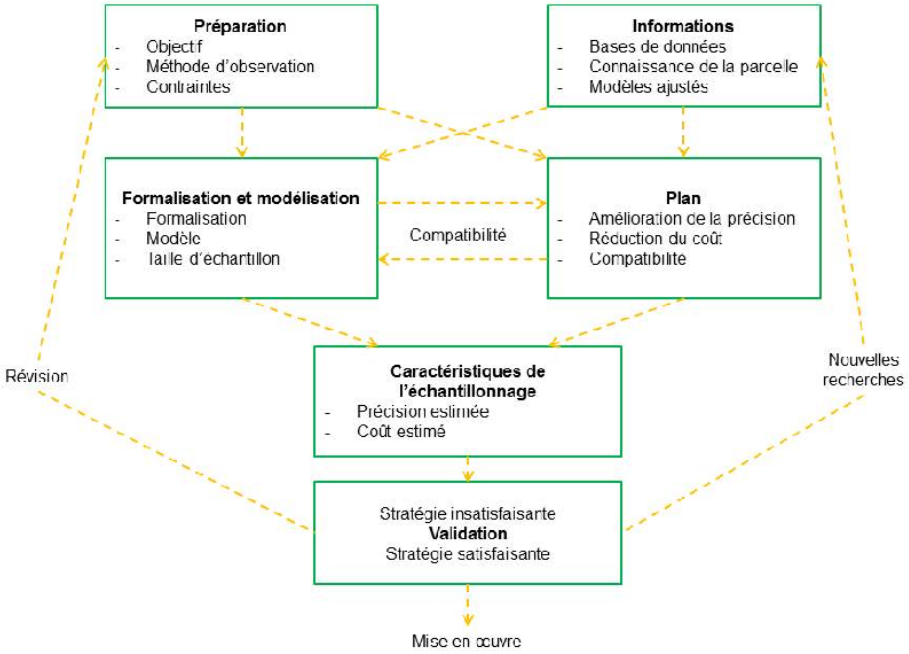

Figure 1: Fonctionnement schématisé de l'outil d'aide à la conception de stratégies d'échantillonnage proposé.

Descriptifs des étapes (Figure 1) :

La préparation : elle précise l'objectif, les méthodes d'observation, et les contraintes (temps, matériels,…) en rassemblant les informations potentiellement utiles pour choisir le type d'échantillonnage.

La recherche d'informations : les informations sur la biologie des bioagresseurs et des ennemis naturels étudiés, sur des données d'échantillonnage existantes, sur des modèles utilisés, seront utiles lors de l'optimisation de la stratégie d'échantillonnage.

La formalisation et la modélisation mathématique : cette étape doit permettre d'aboutir à l'utilisation d'un modèle plus ou moins élaboré permettant d'adapter la taille d'échantillon aux contraintes de précision et de coût.

Le choix du plan d'échantillonnage : en cohérence avec la formalisation mathématique, le plan d'échantillonnage permet d'améliorer la précision et/ou de réduire le coût.

La caractérisation de l'échantillonnage : cette étape émane des étapes précédentes. Les éléments des différentes étapes sont confrontés pour décider si la stratégie obtenue est satisfaisante ou si elle doit être retravaillée. Cela peut impliquer de faire des compromis et de revoir les premières étapes de la réflexion.

La Validation et l'application : la stratégie d'échantillonnage est validée ou non avant d'être appliquée. Si « non », elle est retravaillée.

Chaque étape renferme des questions et différentes tâches à réaliser. A l'issue des différentes étapes du Guide et suite à plusieurs itérations de celles-ci, une stratégie d'échantillonnage répondant au mieux aux objectifs définis peut être mise en place.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

3 La préparation

L'objectif est de poser les bases de la réflexion sur l'échantillonnage que l'on souhaite mettre en place.

3.1 Définition des objectifs de l'observation (8)

Définir les objectifs d'un échantillonnage consiste à déterminer :
Le contexte de l'étude (impacte des éléments extra-parcellaires sur la variable étudiée à l'échelle de la parcelle ou l'évaluation du système de culture) qui va conditionner la répartition des échantillons (ex : inclure ou non les bordures de parcelle).
Les paramètres de la variable qui sera mesurée (moyenne, médiane, …), le niveau de précision ou de certitude souhaité ainsi que le coût acceptable (en termes de temps le plus souvent).

Question : L'objectif de l'observation est-il de déterminer globalement à la parcelle, la présence d'un bioagresseur ou d'un auxiliaire (intensité de la pression biotique ou abondance), et/ou de réaliser un diagnostic?
Non- Cependant, il est possible de s'appuyer sur les pistes de réflexions proposées dans le guide pour élaborer la stratégie d'échantillonnage.
Oui- Il faut maintenant préciser et définir l'objectif de l'échantillonnage.

Remarque : s'il existe plusieurs objectifs, la conception de stratégie d'échantillonnage peut être poursuivie séparément pour chacun d'eux. L'utilisateur peut par la suite s'organiser pour réaliser les échantillonnages simultanément.

3.2 Choix de la méthode d'observation (9)

Choisir la méthode consiste à définir le protocole qui sera mis en place en définissant l'objet qui sera observé (plante, arbre, organe ou quadrat), la méthode de notation (échelle de notation ou comptage, note qualitative ou quantitative). Ces éléments peuvent influencer la précision recherchée, peut rendre des observations incompatibles avec un modèle préexistant élaboré avec un autre protocole, et peut poser des problèmes de comparabilité avec d'autres études.

Question : une méthode d'observation a-t-elle été définie ?
Non- Consulter un expert ou la littérature sur les méthodes (les plus courantes seront mieux documentées).
Oui- Passer à la question suivante.
Plusieurs méthodes envisagées- La conception de la stratégie d'échantillonnage peut être réalisée pour chacune des méthodes. Les conclusions obtenues permettront de les comparer et de repérer la stratégie répondant au mieux aux objectifs et aux contraintes de l'utilisateur.

Question : la méthode d'observation choisie est-elle fiable (biais, répétabilité et reproductibilité) (9.1) ?
Non – Ainsi, l'utilisation de modèles sera hasardeuse car les données d'une session d'échantillonnage à l'autre seront difficilement comparables.
Oui – Ainsi, l'utilisation de modèles sera possible (12).

Remarque : Si la méthode a déjà été utilisée (articles scientifiques, retours d'expérience,...), il est possible d'avoir une idée de sa fiabilité.

3.3 Définition de la limite acceptable du coût en temps de l'observation (9.2)

Le temps disponible influence la taille de l'échantillon (nombre total de relevés) définie dans le protocole. Le plus souvent, les protocoles peu chronophages seront préférés. Cependant, la précision de ces derniers est très souvent réduite ce qui les rend potentiellement moins fiables.

Tâche : Estimer quel temps est nécessaire pour la méthode d'observation choisie. En déduire un nombre maximal d'observations envisageables.

Tâche : Estimer la part du temps de déplacement dans la parcelle sur le temps total nécessaire pour réaliser l'observation.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

4 La recherche d'informations

L'objectif est de recueillir des informations utiles sur la biologie des bioagresseurs et des auxiliaires étudiés, sur des données d'échantillonnage existantes ou sur des modèles ayant déjà été utilisés, afin d'optimiser la stratégie d'échantillonnage.

Tâche : Rassembler des informations sur les conditions environnementales et les pratiques culturales qui pourraient influencer la structuration spatiale de la variable observée (11.2). Trois types de structuration de la parcelle sont à prendre en compte pour établir une stratégie d'échantillonnage.

- La disposition des cultures (11.2.1) ;
- La structuration spatiale de la variable étudiée peut être hétérogène à cause de divers facteurs (11.2.3) ;
- La structuration spatiale de la variable étudiée peut être hétérogène à cause de caractéristiques biologiques de la variable (agrégation ou de corrélation spatiale) (11.2.3.1 ; 11.2.3.2).

Remarque : La structuration spatiale (causée par divers facteurs) des populations de bioagresseurs ou d'ennemis naturels influencent les aspects de la stratégie d'échantillonnage suivant : le plan et/ou les modèles utilisés. Si plusieurs sources d'hétérogénéité sont présentes, il peut être judicieux de ne prendre en compte que la principale.

Question : Existe-t-il une stratégie d'échantillonnage dont la précision des résultats et le coût étaient satisfaisants et répondaient à des objectifs similaires à ceux fixés en (3.1)?

- **Oui-** étudier cette stratégie préexistante pour repérer les améliorations potentielles pouvant être utiles à la stratégie d'échantillonnage en cours d'élaboration.
- **Non-** s'appuyer sur le document ressource et/ou la partie 6 du Guide
  - **le coût était trop élevé -** essayer de réduire le coût sans trop perdre de la précision via la révision du plan d'échantillonnage. S'appuyer sur la partie 11.3 « plan d'échantillonnage ».
  - **la précision était insuffisante -** essayer d'améliorer la précision sans augmenter significativement le coût en temps. S'appuyer sur la partie 10.2 « Indicateurs de précision »

Question : Existe-il des études sur l'agrégation ou la corrélation (11.2.3.2) spatiale des populations des bioagresseurs et auxiliaires ?

- **Oui-** Ces informations permettront de choisir des modèles spécifiques pour analyser les données et/ou d'adapter le plan d'échantillonnage (espacement géographique des observations).
- **Non**- utiliser les connaissances sur la biologie des organismes étudiés afin d'évaluer grossièrement leur agrégation ou leur corrélation spatiale.

Question : Existe-il des modèles permettant d'étudier des phénomènes de stress biotiques ou de régulations biologiques ?

- **Oui-** ils pourront être mobilisés. L'utilisateur pourra s'appuyer sur les paramètres existants des modèles ou sur d'autres paramètres issus de données proches des conditions d'échantillonnage de la parcelle étudiée (type de cultures, stade de développement de la culture,…).
- **Non-** cependant, des bases de données issues d'échantillonnages similaires à celui envisagé, peuvent être recherchées afin de tester ou ajuster un modèle éventuel.

Tâche : Essayer de pré-estimer les résultats qui peuvent être obtenus suite à l'échantillonnage. Cela permet souvent d'optimiser la taille de l'échantillon. Des mesures antérieures réalisées éventuellement sur d'autres parcelles ou l'utilisation des modèles épidémiologiques peuvent être valorisés.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

5 Formalisation mathématique du résultat souhaité et modèles

Cette approche permet de quantifier la précision que l'on espère obtenir, et parfois de la prévoir. Cette étape peut être négligée si on se contente d'optimiser l'échantillonnage sur des critères plus qualitatifs et empiriques.

5.1 Formalisation

La variable d'intérêt et la méthode d'observation associée déterminent le type de données obtenu. Ces dernières vont conditionner l'utilisation des modèles de distribution, des modèles de variance et des quantificateurs de précision existants.

Tâche : Formaliser l'objectif selon la typologie des variables proposée ci-dessous :

5.1.1 Valeurs continues

Définition	Une variable est dite continue lorsqu'elle peut prendre une infinité de valeurs. En pratique, aucune mesure n'a la précision de l'infiniment petit. Alors, on considère une variable comme continue dès lors qu'on lui attribue un grand nombre de valeurs possibles. Le paramètre de la variable étudiée est la moyenne.
Modèles pouvant être utilisés	L'estimateur de la moyenne $\hat{m}$ peut suivre une loi normale ou une loi de Student et permettre de construire des intervalles de confiance pour la moyenne (12.2.1 ; 12.2).
Améliorer la précision	La précision de la moyenne calculée sera caractérisée par un écart-type, un coefficient de variation ou un intervalle de confiance. Si elle est comparée à un (ou plusieurs) seuil(s), des tests basés sur la loi normale ou la loi de Student permettent de calculer une probabilité d'erreur.

5.1.2 Valeurs discrètes entières bornées

Définition	Une variable est dite discrète lorsqu'elle peut prendre un nombre fini de valeurs. On dit que ces valeurs sont bornées lorsque la valeur maximale qu'elles peuvent prendre est connue. Des valeurs entières bornées peuvent par exemple être obtenues lors de comptages de plantes malades dans des grappes de plantes de taille fixée n (par exemple 10 plantes consécutives). Ce sont en général les résultats de comptage qui sont modélisés même si l'information intéressante sera l'incidence exprimée en pourcentage**.** Le paramètre de la variable étudiée est la moyenne.
Modèles pouvant être utilisés	Les modèles utilisés pour ce type de valeurs sont la loi bêta-binomiale ou une loi de puissance sur la variance (12.3.3 ; 12.4).
Améliorer la précision	La sur-dispersion parfois constatée peut être modélisée par la loi bêta-binomiale ou une loi de puissance sur la variance afin de mieux évaluer la variance de $\hat{m}$, voire de la prévoir. Les intervalles de confiance pourront être déterminés de manière très approximative avec la loi normale ou de manière plus fine avec un les modèles évoqués.

5.1.3 Valeurs discrètes entières non-bornées

Définition	Des valeurs entières sont dites non–bornées lorsque la valeur maximale potentielle qu'elles peuvent prendre n'est pas connue. Elles peuvent par exemple être obtenues lors de comptages de symptômes ou de ravageurs sur des organes (sévérité) ou lors de comptages de plantes adventices dans des quadrats. Le paramètre de la variable étudiée est la moyenne.
Modèles pouvant être utilisés	Les modèles utilisés pour ce type de valeurs sont la loi binomiale négative, la loi de Poisson, la loi de puissance de Taylor, et parfois les indices de Lloyd et Iwao (12.3.1 ; 12.4.1 ; 12.4.3).
Améliorer la précision	La sur-dispersion sur les moyennes estimées parfois constatée peut être prise en compte par certains modèles. Les indicateurs de précision et les intervalles de confiance sont calculés comme pour les variables à valeurs entières bornées.

Classes

Définition	Le paramètre de la variable étudiée est une note déterminée parmi un ensemble fini de *k* classes.
Modèles pouvant être utilisés	Même si des phénomènes d'agrégation spatiale existent, aucun modèle utilisé ne

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

	semble les prendre en compte pour ce type de variable. Le seul modèle rencontré dans la littérature scientifique est la loi multinomiale (12.3.5). Ses propriétés permettent de déduire un coefficient de variation ou un intervalle de confiance qui exprime la proportion d'observations dans chacune des classes.
Améliorer la précision	Pour éviter de calculer abusivement une variance ou une moyenne une conversion de la variable ordinale doit être réalisée : ex : attribution d'un score à chaque classe. Cette conversion permet de retrouver une variable du style « valeurs continues », sans pour autant que les modèles applicables aux « valeurs continues » ne soient appropriés aux classes.

Question : L'une des variables proposées convient-elle à la méthode de relevé choisie ?

- **Non –** pour s'assurer de la précision du résultat il faudra se passer de la modélisation qui permet de déterminer la taille optimale de l'échantillon, et ainsi optimiser la répartition des observations via le plan d'échantillonnage (11.3).
- **Oui-** Un des types de données (10.3) et un des indicateurs (10.2) de précision conviennent. Les modélisations possibles peuvent maintenant être étudiées.

5.2 Modèles (12)

Lors de l'élaboration de la stratégie d'échantillonnage, l'utilisation de modèles permet de prévoir la précision obtenue pour une certaine taille d'échantillon. De plus, suite à l'échantillonnage, utiliser un modèle permet de mieux évaluer la précision obtenue. Les modèles proposés dans la partie « Ressources théoriques sur les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes », sont ceux rencontrés dans la littérature scientifique, dans le cadre d'études en agronomie ou en écologie. La liste proposée n'est pas exhaustive.

Tâche : Choisir puis paramétrer un ou des modèles en valorisant les informations rassemblées (biologie et écologie) concernant la variable observée. En effet, ces informations peuvent influencer le paramétrage du modèle. Ex : l'agrégation spatiale (11.2.3.2). S'appuyer sur le document ressource (« Ressources théoriques sur les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes »,) pour le choix du modèle.

5.3 Taille d'échantillon (11.1)

En fonction de l'objectif de l'étude, la taille de l'échantillon résulte d'un compromis entre coût (temps et/ou matériel) et la précision souhaitée. Cette partie permet, en fonction du type de données et des modèles choisis, de définir des formules mathématiques permettant de calculer la taille d'échantillon optimale pour obtenir des résultats précis. Le calcul quant à lui sera réalisé dans la partie 7 du Guide.

Question : Est-ce que les informations disponibles avant l'échantillonnage permettent, en faisant ou non appel à un modèle, de prévoir la précision associée à une taille d'échantillon donnée ?

- **Oui** En déduire la plus petite taille d'échantillon permettant d'obtenir la précision souhaitée.
- **Non** Choisir entre :
  - Échantillonner sans contrôler la précision, avec une taille d'échantillon choisie pour sa faisabilité (11.1.2.1).
  - Mettre en place un échantillonnage adaptatif (11.1.2.2) pour ajuster le nombre d'observations au cours de l'échantillonnage.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

6 Le plan d'échantillonnage (11.3)

L'objectif est de tenir compte (i) des différentes hétérogénéités spatiales supposées exister de la variable étudiée et (ii) du coût d'échantillonnage, afin de définir la répartition spatiale des observations à réaliser sur la parcelle. Les plans d'échantillonnage se distinguent les uns des autres par leur caractère aléatoire ou non et par leur complexité.
Le choix du plan d'échantillonnage est adapté au compromis coût/précision pour éviter d'avoir un biais sur le résultat final. Le plus souvent, il est préférable de réaliser, si possible, des observations réparties aléatoirement dans la parcelle. La répartition des observations influençant fortement la précision, un plan d'échantillonnage choisi judicieusement permettra parfois de réduire la taille de l'échantillon en conservant une précision acceptable.

6.1 Si l'on souhaite améliorer la précision

Question : Des informations sont-elles disponibles concernant l'hétérogénéité ou l'homogénéité spatiale de la variable observée (11.2) ?

- **Non** Privilégier un échantillonnage systématique (11.3.3.2), sauf si on souhaite mettre en place un échantillonnage séquentiel.
- **Oui**
  - S'il existe une hétérogénéité spatiale : s'intéresser à l'échantillonnage stratifié (11.3.4.1)
  - S'il existe une homogénéité spatiale : tous les types de plan d'échantillonnage sont envisageables. Le choix du plan dépendra des objectifs de précision et des limites fixées en termes de coût.
  - En cas de corrélation spatiale (11.2.3.2) notable, espacer les observations par un échantillonnage systématique (11.3.3.2)

6.2 Si l'on souhaite réduire le coût

Question : Est-ce que le coût des déplacements dans la parcelle est élevé par rapport au coût des observations ?

- **Oui** S'intéresser aux échantillonnages par grappes (11.2.4.2), selon un parcours… et à plusieurs degrés (11.3.4.3).
- **Non** En profiter pour bien espacer les observations. Ex : échantillonnage systématique (11.3.3.2).

Remarque : Si chaque observation demande une analyse coûteuse (temps ou matériel), envisager d'analyser des échantillons composites (11.3.5) et éviter l'échantillonnage par grappes. Si le coût d'échantillonnage est très variable au sein de la parcelle, évaluer la possibilité de stratifier (11.2.4.1) la parcelle selon le coût.

6.3 Compatibilités entre formalisation, modèles et le plan d'échantillonnage

Question : Le modèle et le plan d'échantillonnage envisagés sont-ils compatibles ?

- **Non** Adapter le plan dans la mesure du possible. Sinon, les informations fournies par le modèle donneront au mieux des ordres de grandeur.
- **Oui** Une taille d'échantillon peut maintenant être déterminée.

Question : Est-ce qu'un modèle imposant un échantillonnage par grappe a été choisi ?

- **Non** Le modèle s'appliquera indépendamment du plan choisi.
- **Oui** L'échantillonnage par grappes simple peut être choisi, les grappes peuvent aussi être disposées de manière stratifiée ou systématique dans la parcelle.

Question : Est-ce qu'un échantillonnage adaptatif (11.1.2.2) est envisagé ?

- **Non** le choix repose sur une taille d'échantillon déjà définie (biblio, modélisation ou par rapport au coût) de ce fait, l'échantillonnage aléatoire simple (11.3.3.1) n'apportera pas de bénéfice justifiant son utilisation par rapport à son coût.
- **Oui** l'échantillonnage systématique selon une grille est à éviter (11.3.3.2), excepté pour les échantillonnages par phases (11.3.4)

Question : Est-ce que la taille de l'échantillon (ou la démarche adaptative) permet de mettre en place le plan d'échantillonnage envisagé de manière cohérente (pas de strates avec peu d'observations, pas d'échantillonnage systématique avec des observations très serrées,...) ?

- **Non** Ajuster la taille d'échantillon ou adapter le plan, selon ce qui semble le plus important.
- **Oui** Il faut maintenant valider ou invalider la stratégie définie par cette taille d'échantillon et le plan.

Remarque : Si la taille minimale imposée par la contrainte de précision, et la taille maximale imposée par la contrainte de coût se révèlent incompatibles, il est nécessaire de revoir les objectifs ou de changer de méthode d'observation.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

7 Estimation de la précision et du coût puis Validation

Après avoir élaboré la stratégie d'échantillonnage (formalisation, modélisation et définition du plan d'échantillonnage), l'objectif de cette partie est d'estimer la précision obtenue et le coût en temps.

Tâche : En gardant à l'esprit l'effet éventuel du plan d'échantillonnage choisi et en se basant sur les formules fournies pour chaque modèle, si nécessaire, calculer :

- La précision compte tenu de la taille de l'échantillon déterminée en fonction du coût maximal acceptable
- Le coût compte tenu de la taille de l'échantillon nécessaire pour atteindre la précision minimale acceptable (3.1), du temps estimé (3.3) et du temps estimé pour le déplacement dans la parcelle
- Les diagrammes lorsqu'il s'agit d'une approche séquentielle (11.1.2.2)

Question : est-ce qu'il existe un compromis acceptable entre taille d'échantillons, précision et coût ?

- **Oui** La stratégie peut être appliquée au champ.
- **Non**
  - **Le coût est trop élevé-** voir les optimisations possibles. Revoir les exigences en termes de précision ainsi que le choix de la méthode d'observation.
  - **La précision est trop faible-** voir les optimisations possibles. Revoir le seuil de coût à ne pas dépasser (définit par l'observateur) ainsi que le choix de la méthode d'observation.

  ➔ Revoir à la baisse les objectifs (trop ambitieux) pour redéfinir la taille de l'échantillon et/ou réaliser une approche adaptative.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Ressources théoriques sur les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes

Le document qui suit propose des pistes de réflexion pour élaborer des stratégies d'échantillonnage visant à obtenir des informations de type diagnostic. Ainsi, les méthodes d'observations, les types de variables et les stratégies abordés sont limités.

8 Influence de l'objectif de l'échantillonnage

La stratégie d'échantillonnage dépend des informations que l'on souhaite tirer de l'observation et des contraintes pratiques qui s'imposent. Définir clairement les objectifs d'un échantillonnage évite de récolter des données qui ne pourront pas être valorisées par la suite. Après avoir identifié la variable cible, il faut déterminer le(s) paramètre(s) de cette variable que l'on souhaite obtenir (une moyenne, une variance, une médiane, un variogramme, une tendance, le dépassement d'un seuil [de Gruijter et al., 2006, page 29]) ainsi que son niveau de précision ou de certitude souhaité. Enfin, il faut fixer la limite acceptable du coût matériel et en particulier du coût en temps.

8.1 Diagnostic

Pour réaliser un diagnostic (ex : caractériser une infestation) ou obtenir une information qualitative (*Est-ce que les stress biotiques sont maîtrisés sur la parcelle ?*), on peut privilégier un plan d'échantillonnage qui soit représentatif *a minima* de la parcelle (ex : échantillonnage simple [Moura et al., 2007] ou un parcours le long d'une diagonale) ou qui se concentre sur des zones à risque (près des bords si la densité de ravageurs y est plus forte [Brown et al., 1993]).
De manière générale, un diagnostic consiste à déterminer l'intensité moyenne d'un stress biotique ou d'un phénomène de régulation sur la parcelle. Des plans d'échantillonnage variés peuvent participer à l'obtention d'une moyenne précise.

Remarque : Il est important de définir si on souhaite obtenir une information sur l'état de la parcelle dans son contexte paysager ou étudier l'effet du système de culture indépendamment du contexte paysager. Dans un cas les bords de la parcelle sont à surveiller spécialement, dans l'autre ils sont plutôt à éviter.

8.2 Études des variabilités et corrélations

Afin d'obtenir des informations sur les variabilités et les corrélations spatiales et/ou temporelles, il faut pouvoir disposer d'une quantité d'échantillons suffisante pour garantir la résolution spatiale ou temporelle. Deux propositions de stratégie d'échantillonnage peuvent être envisagées : réaliser un plan d'échantillonnage selon un quadrillage [Pulakkatu-Thodi, 2014] (Figure 2) ou bien selon des transects [Parker et al., 1997, Clark et al., 2007].

8.3 Cartographie

En agriculture de précision, pour réaliser une cartographie ou un suivi spatio-temporel (ex : suivi de zones traitées), les relevés sont disposés sur un quadrillage (exemples : Aubertot et al. [2004], Barroso et al. [2005]) dont l'échelle doit être compatible avec celle du phénomène étudié ou du traitement appliqué. On parle d'échantillonnage systématique (11.3.3.2)

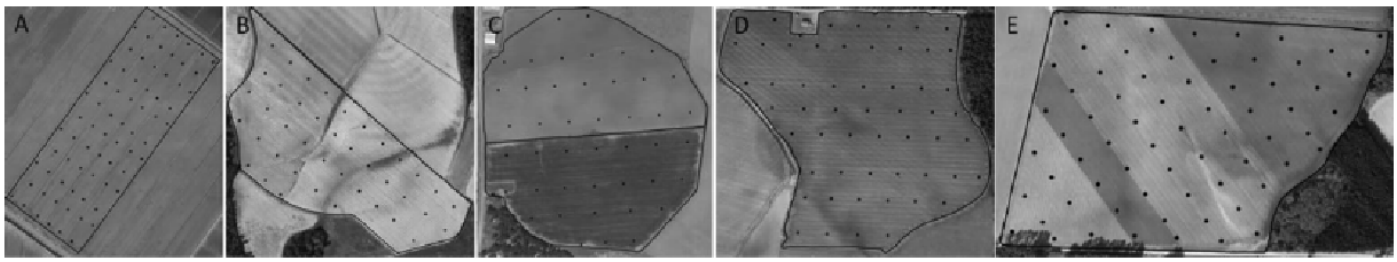

Figure 2: Exemple de plans d'échantillonnage pour une étude de variabilité spatiale, l'échantillonnage est systématique et adapté à la forme des parcelles étudiées - Source : Pulakkatu- Thodi [2014]

9 Méthodes d'observation

La conception d'un plan d'échantillonnage ne doit pas être dissociée de la méthode utilisée pour observer et noter chaque plante, chaque arbre ou chaque quadrat. En effet, cette dernière influence la précision et peut rendre les observations réalisées incompatibles avec un modèle ayant été élaboré à l'aide de données issues de l'utilisation

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

d'autres méthodes. Exemple : dans le cas d'une évaluation de la sévérité d'une attaque, l'échelle choisie influence la précision et peut poser des problèmes de comparabilité avec d'autres études utilisant d'autres échelles de notation.

9.1 Caractériser la qualité des méthodes d'observations

Plusieurs critères décrivent la qualité d'une méthode d'observation : les biais, la sensibilité, la résolution, la répétabilité et la reproductibilité de la méthode. La répétabilité et la reproductibilité sont caractéristiques d'une méthode d'observation fiable permettant d'obtenir des résultats précis. Ce sont des éléments indispensables pour modéliser sans trop de difficultés les données.

9.1.1 Le biais

Le biais est l'écart moyen entre la valeur issue d'une méthode d'observation de référence (réputée fiable) et la valeur de la variable aléatoire estimée. Un biais peut nuire fortement à la qualité des résultats s'il ne peut pas faire l'objet d'une correction (variogramme pour correction annexe…).
Exemple : les changements climatiques peuvent être des sources de biais c'est-à-dire qu'ils contribuent à l'écart moyen des résultats par rapport à une référence.

9.1.2 La sensibilité

La sensibilité est le seuil en dessous duquel l'observateur ne sera pas capable de détecter le phénomène observé avec la méthode d'observation choisie. La sensibilité requise dépend plus des objectifs de l'échantillonnage que de sa mise en œuvre.

9.1.3 La résolution

La résolution est l'écart limite en dessous duquel la méthode d'observation ne permet pas de différencier deux situations.

9.1.4 La répétabilité

La répétabilité est la possibilité pour un même observateur d'obtenir le même résultat s'il applique plusieurs fois la méthode d'observation. La répétabilité est meilleure lorsque les méthodes sont simples ou automatisées.

9.1.5 La reproductibilité

La reproductibilité est la possibilité pour plusieurs observateurs utilisant la même méthode d'observation, d'obtenir les mêmes résultats. Elle est d'autant plus mauvaise que la méthode est subjective et basée sur le jugement de l'observateur.

9.2 Influence du temps disponible

Le temps disponible influence la taille de l'échantillon et le choix de la méthode d'observation. En effet, les méthodes d'observation peu exigeantes en termes de temps sont très souvent recherchées au détriment de la précision et de la fiabilité des résultats obtenus.

Utiliser des méthodes d'observation dont on connait la fiabilité et le temps de mise en œuvre, limite les pertes de précision. En cas d'hésitation entre deux méthodes d'observation, des stratégies d'échantillonnages adaptées à chacune peuvent être conçues, puis comparées en fonction du coût et de la précision prévus.

9.3 Positionnement dans le temps

La période de relevée peut influencer le contexte de mis en place de la méthode d'observation (ex : observation plus ou moins difficile selon le stade de la culture) et donc influencer la répétabilité et la reproductibilité de celle-ci.
La meilleure option pour le diagnostic est d'échantillonner au moment où les variables observées (bioagresseurs, dégâts/symptômes, auxiliaires) sont les plus visibles. Afin de définir cette période d'observation, il est important de connaitre la dynamique de ces variables, en réalisant une surveillance sommaire de leur présence ou en utilisant toute autre source d'informations (bibliographie, retour d'experts,…) pouvant être mises à profit.

10 Types de données et précision associée

La variable d'intérêt (bioagresseurs ou auxiliaires) et la méthode d'observation associée déterminent le type de données obtenues. Pour chaque type de données, divers modèles de distribution, modèles de variance (12.4) et quantificateurs de précision (voir 10.2) existent. De fait, la traduction formelle de l'objectif de l'échantillonnage en données (qui seront manipulées par la suite) est déterminante.

10.1 Paramètre d'intérêt

On considère qu'une stratégie d'échantillonnage se compose de N observations, ou grappes d'observations, qui aboutissent à N valeurs $X_1, \ldots, X_N$. L'utilisation éventuelle de modèles ainsi que le choix d'un indicateur de précision dépendent du paramètre d'intérêt qui sera calculé à partir de ces valeurs. Ainsi, l'objectif de l'échantillonnage revient à

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

déterminer un paramètre à partir des données recueillies. Ex : *Quelle est l'incidence moyenne de tel bioagresseur sur la parcelle ?* Dans le cadre d'un diagnostic, les données sont essentiellement traitées pour déterminer une **moyenne**, une **répartition entre classes**, ou évaluer le **dépassement d'un seuil**. Les modèles et les indicateurs de précision proposés dans la suite de ce document sont restreints à ces trois cas.

Si on souhaite estimer la moyenne m de la variable mesurée sur la parcelle, *l'estimateur* (10.1) est :

$$\widehat{m} = \frac{1}{N}\sum_{i=1}^{N} Xi$$

Si on souhaite estimer la proportion pi d'individus dans la i-ème classe d'un ensemble de k classes, l'estimateur est $\widehat{p_i} = \frac{Ni}{N}$ où Ni est le nombre d'observations dans la classe i

10.2 Indicateurs de précision

La précision des résultats est caractérisée par une faible variabilité entre les répétitions d'un échantillonnage. Ainsi pour s'assurer de la précision des résultats il faut pouvoir quantifier les variations possibles, d'une répétition à l'autre, par rapport au résultat attendu. Pour cela, il existe divers indicateurs de précision. Le choix d'un indicateur dépend à la fois du type de variable observée et des objectifs de l'échantillonnage.

10.2.1 La variance

La variance n'indique pas très clairement la précision obtenue, toutefois elle peut être conservée pour affiner une analyse ultérieure ou préparer un futur plan d'échantillonnage dans des conditions similaires. L'estimation précise de la variance demande une grande taille d'échantillon, mais en réalité on se satisfait le plus souvent d'une estimation médiocre de la variance (16).

Lorsque des valeurs observées numériques sont sommées, la variance sur l'échantillon observé est donnée par la formule :

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (Xi\_\widehat{m})^2$$

Sous certaines hypothèses (indépendance des observations), on peut en déduire la variance de la moyenne estimée $\widehat{m}$ par la formule :

$$s^2_{\widehat{m}} = \frac{s^2}{N}$$

Précision : Si on multiplie la moyenne par un coefficient k alors la variance de km est donnée par la formule $s^2_{km} = k^2 s^2_m$

10.2.2 L'écart-type

L'écart-type, pour une moyenne estimée est la racine carrée de sa variance qui vaut $\frac{s}{\sqrt{N}}$. L'unité de la valeur de l'écart-type est la même que celle de la variable étudiée. Afin de donner un ordre de grandeur de l'incertitude associée à la moyenne calculée, l'écart-type est utilisé pour calculer des intervalles de confiance.

Précision : Si on multiplie la moyenne par un coefficient *k* alors l'écart-type de *km* est donnée par la formule $s_{km} = k s_m$.
k est un coefficient quelconque (par exemple pour convertir un nombre de plantes par quadrat en un nombre de plantes par m²)

10.2.3 Les intervalles de confiance (I_c)

Un intervalle de confiance à 95% (le niveau de confiance choisi le plus souvent) est borné par deux valeurs calculées de telle façon que si on reproduisait l'échantillonnage un grand nombre de fois, la valeur à estimer se trouverait entre ces valeurs dans 95% des cas.
Les intervalles de confiances peuvent être établis différemment en fonction des modèles utilisés. La contrainte de précision porte en général sur la demi-largeur de l'intervalle de confiance (notée d dans la suite du document).

10.2.4 Le coefficient de variation (c_v)

Le coefficient de variation permet de comparer la variabilité des échantillons ayant des moyennes très différentes ou qui ne sont pas exprimées dans les mêmes unités. En effet, c_v est adimensionné car il exprime l'écart-type en pourcentage de la moyenne, ce qui permet de comparer des mesures dans différentes grandeurs physiques [http://ebrunelle.profweb.ca/MQ/Chapitre8.pdf]. C'est un indicateur qui explicite bien la dispersion des données par

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

rapport à la moyenne. En revanche, pour une moyenne proche de zéro, il prend alors des valeurs très grandes ce qui porte à confusion.

Enfin, le coefficient de variation est, pour une moyenne estimée, son écart-type relatif, soit en pratique la valeur de l'écart-type divisé par l'estimation de la moyenne : $\frac{s}{\hat{m}\sqrt{N}}$

10.2.5 La probabilité d'erreur de classification

Lorsque l'on veut comparer la grandeur estimée à un seuil, il est intéressant de savoir avec quelle certitude la valeur réelle est du même côté du seuil. Cette certitude peut parfois être estimée à partir de la théorie des tests statistiques. La courbe sur la figure ci-dessous (Figure 3) représente par exemple un taux d'erreur de classification en fonction du rapport entre la valeur estimée (une densité de parasites) et le seuil. La probabilité d'erreur de classification acceptable est à fixer en fonction des objectifs de l'échantillonnage.

Remarque : dans le cas de l'étude d'un dépassement de seuil la résolution est un élément important pour s'assurer de la justesse du résultat.

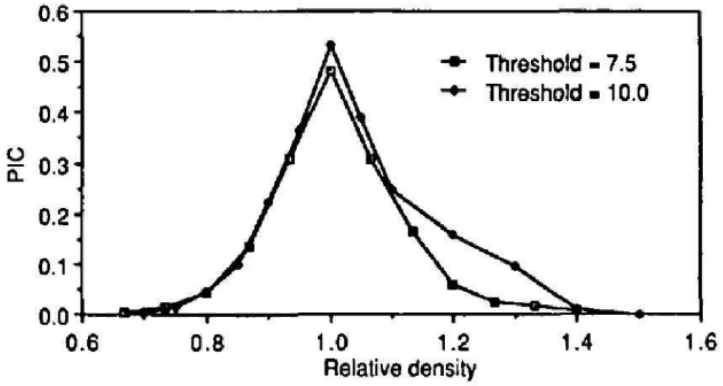

Figure 3: Probabilité d'erreur de classification (PIC pour Probability of Incorrect Classification)

10.3 Types de données retenus

Ce document présente, ci-dessous, quatre types de données couramment utilisées dans le cadre des méthodes d'observation.

10.3.1 Valeurs continues

Une variable est dite continue lorsqu'elle peut prendre une infinité de valeurs. En pratique, aucune mesure n'a la précision de l'infiniment petit. Alors, on considère une variable comme continue dès lors qu'on lui attribue un grand nombre de valeurs possibles.
Par exemple, des valeurs continues peuvent être obtenues par des mesures de taille ou de biomasse. Elles peuvent aussi provenir de mesures exprimées en pourcentage (incidence, sévérité), même si dans ce cas on préfère en général modéliser les résultats de comptage et non le pourcentage obtenu [Madden and Hughes, 1999].
Pour ce type de données, le **paramètre d'intérêts** qui sera utilisé est **la moyenne**.

10.3.1.1 Modèles

Si l'estimateur de la moyenne $\hat{m}$ suit une loi normale ou une loi de Student (12.2.2), on peut construire des intervalles de confiance pour la moyenne via le calcul d'écart-type. Lorsqu'il existe une corrélation spatiale (11.2.3.2) entre différents relevés, l'estimation de l'écart-type peut être faussée, cependant une correction peut être envisagée (15.4).

10.3.1.2 Précision selon la question

La précision de la moyenne calculée pourra être caractérisée par un écart-type, un coefficient de variation ou un intervalle de confiance. Si elle est comparée à un (ou plusieurs) seuil(s), il existe des tests basés sur la loi normale ou la loi de Student qui permettent de calculer une probabilité d'erreur.

10.3.1.3 Remarque

Les outils mathématiques évoqués (sauf ceux impliquant la loi de Student) peuvent aussi être appliquées à des variables discrètes entières. Toutefois il faudra plus d'observations pour que $\hat{m}$ suive la loi normale (12.2.1).

10.3.2 Valeurs discrètes entières bornées

Une variable est dite discrète lorsqu'elle peut prendre un nombre fini de valeurs. On dit que ces valeurs sont bornées lorsque la valeur maximale qu'elles peuvent prendre est connue. Des valeurs entières bornées peuvent par exemple

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

être obtenues lors de comptages de plantes malades dans des grappes de plantes de taille fixée n (par exemple 10 plantes consécutives). Ce sont en général les résultats de comptage qui sont modélisés même si l'information intéressante sera l'incidence exprimée en pourcentage.

10.3.2.1 Modèles

Lors d'un échantillonnage en grappes, si les observations à l'intérieur de chaque grappe sont indépendantes les unes des autres, la loi binomiale (12.3.2) permet de modéliser les valeurs observées. Dans les autres cas, une sur-dispersion est en général observée. Cette dernière est définie comme une variance des données observées supérieure à la variance théorique issue du modèle utilisé. De manière similaire, la sous-dispersion peut se définir comme une variance observée inférieure à celle induite par le modèle. Toutefois, cette dernière situation est très rarement rencontrée dans les problématiques épidémiologiques [Sébastien MARQUE. Prise en compte de la sur-dispersion par des modèles _a mélange de Poisson. Life Sciences. Université Victor Segalen - Bordeaux II, 2003]. La sur-dispersion peut être modélisée par la loi bêta-binomiale ou une loi de puissance sur la variance [Madden and Hughes, 1999] (12.3.3 ; 12.4) ce qui permet de mieux évaluer la variance de $\hat{m}$, voire de la prévoir.

10.3.2.2 Précision selon la question

Les indicateurs de précision sont les mêmes que pour une variable continue. Les intervalles de confiance pourront être déterminés de manière très approximativement avec la loi normale ou de manière plus fine avec un des modèles évoqués.

Remarque : Dans le cas d'un échantillonnage en grappe, si on veut obtenir l'écart-type de l'incidence moyenne en pourcentage, alors on prend l'écart-type du nombre de plantes malades pour les mesures d'incidence, l'écart-type inter-grappe est égal à l'écart-type obtenu sur le compte moyen divisé par n. Le coefficient de variation est le même, et les intervalles de confiance se déduisent directement en divisant les deux bornes pas n.

10.3.3 Valeurs discrètes entières non-bornées

Des valeurs entières sont dites non–bornées lorsque la valeur maximale potentielle qu'elles peuvent prendre n'est pas connue. Elles peuvent par exemple être obtenues lors de comptages de symptômes ou de ravageurs sur des organes (sévérité) ou lors de comptages de plantes adventices dans des quadrats. En théorie, ces valeurs ne sont pas bornées mais en pratique elles sont bornées par des contraintes physiques diverses. L'agrégation spatiale (11.2.3.2) observable sur ce type de variable a été largement étudiée en écologie, elle se traduit par une sur-dispersion sur les moyennes estimées. Cette dernière est parfois prise en compte par les modèles excepté celui de la loi de Poisson.

10.3.3.1 Modèles

En écologie et en protection des cultures, les modèles utilisés pour ce type de variable sont la loi binomiale négative, la loi de Poisson, la loi de puissance de Taylor, et parfois les indices de Lloyd et Iwao (12.4.3).

10.3.3.2 Précision selon la question

Les indicateurs de précision et les intervalles de confiance sont calculés comme pour les variables à valeurs entières bornées.

10.3.4 Classes

Le résultat de l'observation est une note déterminée parmi un ensemble fini de *k* classes. On retrouve ce type de notation par exemple pour évaluer la sévérité d'une attaque de bioagresseur avec l'utilisation d'une échelle ordonnée (cas de Aubertot et al. [2004]). Parfois, par confusion entre variables ordinales et numériques, une variance ou une moyenne sont calculées abusivement avec des variables ordinales considérées comme des variables numériques [Madden et al., 2006, page 20]. Pour éviter ce type d'erreur, les calculs doivent être effectués après une conversion, exemple : on attribue à chacune des classes un score qui peut être la valeur moyenne attendue.

10.3.4.1 Modèles

Même si des phénomènes d'agrégation (11.2.3.2.1) existent, aucun modèle utilisé dans le domaine de la protection des cultures ou en écologie ne semble les prendre en compte pour ce type de variable. Le seul modèle rencontré est la loi multinomiale (12.3.5), qui ne tient pas compte de l'ordre éventuel des classes. Le développement de modèles plus élaborés est limité par la faible répétabilité des estimations de sévérité en général.

Remarque : attention, attribuer un score à une classe permet de retrouver une variable du style « valeurs continues », sans pour autant que les modèles applicables aux « valeurs continues » (Cf : paragraphe précédent) ne soient appropriés aux classes.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

10.3.4.2 Précision selon la question

Lorsque l'on souhaite déterminer la proportion de la population totale qui se trouve dans chaque classe, en général, on caractérisera la précision pour chaque proportion par l'intervalle de confiance. Si l'on souhaite déterminer les scores moyens sur la population (dans le cas où chaque classe est associée à un score), alors on caractérisera la précision par le coefficient de variation ou l'écart-type. Ils peuvent être calculés grâce à la loi multinomiale (voir partie modèles), même si la formule est exacte uniquement lorsque les relevés sont indépendants.

Remarque : Une notation en classe sur toute la parcelle (ex : protocoles Vigiculture et CASIMIR) n'est pas vraiment un échantillonnage et la fiabilité du résultat dépend surtout de la compétence de l'observateur.

11 L'échantillonnage

11.1 Taille de l'échantillon

La taille de l'échantillon est déterminée en fonction de l'objectif qui a été choisi en termes de précision et selon les contraintes de coût (temps, matériel*)*. Des informations disponibles avant l'échantillonnage peuvent être mises à profit pour définir un bon compromis entre le coût et la précision. Ainsi, si la taille minimale imposée par la contrainte de précision, et la taille maximale imposée par la contrainte de coût se révèlent incompatibles, il est nécessaire de revoir les objectifs ou de changer de méthode d'observation.

11.1.1 Détermination de la taille de l'échantillon par le coût d'échantillonnage

Le choix du nombre de relevés est limité par des contraintes de disponibilités en temps et en matériel. En tenant compte de ces contraintes, après avoir défini le nombre maximal de relevés qui sera réalisé, il est possible de prévoir la précision obtenue afin de juger de la qualité de la stratégie d'échantillonnage.

Lors de la détermination de la répartition optimale des relevés (11.3), le temps de déplacement dans la parcelle (fonction du type de la culture et du matériel à transporter) doit être pris en compte. Les modèles permettant de faire le lien entre un plan d'échantillonnage, une taille d'échantillon et le coût associé, sont peu génériques. En effet, pour chaque observation, le temps passé dépend de la compétence de l'observateur, des conditions météorologiques, de l'état de la culture, etc. Ainsi, c'est au concepteur de la stratégie d'échantillonnage de déterminer quels sont les déplacements et quel est le nombre d'observations acceptables.

11.1.2 Détermination de la taille de l'échantillon par la contrainte de précision

La taille de l'échantillon influence fortement la précision des résultats obtenus. Ainsi, connaitre *a priori* les caractéristiques de la variable à observer, permet de déduire une taille optimale de l'échantillonnage avant sa mise en place sur le terrain. Ou bien, un échantillonnage adaptatif permet d'adapter la taille finale de l'échantillon en fonction des informations recueillies pendant l'échantillonnage.

Dans la plupart des cas, la variance permet de prévoir la précision des résultats pour la variable étudiée. Cette variance peut être (i) connue à l'avance grâce à l'expérience acquise lors d'échantillonnage précédents, (ii) déduite d'un modèle et d'hypothèses sur la moyenne qui sera observée et/ou (iii) déterminée au cours d'un échantillonnage via l'utilisation d'une méthode adaptative.

11.1.2.1 Taille d'échantillon fixée à l'avance

Les différents indicateurs de précision (10.2) peuvent variés en fonction de la taille de l'échantillon ou du paramétrage du modèle utilisé. Par conséquent, si on dispose d'une prévision de variance ou des paramètres adéquats pour un modèle fiable, on peut déduire un nombre minimal d'observations à réaliser pour que la précision soit satisfaisante (11.1).

11.1.2.2 Échantillonnage adaptatif

L'échantillonnage adaptatif s'impose lorsqu' il est nécessaire de contrôler la précision de l'échantillonnage et que celle-ci ne peut pas être estimée avant le début des observations. Cela implique la réalisation de quelques calculs au cours de l'échantillonnage. Cette méthode n'est pas applicable dans le cadre d'observation durant dans le temps tel que le piégeage (temps entre la pose du piège et le relevé du piège).
Ci-dessous, sont présentées deux méthodes d'échantillonnage adaptatif : la méthode par phases et la méthode séquentielle.

11.1.2.2.1 Échantillonnage par phases

Un échantillonnage par phases consiste à réaliser un premier échantillonnage sommaire, puis un deuxième plus poussé avec un nombre et une disposition des relevés adaptés selon les informations obtenues lors de la première phase. Les résultats obtenus lors de la première phase d'observation sont complétés par ceux de la deuxième phase.
Ainsi, la première phase peut :

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

- permettre d'estimer la variance, même si cette estimation est très grossière de par le faible nombre d'observations.
- servir à déterminer une moyenne approximative, dont on déduit la précision potentielle que l'on peut atteindre à l'aide d'un modèle.
- révéler des hétérogénéités de répartition spatiale de la variable étudiée au sein de la parcelle permettant ainsi d'adapter le plan d'échantillonnage.

11.1.2.2.2 Échantillonnage séquentiel

En général, les procédures d'échantillonnage séquentiel sont proposées pour des estimations d'incidence ou des comptages. Le principe est d'échantillonner jusqu'à ce que la précision obtenue soit satisfaisante. De telles procédures permettent, de manière rigoureuse, d'interrompre l'échantillonnage dès que l'information obtenue convient. Elles se basent sur un modèle qui permet d'estimer la variance d'une variable au sein d'une parcelle à partir de la moyenne estimée sur un certain nombre d'observations.

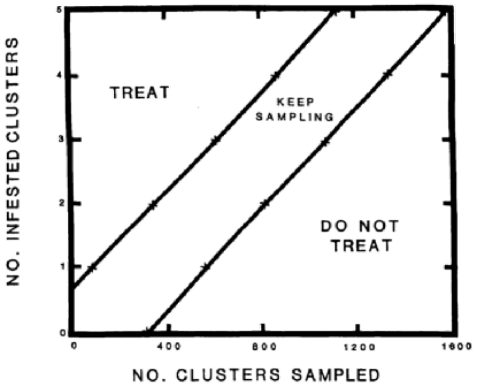

Figure 4 : Diagramme pour simplifier l'échantillonnage séquentiel - Source : Ring et al. [1989]

Concrètement, à partir du modèle choisi, on déduit une relation entre l'indicateur de précision choisi, le nombre total d'observations réalisées N, et le nombre d'observations dont le résultat est positif (ou le nombre total d'individus comptés) T_N (exemple : présence de la maladie, nombre d'insecte sur les N premiers fruits examinés). Ainsi, pour un couple de valeurs N,T_N, l'indicateur de précision peut alors être calculé ; s'il est satisfaisant l'échantillonnage est arrêté, et si la précision n'est pas suffisante l'échantillonnage est poursuivi. Sur le diagramme (Figure 4) sont représentés N en abscisses et T_N en ordonnées ainsi que les zones ou la précision est satisfaisante.

Exemple : Ring et al. [1989] propose le diagramme figure 4, avec pour objectif de savoir si la densité de ravageurs dépasse un seuil. De même, deux exemples de diagrammes sur l'étude des carpocapses se trouvent en exemple à la fin du document (0). Enfin, le code R pour tracer de tels diagrammes est disponible en (16).

11.2 Influence de la structure spatiale sur l'échantillonnage

Pour établir une stratégie d'échantillonnage dans une parcelle, trois types de structuration spatiale doivent être considérées :

- Premièrement, la disposition des cultures qui influence le type de plan d'échantillonnage ;
- Deuxièmement, les valeurs de la variable étudiée qui peuvent être hétérogènes à cause de divers facteurs environnementaux (climat, pratiques culturales,…) ;
- Troisièmement, les phénomènes d'agrégation ou de corrélation spatiale pouvant entrainer une hétérogénéité des valeurs.

Les caractéristiques biologiques de la variable étudiée (bioagresseur ou auxiliaire) structurent sa répartition spatiale, ce qui influence le choix du plan d'échantillonnage et des modèles utilisés lors de la conception de la stratégie d'échantillonnage. Si plusieurs sources d'hétérogénéité spatiale sont présentes, il peut être judicieux de ne tenir compte que de la principale.

11.2.1 Influence du type de culture sur l'organisation de la parcelle

Le type de culture structure spatialement les parcelles étudiées (rang, arbre, ...) et influence la stratégie d'échantillonnage. En effet, l'organisation spatiale des cultures impose le choix des unités statistiques à observer. Les champs et les vergers sont les structures les plus fréquemment rencontrées, mais il existe aussi les serres, les vignobles,...

Remarque : lors du choix et de la mise en place du plan d'échantillonnage, il faut tenir compte des motifs spatiaux qui structurent la répartition spatiale des unités échantillonnées pour ne pas biaiser les résultats (notamment dans le cas de cultures associées en bandes).

Dans les cultures de pleins champs, l'observation des pressions biotiques et le prélèvement de plantes se fait :

- sur quadrats, c'est-à-dire sur une délimitation (virtuelle ou non) de forme prédéterminée. Les quadrats peuvent être repérés le long de rangs, placés totalement aléatoirement ou encore positionnés sur un quadrillage.
- sur un rang ou par grappe
- sur un transect selon un parcours bien choisi avec des observations faites aléatoirement ou régulièrement le

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

long du parcours (parcours aléatoire, en X [Burts and Brunner, 1981], en zigzag [Workneh et al., 1999], en losange).
- sur la globalité du champ, le long d'un parcours, en réalisant par exemple une évaluation continue [Barroso et al., 2005].

Dans un verger, les rangs, les arbres, les branches, les bourgeons ou les fruits, structurent naturellement la répartition des observations. En effet, un *échantillonnage par degré* est réalisé en choisissant dans un premier temps les arbres (niveau 1), puis les branches dans les arbres (niveau 2), etc (par exemple Brown et al. [1993] et Navarro-Campos et al. [2012], Ojiambo and Scherm [2006] pour les myrtilles). A chaque niveau d'observation, les unités observées sont faciles à distinguer et à numéroter, ce qui peut favoriser un tirage aléatoire. Pour chaque niveau d'observation, le nombre d'unités observées (qui dépend des différentes variances observées) peut être optimisé, mais ce sujet ne sera pas développé dans ce document (voir Ojiambo and Scherm [2006])

11.2.2 Parcelles vraisemblablement homogènes

Dans le cas d'une structuration spatiale homogène de la variable étudiée, les critères déterminants pour choisir le plan d'échantillonnage seront surtout les contraintes pratiques et la possibilité d'utiliser un modèle pour prévoir la précision. Par exemple, si les déplacements sont difficiles au sein de la parcelle, il peut être tentant de regrouper toutes les observations dans de petites zones (échantillonnage par grappes avec peu de grappes). Il est alors préférable d'observer au moins 3 ou 4 grappes séparées pour vérifier l'homogénéité. De plus, l'échantillonnage par grappes (11.2.4.2) permet de gagner du temps lors de l'observation. Ex : protocoles Vigiculture.

Si l'homogénéité soupçonnée de la structuration spatiale d'une variable observée est définit par une moyenne identique sur toutes les zones de la parcelle, une variance stable et de faibles corrélations spatiales alors, l'échantillonnage choisi sera de préférence aléatoire simple (garanti fiable en théorie) (11.3.3.1),

Dans le cas d'une homogénéité soupçonnée de la structuration spatiale d'un variable observée et en présence de corrélation spatiale, l'échantillonnage choisi sera systématique (pratique et adapté en cas de corrélation spatiale) (11.3.3.2), ou un autre plan permettant d'espacer les observations (losange, Z, W, X, U).

11.2.3 Parcelles vraisemblablement hétérogènes

11.2.3.1 Hétérogénéité spatiale causée par des facteurs environnementaux et techniques

L'hétérogénéité spatiale due aux facteurs environnementaux et/ou liée aux pratiques culturales peut induire un biais important dans la précision des résultats si le plan d'échantillonnage n'est pas bien choisi.

11.2.3.1.1 Hétérogénéités due à l'environnement

Le contexte pédoclimatique de la parcelle (ensoleillement, la texture et la structure du sol) et/ou la proximité d'une structure paysagère (haie, friche,…) peuvent être source d'hétérogénéité de la structuration spatiale de la variable observée (dégâts, bioagresseurs ou auxiliaires). La variabilité de ces facteurs entraine une variation des populations des organismes observés ou des phénomènes étudiées (pressions biotiques et régulations biologiques).
L'hétérogénéité peut être présentée sous forme de gradient ou de taches (zone bien délimitée). Dans les deux cas, cette hétérogénéité va générer une surestimation de la variance. De fait, la précision d'une moyenne calculée pourrait sembler peu fiable à cause de la variance élevée entre les observations.

11.2.3.1.1.1 Repérer les hétérogénéités spatiales

Afin de caractériser la structure spatiale d'une variable observée un premier échantillonnage ainsi qu'une évaluation visuelle rapide de la parcelle, peuvent permettre de révéler des hétérogénéités dont il faudra tenir compte lors de l'élaboration du plan d'échantillonnage.

Remarque : Si une variable explicative fortement corrélée à la variable mesurée est connue avec précision, elle peut être utilisée pour mettre en place un *échantillonnage à probabilités inégales.* Cette méthode, trop peu utilisée, n'est pas présentée dans ce document cependant, vous trouverez des renseignements dans les ouvrages de Gregoire and Valentine [2007] et de Gruijter et al. [2006].

11.2.3.1.1.2 Adapter le plan d'échantillonnage

Quand un gradient est vraisemblablement présent à l'échelle de la parcelle, les observations peuvent être rassemblées le long de transects (…) dans la direction principale du gradient [Gruijter et al., 2006, page 78] ou, un échantillonnage systématique (11.3.3.2) garantissant la représentativité de la variable observée peut également être mis en place.

Pour améliorer la précision et pouvoir mieux l'estimer *a posteriori*, il est possible de découper la parcelle en strates (11.3.4.1) dans lesquelles la structure spatiale de la variable étudiée serait plus homogène.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Exemples dans la littérature : Brown et al. [1993] (parasites du pommier), Goodell and Ferris [1981] (nématodes)

11.2.3.1.2 Hétérogénéité due aux pratiques culturales

Les diverses pratiques culturales (travail du sol, semis, cultures antérieures, traitements phytosanitaires) sont parfois source d'hétérogénéité de la structure spatiale de la variable observée. De plus, le caractère périodique de certaines de ces pratiques doit être pris en compte lors de l'élaboration de la stratégie d'échantillonnage.

11.2.3.1.2.1 Repérer les hétérogénéités spatiales

Lorsque l'itinéraire technique n'est pas le même sur l'ensemble de la parcelle étudiée, l'hétérogénéité de la structuration spatiale de la variable observée peut être constatée visuellement. Ex : une forte concentration de plantes adventices sur certaines zones de la parcelle.

11.2.3.1.2.2 Adapter le plan d'échantillonnage

Dès que la parcelle peut être découpée en zones (représentant une proportion connue de la surface totale) au sein desquelles la structuration spatiale de la variable observée est plus homogène que dans l'ensemble de la parcelle, alors une stratification (11.3.4.1) peut être mise en place (de Gruijter et al. [2006, page 78], Clostre et al. [2014]).

Ex : Lorsque les pratiques culturales risquent de présenter des variations récurrentes spatialement et à des distances égales, il faut veiller à ce que cela ne soit pas la source de biais qui remettrait en cause la fiabilité des résultats. Ainsi, le plan d'échantillonnage ne doit pas suivre le schéma spatial de la variable observée (cas de l'échantillonnage systématiques (11.3.3.2)).

Si l'objectif du diagnostic est de participer à l'évaluation d'un système de culture, certaines zones de la parcelle qui ne correspondent pas au système étudié peuvent être écartées. **Exemples** dans la littérature : Clostre et al. [2014] (Pollution des sols).

11.2.3.2 Hétérogénéité spatiale liée aux caractéristiques biologiques de la variable observée

11.2.3.2.1 L'agrégation spatiale

Largement étudiée en écologie (étude des populations animales [Taylor, 1984]), l'agrégation spatiale est constatée lorsque la variable observée n'est pas répartie aléatoirement dans l'espace mais qu'elle forme des agrégats.

Remarque : l'agrégation est plus facilement observable lors d'observations sur quadrats.

Un phénomène d'agrégation ne sera pas forcément rendu visible au moment de l'échantillonnage. En effet, si les objets observés forment des agrégats de quelques centimètres et que les relevés se font sur des quadrats de 1 mètre de côté, l'hétérogénéité due à l'agrégation ne sera pas mesurée. Si les agrégats sont beaucoup plus grands que les quadrats, on observera plutôt une corrélation (voir partie suivante) entre les résultats sur des quadrats proches."

En s'appuyant sur les données d'une observation déjà réalisée sur quadrats ou sur des organes, il est possible d'ajuster un modèle pour qu'il tienne compte de l'agrégation observée, ainsi que de déterminer un paramètre d'agrégation spécifique à la variable observée afin de prévoir la variance de l'échantillonnage à venir. La variance permettra in fine de déterminer la taille de l'échantillon.

11.2.3.2.1.1 Sources d'agrégation

L'agrégation spatiale des individus peut avoir des origines diverses, souvent liées à la biologie des espèces observées (les sources d'alimentation et la capacité de déplacement des insectes, le mode de reproduction d'une plante adventice, …).

Remarque : une forte densité d'individus implique en général une réduction de l'agrégation ou la rend moins visible. A l'inverse, les populations émergentes montrent fréquemment des phénomènes d'agrégation marqués. En 1984, Taylor [1984].

11.2.3.2.1.2 Modélisation

En présence d'agrégation spatiale, la loi binomiale négative (12.3.4) est souvent un bon modèle pour la distribution des résultats de comptages. Cette loi est paramétrée par sa moyenne et un paramètre *k* qui est petit pour une population très agrégée. Quand *k* est très grand, cette loi approche la *loi de Poisson*. Le paramètre *k* est parfois caractéristique d'une espèce et dans certains cas il peut être utilisé pour faire des prévisions. Il est important de vérifier la robustesse du modèle, c'est à dire la stabilité du paramètre sur plusieurs échantillonnages antérieurs.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

De même, la loi de puissance de Taylor (12.4.1), qui établit une relation entre la moyenne et la variance des résultats de comptages (pour une espèce et une zone de comptage données) est plus robuste que d'autres modélisations et permet par conséquent de bien prévoir la variance en cas d'agrégation spatiale.
Enfin, deux indices d'agrégation spatiale de Lloyd, quand ils se révèlent stables pour une stratégie d'échantillonnage donnée, peuvent aider à prévoir la variance. Chacun d'eux a une interprétation biologique qui pourrait permettre de l'estimer « à dire d'expert » pour ensuite faire des prévisions grossières, quand aucune autre information n'est disponible (12.4.3). Des correspondances existent entre certains paramètres de la puissance de Taylor et les indices, pour avoir une meilleure idée de l'agrégation de la variable étudiée. Si ces modèles ont initialement été prévus pour des comptages sur des zones, ils peuvent très bien s'adapter à d'autres situations, comme des comptages sur des organes, ou encore des piégeages.

11.2.3.2.1.3 Lien avec le plan d'échantillonnage

Si la taille des unités d'observation (quadrats, organes ou autres) varie, alors les paramètres des différents modèles d'agrégations risquent de varier aussi.
Exemple, plus les quadrats sont grands, plus les résultats de comptage deviennent proches de la moyenne sur toute la parcelle. La forme des quadrats peut aussi influencer la visibilité de l'agrégation. Pour maximiser la qualité des informations recueillies avec le moins de surface à examiner, il vaut mieux de nombreux petit quadrats plutôt que peu de grands. De cette façon, l'agrégation peut, en plus, être mieux quantifiée.

11.2.3.2.2 Corrélation spatiale

Les phénomènes de corrélation spatiale sont présents lorsque les relevés sont proches spatialement les unes des autres. La corrélation spatiale peut conduire à sous-estimer la variance de la variable observée, et donc à surestimer la précision obtenue. En fonction de l'objectif, de la variable étudiée et des informations disponibles sur la corrélation, plusieurs approches de modélisation pourront être adoptées.

11.2.3.2.2.1 Corrélation et incidence

En absence de corrélation spatiale, la loi binomiale est un bon modèle pour étudier l'incidence. En revanche, en présence de corrélation spatiale, cette loi conduit à une sous-estimation de la variance de la moyenne Cependant, il est possible de modéliser la corrélation par une expression corrigée de la variance c'est-à-dire avec un coefficient qui traduit la sur-dispersion.

La loi bêta-binomiale et d'autres modèles quant à eux [Collett, 1991, page 194], [Madden and Hughes, 1999] permettent d'obtenir une expression de la variance proche de celle que l'on pourrait obtenir avec la loi binomiale. Ils impliquent l'utilisation d'un paramètre d'ajustement (lien vers partie modèle) qui quand il est positif permet d'indiquer la sur-dispersion et quand il est nul indique que tous les individus observés sont indépendants.

11.2.3.2.2.2 Adaptation du plan d'échantillonnage

Pour étudier l'incidence d'une attaque, l'échantillonnage par grappes permet une modélisation qui rend compte de la différence de variance par rapport à la loi binomiale [Madden and Hughes, 1999]. A partir des paramètres ajustés du modèle (issus de la bibliographie ou d'échantillonnages passés), cette modélisation permet de prévoir la précision des résultats et d'estimer grossièrement par exemple l'incidence réelle (qui peut être obtenue par expérience ou avec un échantillonnage adaptatif (11.1.2.2)).
Cependant, pour la même précision, un échantillonnage par grappes aura un coût plus élevé qu'un échantillonnage simple ou systématique en présence de corrélation spatiale [Vaillant, 1996, page 35].

Exemple : ainsi, pour une étude de sévérité, il n'existe pas de modèle de corrélation, il vaudra donc mieux espacer les observations autant que possible par un échantillonnage systématique.

11.2.4 Adaptation à des contraintes pratiques

Les contraintes pratiques rencontrées peuvent être très diverses, dans le cas où les déplacements dans la parcelle sont très coûteux, deux possibilités peuvent être étudiées.

11.2.4.1 La stratification

La parcelle peut être découpée selon la difficulté d'échantillonnage dans chaque zone. Ainsi, l'échantillonnage peut être intensifié dans les zones plus faciles d'accès sans qu'un biais ne soit créé (9.1.1).

11.2.4.2 Les observations en grappes

Dans la mesure où la corrélation spatiale n'est pas trop importante, un échantillonnage par grappes (11.2.4.2) peut être adopté de façon à limiter les déplacements nécessaires.

11.3 Plan d'échantillonnage

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Le plan d'échantillonnage indique la répartition spatiale des observations à réaliser sur la parcelle. Pour éviter d'avoir un biais sur le résultat final, la localisation des observations devrait avoir un aspect aléatoire. Parfois, on néglige ce problème en faisant l'hypothèse (raisonnable, 15) que le choix de l'emplacement des observations donne un résultat représentatif de la parcelle.

Les plans d'échantillonnage se distinguent les uns des autres par leur caractère aléatoire ou non, ainsi que par leur complexité. Chacun des plans présentés ci-dessous ont des avantages et des inconvénients relatifs au coût et à la précision. Le plus souvent, le choix d'un plan est donc fondé sur un compromis.
La classification proposée ainsi que les noms proposés dans le document peuvent ne pas être partagés dans d'autres documents scientifiques.

11.3.1 Définitions

11.3.1.1 La population statistique

La population statistique est l'ensemble formé de ce qui va être tiré aléatoirement au moment de l'échantillonnage. Ce peut être l'ensemble des plantes d'une parcelle, l'ensemble des fruits, l'ensemble des quadrats possibles,... Chaque élément de la population statistique peut être appelé unité statistique, on parle aussi ici de relevés.

11.3.1.2 Quadrats

Des quadrats seront souvent utilisés pour des comptages ou des déterminations de biomasse. Le choix de leur taille pourra se faire selon des contraintes pratiques, selon les spécificités biologiques de ce qui est étudié et/ou en s'appuyant sur des protocoles déjà existants. Pour équilibrer la taille et le nombre de quadrats à examiner (la réflexion est la même que pour un échantillonnage par grappes et si la variable étudiée est corrélée spatialement, avoir un plus grand nombre de quadrats mais plus petits, permet un gain de précision. En contrepartie, cela augmente le coût de déplacement.

11.3.2 Échantillonnage raisonné

Un échantillonnage raisonné est un échantillonnage sans choix aléatoire. Il implique le choix réfléchi de certaines unités statistiques supposées représentatives. Il est intéressant car il permet simplement de tenir compte d'une expertise dans le domaine étudié néanmoins, il y existe un risque de biais lié à la subjectivité du concepteur du plan l'échantillonnage. En effet, contrairement aux échantillonnages aléatoires, on ne peut pas prouver théoriquement qu'un échantillonnage raisonné ne soit pas biaisé.

En pratique, une version raisonnée des plans d'échantillonnage proposés est parfois mise en place (ex : uniformisation de la répartition des points d'échantillonnage pour être représentatif de la parcelle), ce choix est moins préjudiciable dans le cadre de la mise en place de plans systématiques car le risque de créer un biais est moins important.

11.3.3 Échantillonnage aléatoire à un niveau

Pour pouvoir exploiter statistiquement les résultats des mesures, choisir de mettre en place un échantillonnage aléatoire fournit des garanties théoriques, ex : cela permet d'utiliser les résultats sur la précision de la moyenne. Les mêmes résultats sont parfois utilisés pour des échantillonnages raisonnés, mais cela présente un certain risque d'erreur. En effet, une moyenne est estimée non-biaisée (9.1.1) si chaque unité statistique a une probabilité non-nulle d'être observée (c'est une des définitions d'un échantillon représentatif).

11.3.3.1 Échantillonnage aléatoire simple

Un plan d'échantillonnage aléatoire simple est le choix aléatoire et indépendant d'un certain nombre d'unités statistiques de la population statistique cible (Figure 5).

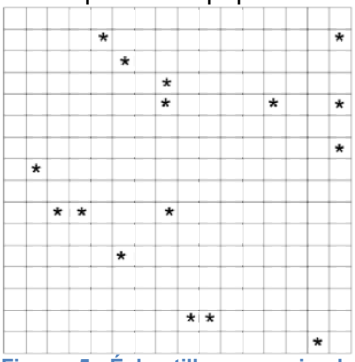

Figure 5: Échantillonnage simple de 16 unités de surfaces sur 256 - Source : Vaillant [1996]

11.3.3.1.1 Avantages

L'échantillonnage aléatoire simple ne requiert aucune connaissance *a priori* sur la population. De plus, son étude théorique est simple et les estimateurs courants pour la moyenne et la variance sont non-biaisés. L'absence de lien entre le choix des différents éléments est un avantage pour mettre en place un échantillonnage adaptatif. De plus, l'échantillonnage peut être interrompu ou poursuivi sans que cela ne cause de biais (à condition d'examiner les éléments dans l'ordre de tirage).

11.3.3.1.2 Inconvénients

L'échantillonnage aléatoire simple offre une précision minimale se traduisant par une variance élevée des estimateurs (à l'inverse rapport d'un plan systématique par exemple). De plus, il peut être difficile à réaliser car il nécessite une procédure de

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

tirage aléatoire qui puisse inclure n'importe quels éléments de la population, il faut donc que ces derniers soient tous repérés individuellement. Enfin, pour un échantillonnage aléatoire simple adaptatif, le coût en temps du déplacement et de repérage des unités statiques peut être élevé puisque elles peuvent être distantes spatialement les unes des autres.

11.3.3.1.3 Variantes approximatives

Si l'observateur choisit aléatoirement les emplacements de ses observations dans la parcelle, on considérera que les résultats théoriques sont toujours valides.
Exemple : les relevés peuvent être réalisés le long d'un parcours (comme ceux évoqués pour les échantillonnages systématiques).

11.3.3.1.4 Choix

Ce plan est le plus souvent utilisé pour réaliser un échantillonnage adaptatif, à condition que les déplacements soient peu coûteux. Par exemple le temps de déplacement sera plus court à l'intérieur de strates de petite taille (échantillonnage en strate). En cas de coût important, ce-dernier peut être réduit avec un plan aléatoire à deux niveaux (11.3.4).

11.3.3.2 Échantillonnage systématique

Un plan d'échantillonnage systématique repose sur un seul choix aléatoire, celui de l'élément de départ (Figure 6). A partir de celui-ci, les unités statistiques sont régulièrement espacées selon un quadrillage ou un parcours (en U, en diagonales,...) avant d'être examinées. Parfois, pour des raisons pratiques, l'unité statistique de départ est choisie sans tirage aléatoire.

Si les observations restent raisonnablement représentatives de la parcelle, on considère que le résultat ne sera pas biaisé.

Pour un échantillonnage systématique, le motif de répartition des relevés est fixé. S'il y a beaucoup de relevés, ils seront plus proches, s'il y en a moins ils seront plus dispersés. Ce type de plan est peu adapté à l'échantillonnage adaptatif, on peut néanmoins échantillonner en plusieurs phases en réalisant un premier échantillonnage systématique et on vérifie la précision obtenue. Si celle-ci ne répond pas aux objectifs de précision fixés, on complète le premier échantillonnage avec un second tout en s'assurant de la répartition systématique et/ou homogène de l'ensemble des relevés. Exemple : réaliser un premier parcours dans la parcelle qui serait complété avec un second parcours (deux diagonales).

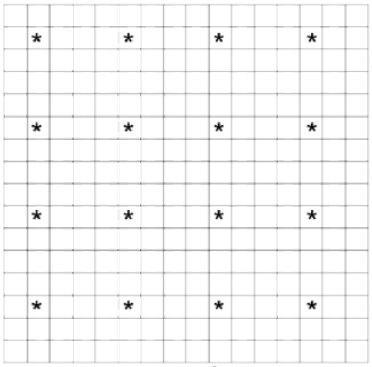

Figure 6 : Échantillonnage systématique de 16 unités de surfaces sur 256 - Source : Vaillant [1996]

11.3.3.2.1 Avantages

Le premier avantage de ce plan d'échantillonnage est la simplicité de repérage des relevés dans la parcelle. Ces derniers, bien espacés et bien répartis dans la parcelle, garantissent une bonne précision, même en cas de forte corrélation spatiale ou d'hétérogénéité à l'échelle de la parcelle [Gregoire and Valentine, 2007, de Gruijter et al., 2006, pages 49 à 56 et page 103 resp.]. Échantillonner selon une grille (Figure 6) a le double avantage d'espacer les observations et de garantir une répartition uniforme sur toute la parcelle. Une grille en quinconce maximise la distance entre les observations, et peut donc augmenter la précision.

Autre utilisation de ce type d'échantillonnage : une liste de classification (ex : croissante ou décroissante) d'objets à observer (unités statistiques) est réalisée selon un caractère spécifique de la variables étudiés. Dans cette liste est réalisé un échantillonnage systématique. Exemple, pour évaluer la sévérité d'une maladie apparaissant avec l'âge des arbres, dans un verger les 20 arbres échantillonnés sont classés par âges décroissants et un échantillonnage systématique est réalisé 1 arbres/2.

Échantillonner selon un parcours dans la parcelle, avec des observations espacées régulièrement, permet d'obtenir un aperçu assez large de la parcelle à faible coût. Ce choix permet en particulier de ne pas perdre de temps pour se repérer dans la parcelle. Le choix du parcours peut être, une diagonale pour aller vite (protocoles Vigiculture), deux diagonales pour une plus grande représentation du centre de la parcelle, un losange qui est pratique et qui représente bien la parcelle, un U pour pouvoir se déplacer le long des rangs (protocoles Rés0Pest), …

11.3.3.2.2 Inconvénients

On ne sait pas bien estimer la précision obtenue, bien qu'en général elle soit meilleure que celle observée pour un échantillonnage simple aléatoire. Un échantillonnage systématique peut induire un biais important si les relevés sont synchronisés spatialement avec les variations de la variable étudiée.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

11.3.3.2.3 Choix

L'échantillonnage systématique est un bon choix par défaut sauf si on souhaite mettre en place une démarche adaptative. Il faut alors vérifier si un ce type de plan laisse suffisamment de libertés (voir paragraphe précédent).

11.3.4 Échantillonnage aléatoire à plusieurs niveaux

L'échantillonnage à plusieurs niveaux consiste dans un premier temps à découper la population statistique totale en sous-populations (disjointes), puis à choisir des unités statistiques dans tout ou partie des sous-populations. Il permet de tirer parti d'une structuration existante de la population ou de la connaissance d'un variable corrélée à la variable étudiée.

11.3.4.1 Échantillonnage stratifié

Un plan d'échantillonnage stratifié est basé sur un découpage de la population statistique totale. Il se compose de plusieurs autres échantillonnages effectués dans chacune des sous-populations (strates) obtenues (Figure 7). Le nombre d'unités statistiques choisies, et la façon de les choisir dans les strates n'est pas forcément le même, ce qui permet de tenir compte des hétérogénéités de moyenne ou de variance dans la population statistique totale. La moyenne de la variable étudiée est calculée dans chaque strate avant d'être rassemblée en une moyenne pour toute la parcelle. Dans chaque strate échantillonnée de manière aléatoire, des échantillonnages simples systématiques ou par grappes peuvent être réalisés.

Taille d'échantillon

Si la taille de l'échantillon (total de tous les relevés toutes strates confondues) est fixée (modèle et couts fixe la taille l'échantillon) et qu'il faut répartir les relevés entre les strates, les deux choix les plus pertinents sont [Gregoire and Valentine, 2007, page 142] :

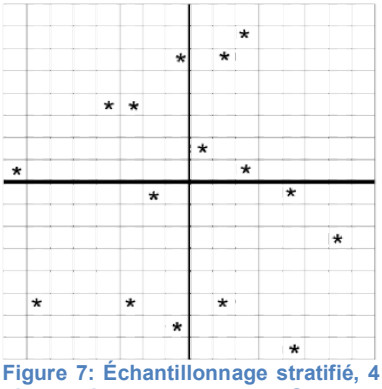

Figure 7: Échantillonnage stratifié, 4 observations par strates - Source : Vaillant [1996]

- Attribuer les relevés proportionnellement à la taille de strates (surfaces, nombre total de plante).
- Attribuer de manière optimale pour la précision finale. Dans ce cas, il faut avoir une estimation de la variance dans chaque strate avant de fixer la taille de l'échantillon. Si on a N observations à réaliser et I strates d'aires (ou d'importance)$a_1,\ldots,a_I$, avec les variances par strate $s_1,\ldots,\ s_I$, alors le nombre N_i d'observations à réaliser dans la strate j est

$$N_j = N\left(\frac{a_j s_j}{\sum_{i=1}^{I} a_i s_i}\right)$$

Il est aussi possible de procéder à des échantillonnages séquentiels (adaptatifs) indépendants dans toutes les strates, ou d'échantillonner de nouveau les strates dans lesquelles la précision s'est révélée être la moins bonne.

Calcul de moyenne

La moyenne sur toute la parcelle est la moyenne pondérée des moyennes dans chaque strate selon leur importance. Par exemple si on a I strates d'aires (ou d'importance)$a_i,\ldots,a_I$, avec les moyennes par strate $m_1,\ldots,m_I$, alors la moyenne pour toute la parcelle est :

$$m = \frac{1}{a_1+\cdots+a_I}\sum_{i=1}^{I} a_i m_i$$

Calcul de variance et précision

La variance de la moyenne de chaque strate doit d'abord être estimée selon le plan d'échantillonnage choisi dans la strate. Ensuite, en considérant que les moyennes sur les différentes strates ne sont pas corrélées. Les I variances $s_1^2,\ldots,s_I^2$, permettent de calculer la variance pour la moyenne sur toute la parcelle par la formule :

$$s_m^2 = \frac{1}{(a_1+\cdots+a_I)^2}\sum_{i=1}^{I} a_i^2 s_i^2$$

La variance obtenue, qui reflète la précision de la moyenne, est plus faible que la variance qui aurait été estimée directement sur toute la parcelle. Cependant, s'il y a peu d'observations dans chaque strate, l'estimation des variances peut être imprécise. D'ailleurs, Gregoire and Valentine [2007] indiquent que la loi de Student est préférable à la loi Normale pour établir les intervalles de confiance pour la moyenne.

Intervalles de confiance

Dans le cas où il y a assez d'observations dans chaque strate, en notant N le nombre total d'observations et I le nombre de strates, la formule proposée pour l'intervalle de confiance de niveau $(1-\alpha)$ est :

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

$$I_c = [m - t_{N-I,1-\alpha/2} \cdot s_m, m + t_{N-I,1-\alpha/2} \cdot s_m]$$

où $t_{N-I,1-\alpha/2}$est la quantile d'ordre $(1 - \alpha/2)$ de la loi de Student à $N - I$ degrés de liberté.

Remarque : Les strates ne sont pas forcément d'un seul tenant. Par exemple, les zones qui ont été couvertes par des andains et les autres peuvent être distinguées dans une parcelle de céréales pour l'évaluation de la densité de repousses spontanées.

11.3.4.1.1 Avantages

Des observations en strates, avec des relevés aux valeurs proches, permettent d'avoir une estimation précise des moyennes intra-strate. Quand elles sont rassemblées en une moyenne sur toute la parcelle, on obtient une variance plus précise que si elles avaient été obtenues avec un échantillonnage directement réalisé sur toute la parcelle.
La situation la plus appropriée pour la mise en place de ce type d'échantillonnage est lorsque les strates sont homogènes et qu'elles présentent de fortes différences entre elles [Vaillant, 1996, page 28].

Une variance homogène intra strate permet d'adapter le nombre d'observations dans chaque strate, pour avoir une précision comparable dans chacune d'elles et une bonne précision pour la moyenne totale, tout en économisant des observations dans les strates à faible variance.

En termes de coût, l'échantillonnage stratifié permet de réaliser les observations dans les zones plus accessibles sans pour autant biaiser le résultat.

Enfin, une stratification avec un échantillonnage aléatoire simple dans chaque strate permet de mieux répartir les observations dans la parcelle qu'avec un échantillonnage simple seul [de Gruijter et al., 2006, page 90].

11.3.4.1.2 Inconvénients

La principale difficulté est de pouvoir déterminer l'importance relative des strates (proportion), en surface ou en nombre total d'unités statistiques car en cas d'erreur le biais créé peut être important.
Exemple : surface ou nombre de plants/arbres, de chaque strate.

D'autre part, afin d'améliorer la précision, [Gregoire and Valentine, 2007, pages 128], il faut veiller au fait que le coût de mise en place d'une stratégie élaborée ne soit pas supérieur aux économies obtenues.

Enfin, dans chaque strate, si le nombre d'observations est trop faible, la variance et, par là même occasion, la précision ne seront pas bien évaluées (10.2, 15).

11.3.4.1.3 Choix des strates

Le choix de la stratification se justifie pour différentes raisons :

- des valeurs de la variable étudiée plus homogènes que sur l'ensemble de la parcelle.
- une variance pour la variable étudiée plus homogènes que sur l'ensemble de la parcelle.
- le coût d'échantillonnage pouvant être réduit (accès difficile en milieu de parcelle,...).

11.3.4.2 Échantillonnage en grappes

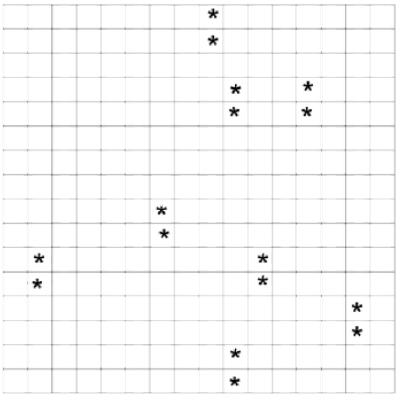

Figure 8: Échantillonnage en grappes, 8 grappes de 2 grains - Source : Vaillant [1996]

Pour un plan d'échantillonnage en grappes, les unités statistiques sont regroupées artificiellement ou selon une structure naturelle en grappes, c'est à dire en groupes de même taille d'unités proches ou consécutives. Ensuite, une partie d'entre elles est échantillonnée par un échantillonnage simple, systématique, ou autre. Enfin tous les éléments des grappes sélectionnées sont examinés (Figure 8).

Apres avoir défini le nombre total de relevés selon le coût ou la précision (via l'utilisation de modèles), la taille et le nombre de grappes peuvent être déterminées en s'assurant qu'il y ait assez de grappes pour que le résultat obtenu soit représentatif de la parcelle. De même, le nombre de grappes peut également être défini de manière à ce que les relevés soient regroupés afin de limiter le coût en temps. L'aide de modèles, le nombre de grappes optimal et la taille des grappes sont fixés comme un compromis entre le coût et la représentativité de la grappe (souvent 5 ou 10 plantes).
Dans certains cas la taille des grappes est fixée par le paramétrage d'un modèle (la loi beta binomiale (12.3.3), la loi binomiale (12.3.2), loi de puissance sur la variance (12.4.2)…) existant ainsi que le choix du nombre de grappes, se fait en

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

fonction de la précision désirée.

11.3.4.2.1 Avantages

Cet échantillonnage est souvent privilégié car il est moins couteux et il met en évidence l'hétérogénéité locale de la variable étudiée. En effet, les observations sont plus faciles à réaliser car elles sont proches les unes des autres (repérage, transport du matériel). Par ailleurs, dans le cas d'une étude d'incidence, celle-ci peut être calculée pour chaque grappe afin d'avoir une meilleure estimation de la précision. Cette dernière est alors basée sur la variance entre grappes (qui reflète les phénomènes de corrélation spatiale) plutôt que sur la loi binomiale). Enfin, les grappes sont propices à une modélisation (ou une simple constatation) des corrélations spatiales [de Gruijter et al., 2006, page 99].

11.3.4.2.2 Inconvénients

Dès que la corrélation spatiale est significative, les observations de la même grappe sont redondantes, elles représentent donc un coût inutile. De plus, le regroupement des observations en grappes se fait au détriment de la bonne représentativité de l'échantillon.

11.3.4.2.3 Choix

L'échantillonnage en grappes est privilégié lorsqu'il faut limiter les coûts et que la corrélation spatiale est négligeable. Il est pertinent d'utiliser cet échantillonnage lorsque les grappes sont semblables entre elles et qu'elles ont une forte variance interne [Vaillant, 1996, page 33].

Remarque : En pratique, l'échantillonnage de plusieurs plantes consécutives sur un même rang est courant dans le domaine de la protection des cultures. Il peut aussi être basé sur un rameau dans le cas de l'arboriculture. Cela limite le biais observateur. En effet, en sélectionnant plus d'une seule plante à un point donné, l'observateur ne peut pas choisir, de manière arbitraire, une plante plutôt qu'une autre.

11.3.4.3 Échantillonnage par niveaux

L'échantillonnage par niveaux est une notion très générale qui englobe tous les cas où il y a plusieurs étapes d'échantillonnage hiérarchisées. Quand la population possède une structure à plusieurs niveaux, une méthode d'échantillonnage peut ainsi être choisie pour le premier niveau. Ensuite, un nouvel échantillonnage est réalisé dans chaque élément sélectionné lors de la première étape et ainsi de suite (Figure 9). Ainsi un échantillonnage stratifié est un échantillonnage à deux niveaux, exhaustif pour le premier degré. Les plans d'échantillonnage par niveaux permettent une adaptation fine à chaque situation, au détriment de la simplicité d'organisation.

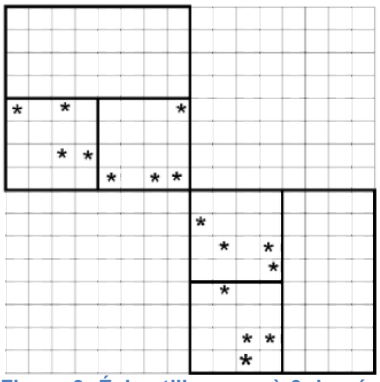

Figure 9: Échantillonnage à 3 degrés - Source : Vaillant [1996]

11.3.4.3.1 Échantillonnage simple à deux niveaux

Deux niveaux consécutifs d'échantillonnage simple, avec des zones choisies au premier niveau et un certain nombre d'observations réalisées dans chaque zone ensuite, réduit les déplacements nécessaires par rapport à un échantillonnage aléatoire simple à un seul niveau. De plus, en cas de corrélation spatiale, la perte de précision est plus faible qu'avec un échantillonnage par grappes [de Gruijter et al., 2006, page 95]. Dans un verger, le premier niveau peut naturellement consister à choisir des arbres, ensuite plusieurs fruits ou feuilles sont examinés sur le même arbre plutôt que de changer d'arbre pour chaque nouvel organe observé. Le traitement statistique peut rester le même que pour un échantillonnage à un niveau sans que cela ne pose de problème.

11.3.4.3.2 Échantillonnage selon des coupes

Des lignes peuvent être définies dans la parcelle le long desquelles les observations seront ensuite réalisées. La position des coupes peut être régulière ou aléatoire, de même que la position des observations le long des coupes, ce qui permet d'introduire un aspect aléatoire de plusieurs manières. Un tel échantillonnage peut être très pratique s'il suit des rangs. Il permet aussi de bien répartir les observations dans le cas où la variable observée présente un gradient sur la parcelle. Enfin, avec des coupes tirées aléatoirement, puis des observations espacées régulièrement, on obtient une sorte d'échantillonnage systématique, mais qui se prête mieux à l'échantillonnage séquentiel.
(Exemple : Barroso et al. [2005] et Clark et al. [2007]).

11.3.5 Échantillonnage composite

L'échantillonnage composite est possible pour des protocoles qui impliquent un prélèvement au niveau de chaque unité statistique pour une analyse ultérieure. A partir de n'importe quel plan, on peut rassembler le matériel prélevé

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

sous forme d'un ou plusieurs échantillons composites dans lesquels plusieurs prélèvements sont mélangés. Les comptages ou autre analyse ont ensuite lieu sur les échantillons composites. Cette méthode ne permet pas d'évaluer la précision du résultat mais peut être source d'économies très importantes, en particulier quand l'examen des unités statistiques coûte cher [Venette et al., 2002, page 158]. Ex : pour analyser la qualité de céréales, on utilise parfois un échantillon composite pour ne pas avoir à répéter des analyses en laboratoire longues et coûteuses.

12 Modèles

Les modèles permettent : (i) de mieux évaluer la précision obtenue lors d'un échantillonnage mais également, (ii) de prévoir la précision obtenue pour une taille d'échantillon. Pour cela, les informations disponibles *a priori,* à propos de l'objet observé, sont mises à profit. La pertinence des modèles est caractérisée par leur ajustement aux données d'échantillonnage existantes. Ainsi, la disponibilité de données *a priori* est cruciale. Celles-ci peuvent être des estimations de moyenne, des estimations de variance ou bien des estimations de répartition entre des classes d'observations. De même, elles peuvent être des données de paramètre d'une loi empirique, d'une distribution ou d'un variogramme empirique, ...
La liste des modèles proposée dans cette partie n'est pas exhaustive et s'appuie sur des études scientifiques en agronomie ou en écologie

12.1 Écarts-types et coefficients de variation

Après l'échantillonnage, dès que la variable mesurée est numérique, on peut calculer une estimation s de l'écart-type et une estimation $\hat{m}$ de la moyenne.

- Si les résultats des observations sont indépendants (*i.e.* non-corrélées), l'écart-type qui est observé en répétant le calcul de $\hat{m}$ sur plusieurs échantillons peut être approché par $\frac{s}{\sqrt{N}}$. C'est un premier indicateur de précision pour la moyenne estimée. On en déduit le coefficient de variation $c_v = \frac{s}{m\sqrt{N}}$.
- Si les résultats sont corrélés, alors le coefficient de variation et l'écart-type calculés sont faux ; ils sous-estiment la dispersion des valeurs $\hat{m}$ qui seraient obtenues par des échantillonnages répétés.

Grâce à (i) de la modélisation, (ii) un échantillonnage adaptatif, ou (iii) toute autre méthode, l'estimation de l'écart-type peut être disponible avant la fin de l'échantillonnage. Ainsi, la taille de l'échantillon N peut être adaptée pour atteindre la précision souhaitée. De fait, pour obtenir un écart-type sur la moyenne inférieure à s_0, il faut prendre :

$$N > \frac{s^2}{s_0^2}$$

De même, pour obtenir un coefficient de variation inférieur à c_v, il faut prendre :

$$N > \frac{s^2}{\hat{m}^2 . c_v^2}$$

Des informations sur l'objet observé permettent à l'observateur d'estimer la moyenne puis, via l'utilisation de modèle, d'estimer l'écart type. Ou bien, l'observateur peut utiliser un échantillonnage adaptatif pour obtenir une moyenne et un écart-type qui lui permettront de déterminer par la suite la taille de l'échantillon en fonction du c_v qu'il aura fixé auparavant.

12.2 Modèles pour la moyenne calculée

Pour N grand, la moyenne calculée après l'échantillonnage $\hat{m}$ suit de plus en plus une loi de probabilité connue. Cette loi permet d'établir des intervalles de confiance ou de donner une certitude pour les tests de dépassement de seuil. Ces modèles utilisent des approximations de la moyenne et de la variance qui leur sont associés (estimée par $\frac{s^2}{N}$). Comme précédemment, ces approximations peuvent être obtenues à la fin de l'échantillonnage, ou bien être disponibles bien avant afin de prévoir quelle sera la précision obtenue pour une taille d'échantillon donnée.

12.2.1 3.5.2.1. Loi Normale

La modélisation par la loi normale est appropriée pour toutes les moyennes de variables numériques à condition que le **nombre de relevés soit grand**. Pour un nombre intermédiaire (de 10 à 30 relevés), il vaut mieux faire preuve de prudence. En particulier, si les observations ne sont pas indépendantes ou si la répartition des valeurs est très asymétrique, les intervalles de confiance calculés risquent d'être faux.

12.2.1.1 Intervalle de confiance

Selon la Loi Normale, l'intervalle de confiance à $100(1-\alpha)\%$ pour la moyenne est :
$I_c=[\hat{m} - z_{1-\alpha/2}\frac{s}{\sqrt{N}}, \hat{m} + z_{1-\alpha/2}\frac{s}{\sqrt{N}}]$

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite donné par le tableau ci-dessous (Tableau 1).

Tableau 1: Quantiles de la loi normale centrée réduite

$\alpha(\%)$	1	5	10	20
$z_{1-\alpha/2}$	2.576	1.960	1.645	1.282

La demi-largeur de l'intervalle de confiance est donc :

$$d = z_{1-\alpha/2}\frac{s}{\sqrt{N}}$$

On en déduit, que pour une demi-largeur maximale d_0, il faut prendre :

$$N > (z_{1-\alpha/2}\frac{s}{d_0})^2$$

12.2.1.2 Etude de dépassement de seuil

Plus $\hat{m}$ est proche de m_0 plus il est difficile de confirmer si l'on est ou non en-dessous du seuil. Ainsi, il faut augmenter le nombre de relevés.

Avec un niveau de confiance où ϕ est la fonction de répartition de la loi normale, si la moyenne calculée $\hat{m}$ est supérieure à un seuil m_0, alors la moyenne réelle est supérieure à m_0 :

$$1-\alpha = \phi(\frac{\hat{m}-m_0}{s}\sqrt{N})$$

De même, la moyenne est entre les valeurs seuil m_1 et $m_2$avec un niveau de confiance :

$$1-\alpha = \phi\left(\frac{\hat{m}-m_1}{s}\sqrt{N}\right) - \phi\left(\frac{\hat{m}-m_2}{s}\sqrt{N}\right)$$

12.2.2 Loi de Student

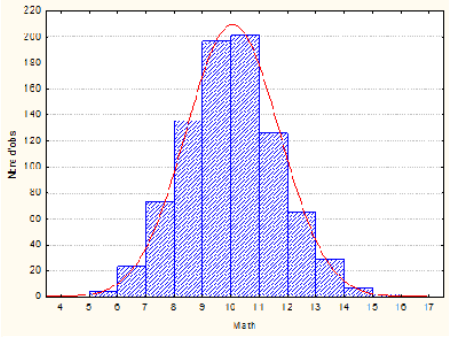

Figure 10: distribution normale d'un échantillon (http://w3.uohpsy2.univ-tlse2.fr).

Dans le cas où le nombre d'observations est faible, l'estimation de la variance est imprécise (15), ce qui perturbe l'estimation de la précision pour la moyenne. La Loi de Student utilisée pour modéliser la moyenne obtenue permet de tenir compte de cette imprécision. En revanche, elle n'est pertinente que si la variable X suit elle-même approximativement une répartition normale. Cette condition est vérifiée par des tests statistiques (Kolmogorov-Smirnov), ou plus approximativement grâce à un histogramme.

Si la variable est discrète avec suffisamment de valeurs différentes et si elle a un histogramme symétrique (Figure 10), alors on pourra également utiliser la loi de Student.

Les formules pour calculer les intervalles de confiance et la confiance associée au dépassement de seuil, sont les mêmes que celles utilisées dans le cadre de la loi normale, à ceci près que les quantiles et la fonction de répartition de la loi normale centrée réduite sont remplacés par ceux de la loi de Student à N degrés de liberté (= taille échantillon) (15).

12.3 Modèles sous forme de lois de probabilité pour la variable mesurée

Dans certains cas, des modèles plus élaborés permettent d'avoir une meilleure prévision de la précision, et donc d'ajuster la taille des échantillons. Un complément sur l'ajustement des modèles est disponible en annexe (15). Les quatre modèles développés reposent sur l'hypothèse que la taille de l'échantillon est petite par rapport à la taille de la population totale (population infinie).

12.3.1 Loi de Poisson

Dans des zones d'observation de même dimension, la loi de Poisson modélise en théorie le résultat d'un comptage d'individus répartis aléatoirement dans l'espace (sans agrégation et corrélation spatiale). Elle a la particularité d'avoir une variance égale à sa moyenne. On l'évoque ici à titre d'exemple puisqu'elle est rarement utilisée dans le domaine de la protection des cultures.

12.3.2 Loi binomiale

La loi binomiale de paramètres n et p, modélise le nombre de 1 obtenus pour n tirages indépendants d'une variable de Bernoulli de paramètre p (qui vaut 1 avec la probabilité p et 0 avec la probabilité $(1-p)$) (ça n'est pas la variable d'étude). Son espérance et sa variance sont :

$m = np \quad s^2 = np(1-p)$

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

La loi permet de modéliser des comptages de plantes sur lesquelles des organismes sont présents ou non (bioagresseurs, auxiliaires, …). Pour un nombre de plantes fixes connue n, p est alors la fréquence moyenne de plante avec la présence des organismes étudiés (ex : pour les bioagresseurs p est l'incidence).

En pratique, les observations peuvent parfois ne pas être indépendantes, et l'estimation de p sera surestimée.

12.3.2.1 Estimation et précision pour l'incidence

Ici n correspond au nombre d'observations N. ainsi l'estimation $\hat{p}$ de la fréquence sur l'ensemble des plantes est égale à la proportion d'observation positives (i.e. de plantes atteintes), soit en notant N_+le nombre de plantes atteintes parmi les N examinées :

$$\hat{p} = \frac{N_+}{N}$$

L'écart-type théorique pour $\hat{p}$ est alors:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Sous conditions, on en déduit un intervalle de confiance valide :
$I_{cp} = [p_i, p_s] = [\hat{p} - z_{\frac{1-\alpha}{2}} s_{\hat{p}}, \hat{p} + z_{\frac{1-\alpha}{2}} s_{\hat{p}}]$

Où $z_{\frac{1-\alpha}{2}}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite donné par le Tableau 1.

Cet intervalle de confiance est valable à condition que N_{p_i}, N_{p_s} , $N(1-p_i)$ et $N(1-p_s)$ soient supérieurs à 5, dans le cas contraire on ne peut pas utiliser les quantiles de la loi normale centrée réduite. Des intervalles de confiance exacts peuvent alors être calculés par des algorithmes spécialisés, comme ceux du package binom sous R [R Core Team, 2014, Dorai-Raj, 2014].

Dès qu'une pré-estimation de p est disponible, en faisant l'hypothèse d'indépendance des observations, on peut prévoir la variance de $\hat{p}$ et donc adapter la taille de l'échantillon. Ainsi, pour déterminer p avec un coefficient de variation c_v ou un demi-intervalle de confiance d donnés, le nombre minimal d'observations N_{min}devra vérifier :

$$N_{min} = \frac{1-p}{pc_v^2} \; ou \; N_{min} = z_{1-\alpha/2}^2 \frac{p(1-p)}{d^2}$$

12.3.2.2 En l'absence de modèle déjà ajusté

Avant l'échantillonnage, il est possible de réaliser une estimation vague de l'incidence qui sera observée, puis d'utiliser la loi binomiale pour en déduire la variance qui pourrait être observée. Cette approximation de la variance permet d'obtenir un ordre de grandeur de la variance. On remarque qu'avec ce modèle, celle-ci est plus élevée quand l'incidence est proche de 50%.

12.3.2.3 En présence de corrélation et sur-dispersion

En cas de corrélation spatiale (11.2.3.2), si les observations sont trop proches les unes des autres, alors elles ne seront pas indépendantes. Par conséquent la variance estimée comme précédemment pour $\hat{p}$ sera inférieure à celle observée en répétant l'échantillonnage.
Une corrélation spatiale peut être quantifiée lors d'un unique échantillonnage si les observations sont rassemblées en grappes (11.3.4.2). On considère alors N grappes de n observations au sein desquelles la probabilité p de faire une observation positive serait la même. Si les observations étaient indépendantes, la variance au sein de l'ensemblep_1,..., $p_N$des estimations de l'incidence dans chaque grappe devrait être égale à $\frac{p(1-p)}{n}$. Quand ce n'est pas le cas, un modèle de loi bêta-binomiale peut être pertinent dans le cadre de comptages dans des grappes.

12.3.3 Loi bêta-binomiale

La loi bêta-binomiale modélise le nombre de 1 obtenu pour n tirages d'une variable de Bernoulli dont le paramètre aléatoire et changeant à chaque tirage, suit une loi bêta. Elle permet de rendre compte de phénomènes de corrélation spatiale pour des variables telles que le nombre de plantes atteintes d'une maladie dans des grappes de n plantes (ce modèle est donc principalement adapté à un échantillonnage par grappes, 11.3.4.2).

Le paramètre intéressant est alors la proportion p d'individus malades (ou autre distinction). Si on a N grappes de n individus et donc N observations notées $X_1, \ldots, X_N$ à valeurs entre 0 et n, l'estimateur pour le paramètre p est :

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

$$\hat{p} = \frac{1}{nN}\sum_{i=1}^{N} X_i$$

Il existe plusieurs écritures classiques pour le paramétrage suivant ce que l'on veut mettre en évidence (espérance, variance, asymétrie,...). Si on prend des paramètres α et β de façon à ce que la densité de la loi bêta soit de la forme : $f(x) = kx^{\alpha-1}(1-x)^{\beta-1}$
(où α et β déterminent la forme et K est un coefficient de normalisation), alors l'espérance et la variance de la loi bêta-binomiale sont :

$$m = np = \frac{n\alpha}{\alpha+\beta}$$

$$s^2 = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

La loi bêta-binomiale peut aussi être paramétrée par n, p et un coefficient ρ qui reflète la sur-dispersion liée aux corrélations spatiales (Figure 11). C'est le paramétrage par défaut proposée par le package R VGAM [Yee, 2015].

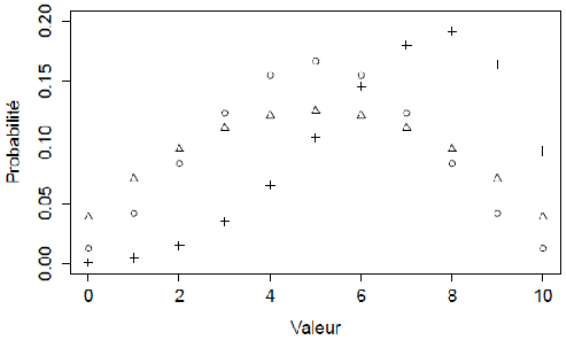

Figure 11: Densités de la loi bêta-binomiale pour n = 10, p = 0.5, ρ = 0.1 (◦) ; pour n = 10, p = 0.5, ρ = 0.2(Δ) ; pour n = 10, p = 0.7, ρ = 0.1(+)

12.3.3.1Intérêt

La loi bêta-binomiale est intéressante dans la mesure où le paramètre ρ est parfois stable entre plusieurs parcelles pour un stress biotique et une méthode d'observation donnés, ce qui permet d'utiliser le modèle pour prévoir la précision de l'échantillonnage à venir (ou mettre en place un échantillonnage adaptatif, 11.1.2.2). L'expression de la variance de la variable modélisée X est alors :

$$s^2 = np(1-p)(1+\rho-n-1))$$

On en déduit l'expression de la variance pour le paramètre p estimé à partir d'un ensemble de N valeurs (supposées indépendantes) :

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{nN}\big(1+\rho(n-1)\big)$$

On en déduit aussi une méthode d'estimation de ρ (valeur théorique) pour l'ajustement du modèle sur des données disponibles :

$$\hat{\rho} = \frac{1}{n-1}\left(\frac{s_X^2}{n\hat{p}(1-\hat{p})} - 1\right)$$

Cette modélisation est abordée plus en détail dans Madden and Hughes [1999]. Un exemple d'application dans le cadre de l'étude de la prédation des graines, se trouve en annexe (4.1).

12.3.3.2 Taille d'échantillon selon l'écart-type

A partir des formules précédentes, on déduit le nombre N_{min} de grappes qu'il faut observer pour obtenir en théorie un écart-type s sur l'estimation de p :

$$N_{min} = \frac{p(1-p)}{ns^2}\big(1+\rho(n-1)\big)$$

12.3.3.3 Taille d'échantillon selon l'intervalle de confiance

Si on dispose d'un nombre assez grand d'observations (en s'inspirant du cas de la loi binomiale, on peut prendre la condition que nNp_i, nNp_s , $nN(1-p_i)$ et $nN(1-p_s)$ soient supérieurs à 5), on peut construire des intervalles de confiance pour p à partir de l'écart-type en utilisant la formule basée sur la loi normale :

$$I_c = \left[\hat{p} - z_{1-\frac{\alpha}{2}} \cdot s_{\hat{p}}, \hat{p} + z_{1-\frac{\alpha}{2}} \cdot s_{\hat{p}}\right]$$

Si l'objectif de précision est d'avoir un intervalle de confiance de niveau (1- α) de demi-largeur inférieure à d il faut prendre :

$$N_{min} = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot p(1-p)}{nd^2}\big(1+\rho(n-1)\big)$$

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

12.3.3.4 Taille d'échantillon selon coefficient de variation

Si l'objectif de précision est un coefficient de variation c_v pour la valeur estimée de p, alors l'expression pour N_{min}est :

$$N_{min} = \frac{(1-p)}{npc_v^2}\left(1+\rho(n-1)\right)$$

12.3.3.5 Taille d'échantillon selon l'intervalle de confiance pour un dépassement de seuil

Comme pour les intervalles de confiance, on se base sur les résultats pour une moyenne suivant une loi normale dans les situations où ce choix est raisonnable.

Remarque : Si les grappes ne font pas la même taille, à cause de difficultés rencontrées lors des observations, on peut adapter les formules précédentes. On peut aussi se contenter de multiplier les valeurs de comptage dans les grappes incomplètes par un coefficient correcteur.

$$\frac{N_{souhaité}}{N_{réel}}$$

12.3.4 Loi binomiale négative

Avec la loi binomiale négative de paramètres k, p modélise le nombre de 1 tirés avant que k 0 aient été tirés pour des répétitions d'une variable de Bernoulli de paramètre p. L'espérance et la variance de cette loi sont :

$$\mu = \frac{kp}{1-p}$$

$$\sigma^2 = \frac{kp}{(1-p)^2}$$

Le paramètre p ainsi que le sens initial de la définition, ne sont pas utilisés en écologie car ils ont peu de sens. Ainsi, on utilise directement μ comme paramètre. Alors :
$p = \frac{\mu}{k+\mu}$ et $\sigma^2 = \frac{\mu(k+\mu)}{k}$

Grâce à sa forme asymétrique, qui attribue une probabilité non nulle à toutes les valeurs positives, la loi négative binomiale s'ajuste à de nombreuses données de comptage non-bornées (10.3.2 ; 10.3.1) pour une densité modérée d'individus à compter (en cas de saturation, la loi est moins pertinente d'après Rew and Cousens [2001]). Parker et al. [1997] soulignent que la loi peu s'ajuster trompeusement à des données trop peu nombreuses. Le paramètre k caractérise des phénomènes d'agrégation à l'échelle des unités d'observation (i.e. organe ou quadrat) [Rew and Cousens, 2001, page 7], k étant d'autant plus petit que l'agrégation est importante.

La loi binomiale négative n'est pas un modèle très robuste, elle doit être utilisée avec prudence pour déterminer des tailles d'échantillon. Si le paramètre k s'est révélé stable sur un bon nombre d'échantillons prélevés par le passé dans des conditions similaires (même stade de végétation, même taille de zone de comptage, même technique de comptage) et représentatives de la diversité de situations d'échantillonnage pouvant exister, la loi binomiale peut être utilisée pour déterminer la taille et/ou la précision de l'échantillon.

Exemples : La loi négative binomiale a été utilisée pour optimiser une stratégie d'échantillonnage par Moura et al. [2007] dans le cas de ravageurs présents sur des feuilles de pois. Pour un groupe de parcelles de la même région, la loi a expliqué la relation observée entre la moyenne et la variance des comptages. De même, Mukhopadhyay and Banerjee [2015] ont présenté de manière approfondie l'utilisation de la loi binomiale négative pour mettre en place un échantillonnage adaptatif en prenant pour exemple un comptage de doryphores.

12.3.4.1 Évaluer et améliorer la précision de l'échantillonnage via le coefficient de variation

La taille d'échantillon nécessaire pour obtenir une valeur donnée de coefficient de variation c_v est :

$$N_{min} = \frac{1}{c_v^2}\left(\frac{1}{\mu}+\frac{1}{k}\right)$$

où la valeur moyenne μ peut être estimée avant l'échantillonnage (ou au moins bornée inférieurement puisque N_{min}décroît en fonction de μ) ou pendant, par une procédure adaptative ou en plusieurs phases. Si on procède à un échantillonnage séquentiel, en notant T_N le nombre total d'individus comptés lors des N premières observations, on obtient la condition suivante sur T_N :

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

$$T_{N,min} = \frac{Nk}{c_v^2 Nk - 1}$$

Cette condition est valide dès que $N > \frac{1}{c_v^2 k}$, ce qui en pratique veut dire qu'il faudra faire au moins $\frac{1}{c_v^2 k}$ observations. Le diagramme (Figure 12) indique les paires N, T_N (zone grisée) pour lesquelles il est nécessaire de continuer à échantillonner dans le cas k = 1 pour obtenir un coefficient de variation de 25%.

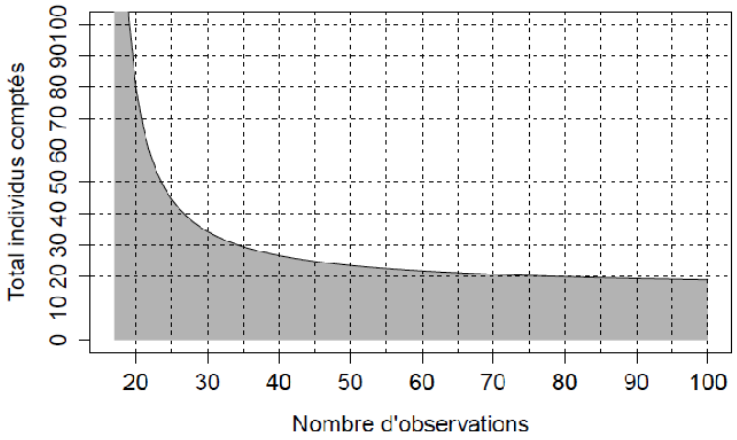

Figure 12: Diagramme permettant de simplifier l'échantillonnage séquentiel avec pour modèle une loi négative binomiale. Il faut poursuivre l'échantillonnage tant qu'on se trouve dans la zone en gris pour obtenir un coefficient de variation de 25%.

12.3.4.2 Évaluer et améliorer la précision de l'échantillonnage via l'écart-type

Pour un objectif de précision en termes d'écart-type (ou de demi-largeur d'intervalle de confiance, ce qui revient au même à un coefficient près), avec un écart-type maximal s, l'échantillonnage doit être poursuivi tant que le nombre total T_N d'individus comptés est supérieur à la limite :

$$T_{N,max} = \frac{Nk}{2}\left(\sqrt{(1 + \frac{Ns^2}{k}}) - 1\right)$$

Le diagramme (Figure 13) indique les paires N, T_N (zone grisée) pour lesquelles il est nécessaire de continuer à échantillonner dans le cas k = 1 pour obtenir un écart-type de 1. Cette démarche n'est pas valide pour un petit nombre d'observations, il sera donc judicieux d'en effectuer un certain nombre (par exemple 10) avant de se référer au diagramme pour l'échantillonnage séquentiel (11.1.2.2).

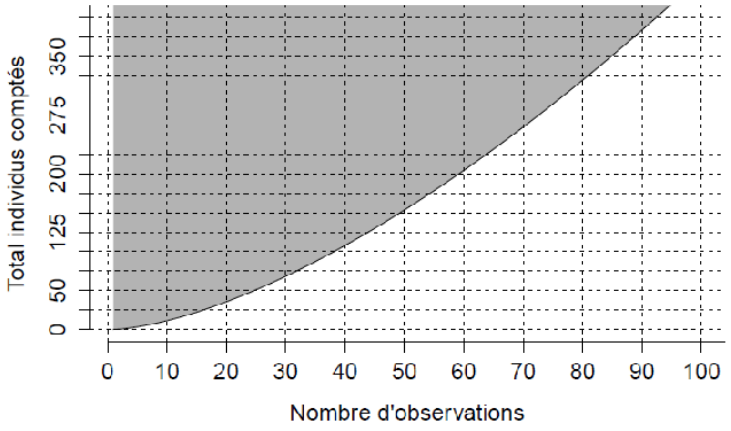

Figure 13: Diagramme pour simplifier l'échantillonnage séquentiel avec pour modèle une loi négative binomiale. Il faut poursuivre l'échantillonnage tant qu'on se trouve dans la zone en gris pour obtenir un écart-type de 1.

12.3.4.3 Ajustement du modèle

On peut déterminer le paramètre k du modèle à partir de la moyenne et de la variance, on peut aussi utiliser la fonction fitdistr du package MASS sous R [Venables and Ripley, 2002]. Connaissant la moyenne et la variance, l'estimation de k est :

$$k = \frac{s^2 - m}{m^2}$$

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Remarque : D'après Iwao [1968], pour un paramètre *k* assez grand et bien ajusté, un lien peut être fait avec l' *Index of Mean Crowding* proposé par Lloyd [1967] (12.4.3.1). Il y a un lien direct entre k et β, le paramètre de la relation entre m et $\dot{m}$: $\beta = 1 + \frac{1}{k}$

12.3.5 Loi multinomiale

La loi multinomiale est une généralisation de la loi binomiale, elle correspond au résultat d'un certain nombre de tirages *indépendants* d'une variable aléatoire avec un nombre fini de valeurs possibles (10.3.2) et une probabilité fixe associée à chaque valeur. Elle modélisera par exemple le résultat de *N* observations pour lesquelles k valeurs (ou classes) sont possibles $(v_1, v_2, .., v_k)$ et $p_1, p_2, \dots, p_k$ sont les probabilités associées (donc en pratique les proportions des observations qui ont une certaine valeur). On a forcément $\sum_{i=1}^{k} p_i = 1$, de plus le nombre de fois où la valeur v_i est tirée suit une loi binomiale de paramètres n, p_i. Si la variable aléatoire tirée est quantitative, alors sa moyenne sur N tirages indépendants et la variance de cette moyenne sont :

$$m = \sum_{i=1}^{k} v_i p_i$$

$$s^2 = \frac{1}{N}\left(\sum_{i=1}^{k} v_i^2 p_i (1 - p_i) - 2 \sum_{i \neq j \in \{1,\dots,k\}} v_i v_j p_i p_j\right)$$

12.3.5.1 Estimation des probabilités pour chaque classe

On peut vouloir estimer les paramètres p_i grâce à des observations. En notant N_i le nombre de fois où la valeur de la classe v_i a été observée, sur N observations, l'estimateur est :

$$\hat{p}_i = \frac{N_i}{N}$$

Des intervalles de confiance exacts existent mais leur expression est trop compliquée pour être utilisée. Wang [2008, page 901] présentent plusieurs intervalles de confiance approchés de niveau $1 - \alpha$. Le premier est basé sur les quantiles d'ordre $1 - \alpha$ de la loi du χ^2 à $k - 1$ degrés de liberté :

$$I_{\hat{p}_i,\alpha} = \left[\hat{p}_i - \sqrt{\chi^2_{k-1,1-\alpha}}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}, \hat{p}_i + \sqrt{\chi^2_{k-1,1-\alpha}}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}\right]$$

Un autre est basé sur les quantiles d'ordre $1 - \alpha/2k$ de la loi normale centrée réduite :

$$I_{\hat{p}_i,\alpha} = \left[\hat{p}_i - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}, \hat{p}_i + z_{1-\alpha/2k}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N}}\right]$$

Des intervalles de confiance peuvent aussi être calculés automatiquement grâce au package R MultinomialIC [Villacorta, 2012].

12.3.5.2 Coefficients sur les classes

Si comme Aubertot et al. [2004], des coefficients sont attribués à chaque classe et qu'il est nécessaire d'optimiser la précision sur la moyenne de ce coefficient, alors l'outil *N-Index* proposé sur le site *Quantipest* [IPM Network, 2013] peut être mis à profit. En replaçant les probabilités p_i par leurs estimations $\hat{p}_i$, les estimations $\hat{m}$ de la moyenne et $\hat{s}^2$ de la variance peuvent être obtenues. Par conséquent, quand une pré-estimation est disponible pour chaque proportion p_i, il est possible de prévoir la précision correspondant à une taille d'échantillon donnée.

12.3.5.3 Observations corrélées

Si les observations ne sont pas indépendantes, la précision exprimée par les indicateurs précédents sera certainement surestimée.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

12.4.1 Loi de puissance de Taylor

Il s'agit d'une loi empirique, du nom du chercheur britannique en écologie Lionel Roy Taylor. Sa pertinence a été largement établie en écologie [Taylor, 1984, page 338], et elle est fréquemment utilisée en agronomie (par exemple par Nyrop et al. [1989] et Navarro-Campos et al. [2012]). Elle se résume à la relation entre variance et moyenne de résultats de comptages pour une population donnée (10.3) :

$$\sigma^2 = am^b$$

Le coefficient a dépend de la taille des zones de comptage et d'autres détails pratiques, en revanche le paramètre b peut être caractéristique d'une espèce ; plus il est élevé plus l'agrégation est forte.
La loi de puissance de Taylor est utilisée pour le même type d'échantillonnages que la loi binomiale négative : quand chaque observation consiste en un comptage (dans des quadrats ou sur des organes de plantes essentiellement).

12.4.1.1 Ajustement du modèle

Les deux coefficients peuvent être déterminés par une régression linéaire du logarithme de la variance par rapport au logarithme de la moyenne. Celles-ci sont calculées à l'aide d'échantillonnage réalisés suite à la mise en place d'un même protocole d'observation (mais pas forcément avec la même taille d'échantillon (0)).

12.4.1.2 Taille d'échantillon selon coefficient de variation

Si une estimation préalable de la moyenne m est disponible, la taille N_{min} de l'échantillon à prendre pour obtenir un coefficient de variation inférieur à c_v est :

$$N_{min} = \frac{am^{b-2}}{c_v^2}$$

Dans les cas où on a une fourchette de valeurs pour m, on peut calculer la taille correspondant aux deux valeurs et prendre la plus grande (celle pour m petit si $b < 2$, celle pour m grand sinon).
Sans estimation préalable de la moyenne, on peut procéder de manière séquentielle, la précision est alors suffisante si le nombre total d'individus comptés T_N après N observations est supérieur à :

$$T_{N,min} = \sqrt[b-2]{\frac{c_v^2 N^{b-1}}{a}}$$

Comme pour la loi binomiale négative, un diagramme peut être réalisé pour mettre en évidence les situations dans lesquelles l'échantillonnage doit être poursuivi pour obtenir la précision satisfaisante.

12.4.1.3 Taille d'échantillon selon l'écart-type ou l'intervalle de confiance

Si une estimation préalable de la moyenne m est disponible, la taille N_{min}de l'échantillon à prendre pour obtenir un écart-type pour la moyenne estimée inférieur à s *(écart-type maximal que l'on ne veut pas dépasser)* est :

$$N_{min} = \frac{am^b}{s^2}$$

Pour une approche séquentielle, il est nécessaire de continuer à échantillonner tant que le nombre total d'individus comptés T_N après N observations est supérieur à :

$$T_{N,max} = \sqrt[b]{\frac{s^2 N^{b+1}}{a}}$$

Dans un objectif de précision, à partir d'un intervalle de confiance connu, on peut utiliser les formules de la loi normale et en déduire l'écart-type qui permettra de déterminer la taille minimale d'échantillon nécessaire pour la précision recherchée. La formule de la loi normale utilisée est la suivante :

$$I_c=[\hat{m} - z_{1-\frac{\alpha}{2}}s; \hat{m} + z_{1-\alpha/2}s]$$

12.4.2 Loi binomiale de puissance

La loi binomiale de puissance est utilisée dans le cadre de l'étude de fréquence (ex : incidence) pour des groupes de taille fixe. Elle établit une relation mathématique entre la variance théorique du cas binomial (indépendance) et la variance observée. Elle s'utilise dans le même contexte que la loi bêta-binomiale [Madden and Hughes, 1999] (10.3).
On a donc N observations aboutissant à des valeurs$X_1,.., X_N$comprises entre 0 et n, et on s'intéresse au paramètre p, correspondant à la moyenne$\frac{1}{nN}\sum_{i=1}^{N} X_i$, c'est à dire à l'incidence moyenne si les valeurs correspondent à des nombres de plantes attaquées dans des grappes de n plantes.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

On s'intéresse alors aux proportions $x_i = \frac{X_i}{n}$ (qui sont les incidences dans chaque grappe). Si l'état de chaque plante au sein d'une grappe est indépendant des autres, la variance entre les valeurs $x_1, \ldots, x_N$ serait donnée par la formule :

$s^2_{bin} = \frac{p(1-p)}{n}$ (Loi binomiale)

La loi binomiale de puissance prend en compte la corrélation entre l'état des différentes plantes d'une grappe grâce au modèle suivant. Ainsi la variance calculée est :

$$s_x^2 = A \cdot (s_{bin}^2)^b \Leftrightarrow \log(s^2) = A + b\log(s_{bin}^2) = A + b\log\left(\frac{p-(1-p)}{n}\right)$$

où *A* et b sont des paramètres plus ou moins caractéristiques d'une espèce et d'une méthode d'observation donnée.

Pour simplifier les expressions, on prendra souvent $a = An^{-b}$. Comme pour les autres modèles, les paramètres a et b peuvent alors être ajustés sur la base d'échantillonnages passés, puis utilisés pour prévoir la précision qui serait observée lors d'un nouvel échantillonnage à partir d'une pré-estimation grossière de p. L'expression de la variance pour l'estimateur $\hat{p}$ de p est alors :

$$s_{\hat{p}}^2 = \frac{s_x^2}{N} = \frac{a}{N}(p(1-p))^b$$

12.4.2.1 *Taille d'échantillon selon le coefficient de variation souhaité*

Si on s'attend *a priori* à observer une proportion p d'individus atteints, la taille de l'échantillon nécessaire pour obtenir un coefficient de variation $c_v$fixé est :

$$N_{min} = \frac{a}{c_v^2}p^{b-2}(1-p)^b$$

La formule de la courbe nécessaire pour mettre en place un échantillonnage séquentiel ne se simplifie pas, un diagramme pourra toutefois être établis numériquement [Madden and Hughes, 1999, page 1095].

12.4.2.2 *Taille d'échantillon selon l'écart-type souhaité*

Si on s'attend à observer une proportion p d'individus atteints, la taille de l'échantillon nécessaire pour obtenir un écart-type s fixé est :

$$N_{min} = \frac{a}{s^2}(p(1-p))^b$$

Concernant l'échantillonnage séquentiel, la situation est la même que pour le cas précédent.

12.4.2.3 *Ajustement du modèle*

Les recommandations sont les mêmes que pour la loi de puissance de Taylor.

12.4.3 Indice de Lloyd et régression d'Iwao

Ces modélisations s'appliquent aux mêmes situations que la loi de puissance de Taylor (12.4.1). Elles sont, elles aussi, basées sur la moyenne et la variance de résultats de comptage. Deux indices distincts ont été introduits par Lloyd [1967] :

12.4.3.1 *Index of mean crowding :*

$$\dot{m} = m + \frac{s^2}{m} - 1$$

12.4.3.2 *Index of patchiness :*

$$P = \frac{\dot{m}}{m} = 1 + \frac{s^2 - m}{m^2}$$

$\dot{m}$ peut être interprété comme le nombre moyen de voisins d'un individu compté. P peut être interprété comme la densité moyenne autour d'un individu compté, il peut être relié à la loi binomiale négative par la correspondance $P = 1 + \frac{1}{k}$ [Lloyd, 1967]. Les deux indices peuvent être caractéristiques d'espèces données, ils dépendent par contre du plan d'échantillonnage et de la taille des quadrats.

Iwao [1968, éq. 4] propose d'utiliser $\dot{m}$ pour modéliser l'agrégation spatiale avec une relation entre $\dot{m}$ et m :

$$\dot{m} = \alpha + \beta m$$

Cette relation est utilisée par Navarro-Campos et al. [2012] et Burts and Brunner [1981] pour optimiser des tailles d'échantillon. Elle est aussi reprise par Waters et al. [2014] qui propose de fixer α = 0, alors β = P. Cette modification se justifie théoriquement en constatant qu'il n'est pas cohérent d'avoir une agrégation non nulle quand il n'y a aucun

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

individu. Waters et al. [2014]. Ce choix l'a conduit à prendre des échantillons plus petits, et donc à réduire le coût, sans que cela se traduise par une perte de la précision qui avait été fixée initialement.

12.4.3.3 Taille d'échantillon selon le coefficient de variation souhaité

La taille d'échantillon minimale pour obtenir un coefficient de variation c_v en connaissant les paramètres α et β est :

$$N_{min} = \frac{\frac{\alpha+1}{m} + (\beta - 1)}{c_v^2}$$

Pour un échantillonnage séquentiel, avec les mêmes notations que pour la loi de Taylor, on obtient la condition :

$$T_{N,min} = \frac{\alpha+1}{c_v^2 - \frac{\beta-1}{N}}, N > \frac{\beta-1}{c_v^2}$$

On remarque que la formule obtenue correspond bien à celle que l'on peut obtenir avec la loi Binomiale négative, à la différence près du coefficient α.

12.4.3.4 Taille d'échantillon selon l'écart-type souhaité

La taille d'échantillon minimale pour obtenir un écart-type s en connaissant les paramètres α et β est :

$$N_{min} = \frac{m}{s^2}(\alpha + 1 + (\beta - 1)m)$$

13 Exemples

13.1 Etude du taux de prédation de graines adventices

Dans le cadre d'une étude sur la prédation de graines adventices par des carabes, des cartes de prédation avec deux types de graines (sur des cartes différents) ont été disposées sur deux parcelles [Trichard et al., 2014]. Sur chaque parcelle, le taux de prédation a été étudié sur 33 stations géo-référencés (Figure 14) pour huit sessions d'observations.

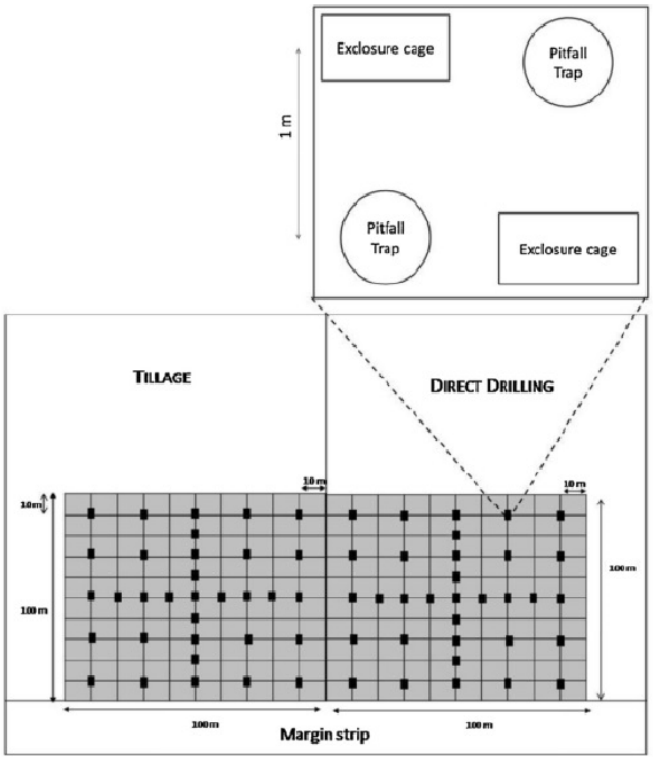

Figure 14: Stations d'observation pour l'étude de la prédation de graines

A partir de ces données, la répartition des observations sera réévaluée et la taille de l'échantillonnage nécessaire pour bien estimer le taux de prédation moyen p de graines sur une parcelle sera recalculé.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

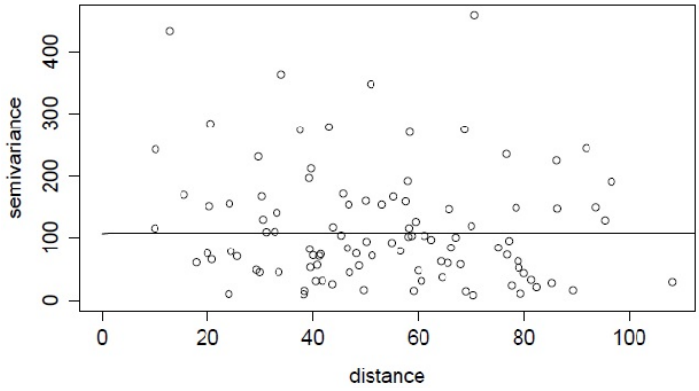

Figure 15: Variogramme pour un groupe de 33 observations simultanées

Des variogrammes (tracés avec le package geoR, Jr and Diggle [2015]) ne montrent pas de corrélation spatiale pour des distances allant de 10m à 100m (Figure 15). A partir de ce résultat deux hypothèses peuvent être émises :

- un résultat comparable aurait peut-être obtenu en plaçant les observations aléatoirement (mais en essayant de garantir une distance minimale de l'ordre de 10m entre deux observation).
- l'estimation s^2 faite de la variance n'est pas biaisée, et le taux de prédation moyen calculé a une précision que l'on peut caractériser par la variances^2/N.

Bien que les données soient des taux moyens de prédation sur deux cartes de la même station, elles représentent le nombre de graines consommé parmi 100 graines. Ainsi, il est possible d'utiliser des modèles compatibles avec des variables entières à valeurs bornées (10.3.2).

En représentant sur un graphe les variances observées et les variances théoriques selon la loi binomiale, on constate qu'elle (ou une loi de puissance) n'est pas pertinente. En effet, elle ne reflète pas la forte corrélation entre la prédation des graines d'une même carte, elle sous-estime par conséquent la variance. En revanche, la loi bêta-binomiale s'ajuste assez bien aux données (fonctions du package VGAM, Yee [2015]). La stabilité approximative du paramètre de sur-dispersion ρ entre les sessions (Figure 17) indique que ce modèle pourrait être utilisé pour prévoir la variance. Ici, à titre d'exemple, la valeur moyenne des ρ ajustés a été réutilisée pour modéliser les résultats des huit sessions (Figure 16), de cette façon on obtient par modélisation un ordre d'idée de la qualité des données obtenues.

Si ρ est connu, on peut déduire la variance entre les observations à partir du taux de prédation moyen. Si le taux de prédation a lui-même été estimé grossièrement avant une session, il est possible de choisir un nombre d'observations adapté à la précision souhaitée pour l'estimation plus précise du taux de prédation. Au vu des graphiques en Figure 17, la variance estimée à partir du taux de prédation et du paramètreρ, pour chaque session et chaque modalité, est

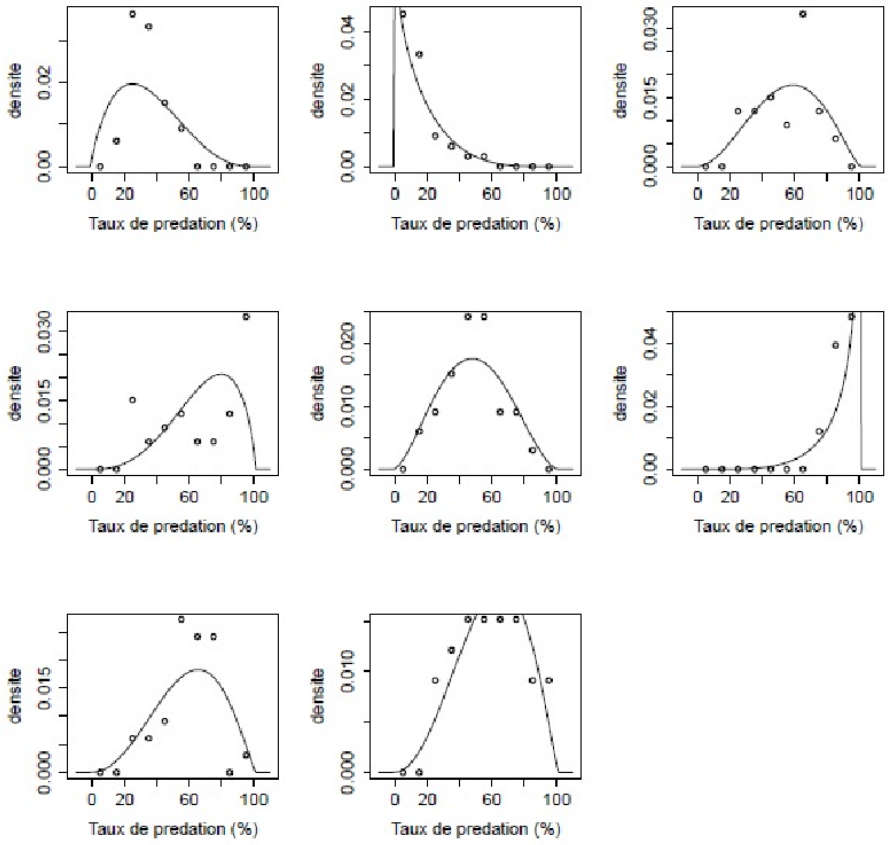

Figure 16: Loi bêta-binomiale ajustée aux 8 sessions de mesure pour une des modalités, avec le même paramètre de sur-dispersion ρ

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

bien de l'ordre de la variance réelle observée. Toutefois, l'erreur de prédiction est importante, il faut donc être prudent au moment du choix de la taille *N* des échantillons.

Ici, une démarche adaptative est difficile à mettre en place car chaque l'observation nécessite la pose d'une carte de prédation et de son relevé sept jours plus tard. En revanche, la taille d'échantillon pourrait être estimée à l'avance en fonction des mesures précédentes et de l'expertise de l'observateur (influence de la météo, du stade de culture,...).

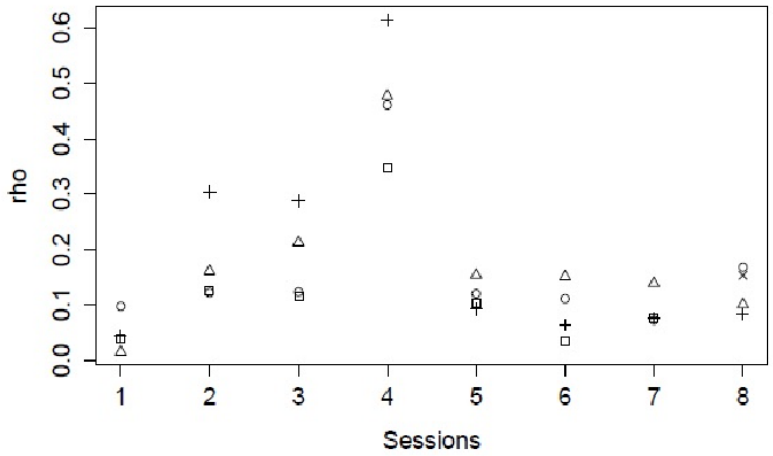

Figure 17: Valeurs estimées du paramètre de sur-dispersion de la loi bêta-binomiale pour les quatre modalités (symboles distincts) et pour les 8 sessions de mesures.

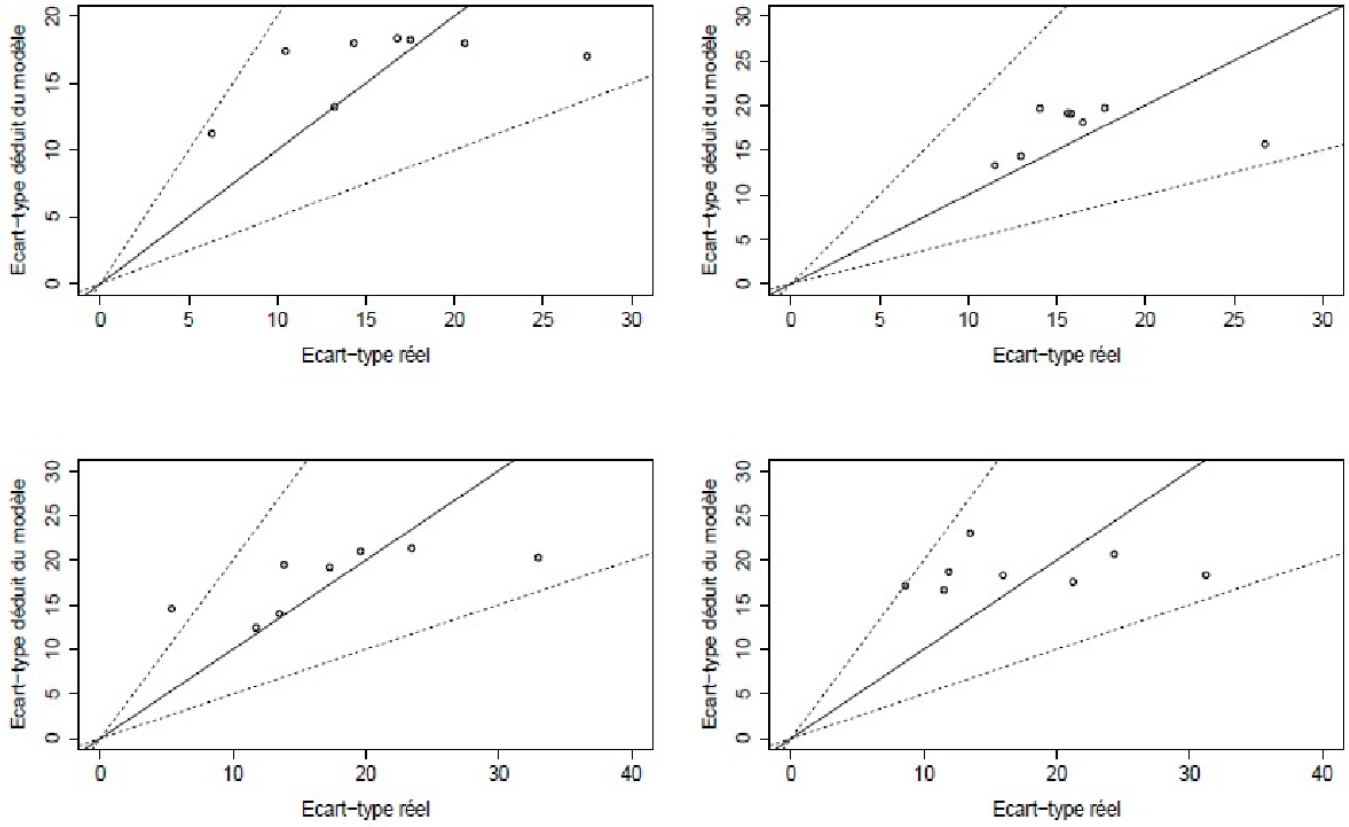

Figure 18: Ecart-types (en %) réels et déduits à partir du modèle pour les taux de prédation selon quatre modalités expérimentales, pour 8 sessions de mesure par modalité. Les trois droites, de pentes 0.5, 1 et 2 indiquent l'erreur réalisée.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

13.2 Etude de suivi des piégeages de carpocapse

L'étude a été menée par l'unité PSH (Plantes et Systèmes de culture Horticoles) de l'INRA Avignon. Les données (fournies par Claire Lavigne) résultent de l'utilisation de pièges à carpocapse qui ont été répartis de manière systématique dans quarante-huit vergers au sud d'Avignon (Figure 19).

Figure 19: Disposition des pièges à insectes dans un verger.

Pour chaque verger et chaque piège, les comptages d'insectes capturés se caractérisent principalement par un très grand nombre de zéros. La loi binomiale négative n'a qu'un sens limité puisqu'elle consiste à rassembler les données de tous les vergers qui semblent très hétérogènes. De plus, le paramètre *k* n'est pas stable sur l'ensemble des vergers. En revanche, une loi de puissance de Taylor (relation entre la moyenne et la variance) semble bien s'ajuster aux données (Figure 20). Ainsi, il existe, entre la moyenne et la variance dans chaque verger, une relation de la forme $s^2 = a.m^b$. Dans ce cas concret, via une régression linéaire sur le logarithme des données, on obtient approximativement a = 2.55 et b = 1.379.

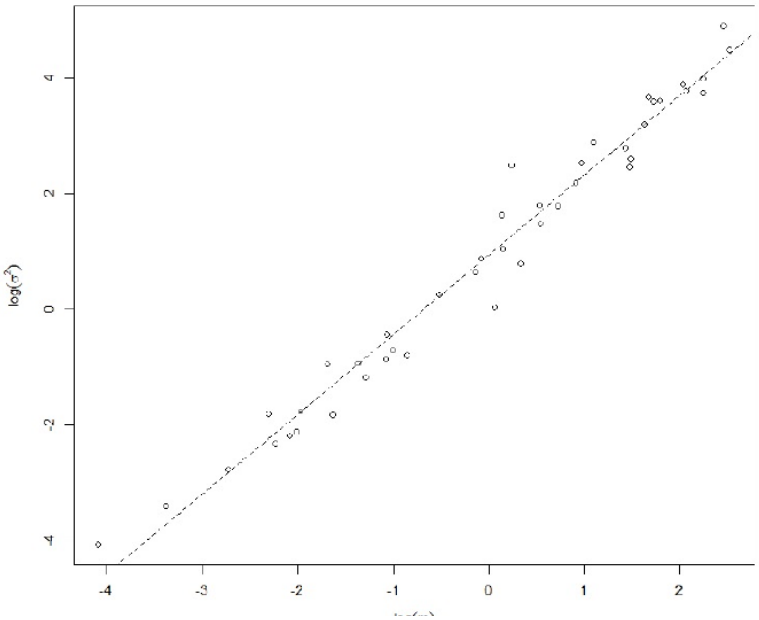

Figure 20: Ajustement d'une loi de puissance de Taylor entre la moyenne et la variance des résultats de piégeage dans 48 vergers

A partir d'une estimation grossière de la moyenne du nombre d'insectes piégés sur une parcelle (L'estimation peut venir de résultat sur des vergers voisins, de connaissances sur la dynamique de la population, ou encore d'une première phase d'échantillonnage), la relation ajustée permet de déduire la variance de ce nombre. Ainsi, un nombre minimal d'observations N_{min} à réaliser peut être calculé en fonction de la précision souhaitée. Par exemple, si le coefficient de variation pour la moyenne doit être inférieur à c_v = 0.25, il faudrait prendre :

$$N_{min} = \frac{am^{b-2}}{c_v^2} = \frac{40.8}{m^{0.62}}$$

Si la précision est définie par un demi-intervalle de confiance à 95% de largeur d = 1 et pour garantir que le nombre d'observations permettent d'avoir une moyenne qui suive une loi normale, alors il faudrait prendre :

$$N_{min} = \frac{1.96am^b}{d^2} = 5m^{1.379}$$

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Remarque : Lors de piégeages, il n'est pas pertinent de parler du nombre total d'insectes sur la parcelle. En effet, les données récupérées combinent la densité d'insectes et l'efficacité des pièges sans qu'il ne soit possible de distinguer l'une de l'autre. Ainsi, l'objectif de l'étude est d'estimer la moyenne du nombre d'insectes capturés par piège, au sens où cette moyenne se stabilise pour un grand nombre de pièges.

Sur certain vergers, la densité d'insecte semble corrélée spatialement, même si cette impression n'est pas confirmée par les variogrammes. Par conséquent, il est important que les observations soient réparties uniformément dans tout le verger, ce qui peut être garanti par un échantillonnage systématique (pratique à mettre en place) comme c'est le cas ici, ou par un échantillonnage par strate de petite taille avec peu d'observation par strate (par exemple : définir des blocs carrés de 16 arbres et observer 1 arbre par bloc au hasard). Ces deux plans d'échantillonnage s'adaptent bien à un changement de taille d'échantillon, il suffit de plus espacer les pièges ou d'agrandir les strates. En revanche, ils ne seraient pas adaptés pour un échantillonnage adaptatif car il serait difficile de choisir dans quel ordre réaliser les observations. L'échantillonnage adaptatif est, de toute façon, inadapté à un protocole avec des pièges devant rester en place un certain temps avant d'être relevés.

13.3 Etude de la sévérité du Phoma du Colza

L'étude porte sur la répétabilité des mesures de sévérité du phoma sur le colza [Aubertot et al., 2004]. Sur trois zones de 9m², la sévérité de la maladie a été relevée sur l'ensemble des plantes Elle a été notée selon six classes de sévérité, auxquelles ont été associés les scores v_1 = 0, v_2 = 1, v_3 = 3, v_4 = 5, v_5 = 7 et v_6 = 9.

Les variogrammes estimés à partir des trois jeux de données indiquent une absence de corrélation spatiale pour des distances allant jusqu'à 3m, comme on le voit sur la Figure 21. Ce résultat corrobore ceux obtenus par Aubertot et al. [2004] pour des distances plus importantes.

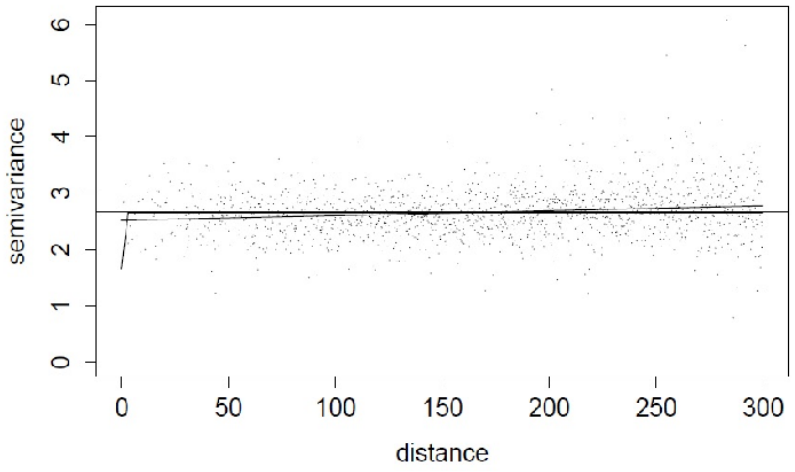

Figure 21 : Nuée variographique et variogrammes ajustés pour le score de sévérité du phoma du colza pour des distances de 20cm à 300cm

Ainsi, la précision des résultats ne serait donc pas dégradée par un échantillonnage des plantes en grappes. De plus, la précision associée à la proportion de chaque classe peut être facilement estimée avec la loi multinomiale, de même que la précision pour un score moyen. Par exemple, pour N observations avec respectivement p_1= 0%,p_2 = 10%, p_3= 40%, p_4= 40%, p_5= 10% et p_6= 0% des observations dans chaque classe, le coefficient de variation serait :

$$c_v = \frac{s}{m\sqrt{N}} = \frac{\sum_{i=1}^{6} v_i^2 p_i(1-p_i) - 2\sum_{i\neq j\in\{1,\dots,6\}} v_i v_j p_i p_j}{\sqrt{N}\sum_{i=1}^{6} v_i p_i}$$

$$\Longrightarrow c_v = \frac{0.4}{\sqrt{N}}$$

Ainsi, pour obtenir une précision donnée avec un c_v = 0.05, il faudrait un nombre d'observations N_{min} donné par le calcul suivant :

$$N_{min} = \frac{0.4^2}{0.05^2} = 64$$

Dans le cas où on n'a qu'une estimation très grossière de la répartition des plantes entre les six classes, il vaut mieux calculer N_{min} avec une estimation favorisant les classes à haut score. N_{min} sera ainsi assez élevé pour compenser la plus forte variabilité.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Finalement, pour obtenir le score moyen de sévérité du phoma sur une parcelle (en absence de corrélations spatiales) tout en vérifiant l'absence de motifs à plus grande échelle, trois recommandations peuvent être faites :

- Les observations peuvent être placées aléatoirement où selon n'importe quel motif couvrant plutôt bien la parcelle (en U par exemple).
- Si les plantes sont prélevées par grappes (ex : 8 grappes de 8 plantes), la précision ne devrait pas être réduite.
- Le nombre d'observation nécessaire pour obtenir une précision définie par un coefficient de variation donné peut être déterminé à l'aide des formules ci-dessus.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

14 Références bibliographiques

J.-N. Aubertot, J.-J. Schott, A Penaud, H Brun, and T. Doré. Methods for Sampling and Assessment in Relation to the Spatial Pattern of Phoma Stem Canker (Leptosphaeria maculans) in Oilseed Rape. *European Journal of Plant Pathology*, 110(2) :183–192, February 2004. ISSN 0929-1873. doi : 10.1023/B:EJPP.0000015359.61910.3b. URL http://link.springer.com/10.1023/B:EJPP.0000015359.61910.3b. 3.1, 3.2.1, 3.3.3.4, 3.3.3.4, 3.5.3.5, 4.3

J Barroso, D Ruiz, E S Leguizamon, P Hernaiz, A Ribeiro, and B Diaz. Comparison of sampling methodologies for site-specific management of Avena sterilis. *Weed Research*, pages 165–174, 2005. 3.1, 3.6.1, 3.7.4.3

M. W. Brown, F. Szentkirályi, and F. Kozár. Spatial and temporal variation of apple blossom weevil populations (Col., Curculionidae) with recommendations for sampling. *Journal of Applied Entomology*, 115(1-5) :8–13, January 1993. ISSN 09312048. doi : 10.1111/ j.1439-0418.1993.tb00358.x. URL http://doi.wiley.com/10.1111/j.1439-0418.1993.tb00358.x. 3.1, 3.6.1, 3.6.2

E C Burts and J F Brunner. Dispersion statistics and sequential sampling plan for adult pear psylla. *Journal of Economic Entomology*, 74(3) :291–294, 1981. 3.5.4.3, 3.6.1 S. J. Clark, P. Rothery, J. N. Perry, and M. S. Heard. Farm Scale Evaluations of herbicidetolerant crops : Assessment of within-field variation and sampling methodology for arable weeds. *Weed Research*, 47(2) :157–163, 2007. ISSN 00431737. doi : 10.1111/j.1365-3180.2007. 00541.x. 3.1, 3.7.4.3

Florence Clostre, Magalie Lesueur-Jannoyer, Raphaël Achard, Philippe Letourmy, Yves Marie Cabidoche, and Philippe Cattan. Decision support tool for soil sampling of heterogeneous pesticide (chlordecone) pollution. *Environmental Science and Pollution Research*, 21(3) : 1980–1992, 2014. ISSN 09441344. doi : 10.1007/s11356-013-2095-x. 3.6.3, 3.6.3

David Collett. *Modelling Binary Data*. Chapman and Hall, 1991. ISBN 0-412-38790-5. 3.6.5

Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics.

John Wiley & Sons, Inc., Hoboken, NJ, USA, September 1993. ISBN 9781119115151. doi :10.1002/9781119115151. URL http://doi.wiley.com/10.1002/9781119115151.ch1http://doi.wiley.com/10.1002/9781119115151. 3.6.5, 5

Jaap J. de Gruijter, Marc F. P. Bierkens, Dick J. Brus, and Martin Knotters. *Sampling for Natural Resource Monitoring*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-22486-0. doi : 10.1007/3-540-33161-1. URL http://link.springer.com/10.1007/3-540-33161-1. 1.2, 1.2.1, 3.1, 3.6.2, 3.6.2, 3.6.3, 3.7.3.2, 3.7.4.1, 3.7.4.2, 3.7.4.3

P. Domburg, J.J. de Gruijter, and D.J. Brus. A structured approach to designing soil survey schemes with prediction of sampling error from variograms. *Geoderma*, 62(1-3) :151–164, 1994. ISSN 00167061. doi : 10.1016/0016-7061(94)90033-7.

Sundar Dorai-Raj. *binom : Binomial Confidence Intervals For Several Parameterizations*, 2014. URL http://cran.r-project.org/package=binom. 3.5.3.2

P B Goodell and H Ferris. Sample optimization for five plant-parasitic nematodes in an alfalfa field. *Journal of nematology*, 13(3) :304–313, 1981. 3.6.2

Timothhy G. Gregoire and Harry T. Valentine. *Sampling Strategies for Natural Resources and the Environment*. Chapman and Hall/CRC, 2007. ISBN 9781584883708. 3.6.2, 3.7.3.2, 3.7.4.1, 3.7.4.1, 3.7.4.1

Jie Guan and Forrest W. Nutter. Quantifying the intrarater repeatability and interrater reliability of visual and remote-sensing disease-assessment methods in the alfalfa foliar pathosystem. *Canadian Journal of Plant Pathology*, 25(2) :143–149, 2003. ISSN 0706-0661. doi : 10.1080/07060660309507062. 3.2.1

IPM Network. Quantipest Website, 2013. URL http://www6.inra.fr/quantipest. 3.5.3.5

S. Iwao. A new regression method for analyzing the aggregation pattern of animal populations. *Researches on Population Ecology*, 10(1) :1–20, 1968. ISSN 00345466. doi : 10.1007/BF02514729. 3.5.3.4, 3.5.4.3

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Paulo J Ribeiro Jr and Peter J Diggle. *geoR : Analysis of Geostatistical Data*, 2015. URL http://cran.r-project.org/package=geoR. 4.1

Monte Lloyd. Mean crowding. *The Journal of Animal Ecology*, pages 1–30, 1967. URL http://www.jstor.org/discover/10.2307/3012?uid=3738016&uid=2&uid=4&sid=21106863488913. 3.5.3.4, 3.5.4.3

L V Madden and Gareth Hughes. Sampling for Plant Disease Incidence. *Phytopathology*, 89 : 1088–1103, 1999. 3.3.3.1, 3.3.3.2, 3.5.3.3, 3.5.4.2, 3.5.4.2, 3.6.5, 3.6.5, 5

Laurence V. Madden, Gareth Hughes, and Frank van den Bosch. *The Study of Plant Disease Epidemics*. APS Press, 2006. ISBN 978-089054-354-2. URL http://www.thestudyofplantdiseaseepidemics.org/. 3.3.3.4, 5

M. F. Moura, M. C. Picanço, R. N. C. Guedes, E. C. Barros, M. Chediak, and E. G. F. Morais. Conventional sampling plan for the green leafhopper Empoasca kraemeri in common beans. *Journal of Applied Entomology*, 131(3) :215–220, April 2007. ISSN 0931- 2048. doi : 10.1111/j.1439-0418.2006.01113.x. URL http://doi.wiley.com/10.1111/j. 1439-0418.2006.01113.x. 3.1, 3.5.3.4

Nitis Mukhopadhyay and Swarnali Banerjee. Sequential negative binomial problems and statistical ecology : A selected review with new directions. *Statistical Methodology*, 26 :34–60, 2015. ISSN 15723127. doi : 10.1016/j.stamet.2015.02.006. URL http://www.sciencedirect.com/science/article/pii/S1572312715000192. 3.5.3.4

C. Navarro-Campos, A. Aguilar, and F. Garcia-Marí. Aggregation pattern, sampling plan, and intervention threshold for Pezothrips kellyanus in citrus groves. *Entomologia Experimentalis et Applicata*, 142(2) :130–139, February 2012. ISSN 00138703. doi : 10.1111/j.1570-7458. 2011.01204.x. URL http://doi.wiley.com/10.1111/j.1570-7458.2011.01204.x. 3.3.2,
3.5.4.1, 3.5.4.3, 3.6.1

F. W. Nutter, Jr. Assessing the Accuracy, Intra-rater Repeatability, and Inter-rater Reliability of Disease Assessment Systems, 1993. ISSN 0031949X. 3.2.1

J P Nyrop, a M Agnello, J Kovach, and W H Reissig. Binomial Sequential Classification Sampling Plans for European Red Mite (Acari : Tetranychidae) with Special Reference to Performance Criteria. *Journal of Economic Entomology*, 82 :482–490, 1989. URL http: //www.ingentaconnect.com/content/esa/jee/1989/00000082/00000002/art00031. 3.2, 3.5.4.1

P. S. Ojiambo and H. Scherm. Optimum Sample Size for Determining Disease Severity and Defoliation Associated with Septoria Leaf Spot of Blueberry. *Plant Disease*, 90(9) :1209–1213, September 2006. ISSN 0191-2917. doi : 10.1094/PD-90-1209. URL http://apsjournals. apsnet.org/doi/abs/10.1094/PD-90-1209. 3.6.1

S. R. Parker, M. W. Shaw, and D. J. Royle. Measurements of spatial patterns of disease in winter wheat crops and the implications for sampling. *Plant Pathology*, 46(4) :470–480, 1997. ISSN 00320862. doi : 10.1046/j.1365-3059.1997.d01-38.x. URL http://doi.wiley.com/10. 1046/j.1365-3059.1997.d01-38.x. 3.1, 3.5.3.4

I Pulakkatu-Thodi. Within-Field Spatial Distribution of Stink Bug (Hemiptera : Pentatomidae)- Induced Boll Injury in Commercial Cotton Fields of the Southeastern United States. *Environmental Entomology*, 43(3) :744–752, 2014. doi : http://dx.doi.org/10.1603/EN13332. URL http://ee.oxfordjournals.org/content/43/3/744. 3.1, 3.1

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.r-project.org/. 3.5.3.2

L. J. Rew and R. D. Cousens. Spatial distribution of weeds in arable crops : Are current sampling and analytical methods appropriate ?, 2001. ISSN 00431737. 3.5.3.4, 3.6.5

D R Ring, M K Harris, and J a Payne. Sequential sampling plan for integrated pest management of pecan nut casebearer (Lepidoptera : Pyralidae). *Journal of Economic Entomology*, 82(3) : 906–909, 1989. ISSN 0022-0493. doi : 10.1093/jee/82.3.906. 3.3, 3.4.2.2

Lionel Roy Taylor. Assessing and Interpreting the Spatial Distributions of Insect Populations, 1984. ISSN 00664170. 3.5.4.1, 3.6.4, 3.6.4, 5

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Aude Trichard, Benoit Ricci, Chantal Ducourtieux, and Sandrine Petit. The spatio-temporal distribution of weed seed predation differs between conservation agriculture and conventional tillage. *Agriculture, Ecosystems & Environment*, 188 :40–47, 2014. ISSN 01678809. doi : 10.1016/j.agee.2014.01.031. URL http://www.sciencedirect.com/science/article/pii/S0167880914000577. 4.1

Jean Vaillant. Notions de base en échantillonage, 1996. URL http://www7.inra.fr/mia/ftp/V/FPstat/module6/version1/Vaillant.pdf. 3.6.5, 3.8, 3.9, 3.10, 3.7.4.1, 3.11, 3.7.4.2, 3.12,5

W N Venables and B D Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. 3.5.3.4

Robert C Venette, Roger D Moon, and William D Hutchison. Strategies and Staitistics of Sampling for Rare Individuals. *Annual Review of Entomology*, pages 143–174, 2002. 3.7.5

Pablo J Villacorta. *MultinomialCI : Simultaneous confidence intervals for multinomial proportions according to the method by Sison and Glaz*, 2012. URL http://cran.r-project.org/package=MultinomialCI. 3.5.3.5

Hsiuying Wang. Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, 99(5) :896–911, May 2008. ISSN0047259X. doi : 10.1016/j.jmva.2007.05.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X07000784. 3.5.3.5

Edward KyleWaters, Michael J. Furlong, Kurt K. Benke, James Robin Grove, and Andrew John Hamilton. Iwao's patchiness regression through the origin : Biological importance and efficiency of sampling applications. *Population Ecology*, 56(2) :393–399, 2014. ISSN 14383896.doi : 10.1007/s10144-013-0417-y. 3.5.4.3

F Workneh, G L Tylka, X B Yang, J Faghihi, and J M Ferris. Regional Assessment of Soybean Brown Stem Rot, Phytophthora sojae, and Heterodera glycines Using Area-Frame Sampling : Prevalence and Effects of Tillage. *Phytopathology*, 89(3) :204–211, 1999. ISSN 0031-949X. doi : 10.1094/PHYTO.1999.89.3.204. 3.6.1

Thomas W Yee. *VGAM : Vector Generalized Linear and Additive Models*, 2015. URL http://cran.r-project.org/package=VGAM. 3.5.3.3, 4.166

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Annexes

15 Statistiques

15.1 Source d'aléa pour l'échantillonnage

Une approche statistique nécessite que les données obtenues soient théoriquement aléatoires, c'est-à-dire qu'elles sont composées de variables aléatoires. L'aspect aléatoire s'explique soit par la procédure d'échantillonnage, soit par le type de phénomène observé qui suit un modèle aléatoire. Cette distinction est importante pour justifier théoriquement l'utilisation de certains modèles.

15.2 Estimateur

En statistiques, un estimateur est la valeur supposée d'un paramètre de loi de probabilité ou de modèle obtenue à partir de données issues d'un sondage, d'un échantillonnage, d'une expérience,... L'estimateur d'un paramètre est caractérisé par sa formule. On s'intéresse à son biais (qui est nul si la formule est choisie judicieusement) et à sa variance (entre des valeurs de l'estimateur qui seraient obtenues à partir de plusieurs échantillonnages indépendants).

15.3 Estimation de la variance

En générale, l'estimation de la variance d'une variable échantillonnée est très imprécise car elle requiert une taille d'échantillon importante.

15.3.1 Variable suivant une loi normale

Lorsque la variable étudiée suit une loi normale (par exemple des résultats de mesures de biomasse), on connait la loi de l'estimateur non-biaisé de la variance :

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{m})^2$$

Plus précisément, la quantité $\frac{(n-1)s^2}{\sigma^2}$ (où σ^2 est la variance réelle) suit une loi du χ^2 ce qui permet d'établir des intervalles de confiance pour la variance et l'écart-type. Pour un niveau de confiance $1-\alpha$, on a l'inégalité :

$$\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}$$

On en déduit les intervalles de confiance suivants :

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}$$

Il est important de noter que ces intervalles de confiance ne sont pas symétriques par rapport à l'estimation s^2 ou s.

Exemple, pour estimer l'ordre de grandeur de la taille d'échantillon nécessaire pour avoir une estimation la plus précise de la variance, si la variance d'une grandeur au sein d'une population est de 1, l'estimation de la variance obtenue à partir de 50 observations sera entre 0.8 et 1.3 dans seulement 75% des cas (sur des répétitions nombreuses de l'échantillonnage aléatoire). Pour 1000 observations, l'estimation de la variance sera entre 0.93 et 1.07 dans 90% des cas.

15.3.2 Variable ne suivant pas une loi normale

Pour une variable de loi continue mais asymétrique, une transformation logarithmique peut permettre de corriger cette asymétrique afin de se trouver dans une situation de loi normale (cas précédent).

15.4 Le variogramme

En présence de corrélation spatiale, un variogramme est un graphique qui représente l'écart au carré moyen des valeurs observées entre deux points, en fonction de la distance entre ces points. Il ne peut être défini convenablement que s'il n'y a pas de tendances à l'échelle de la parcelle et si la variance y est homogène (condition de stationnarité).

En général, un variogramme présente une partie croissante partant ou non de l'origine puis, un palier qui correspond à la variance de la variable concernée (Figure 22). La partie du variogramme avec des valeurs inférieures à la variance correspond aux distances pour lesquelles une corrélation spatiale existe. Ainsi, si le variogramme est plat, on peut supposer qu'il n'y a pas de corrélations spatiales pour les distances étudiées. La distance à partir de laquelle la valeur du variogramme est comprise dans un intervalle de ±5% autour du palier est appelée portée, et on considère qu'il n'y a plus de corrélation significative à partir de cette distance.

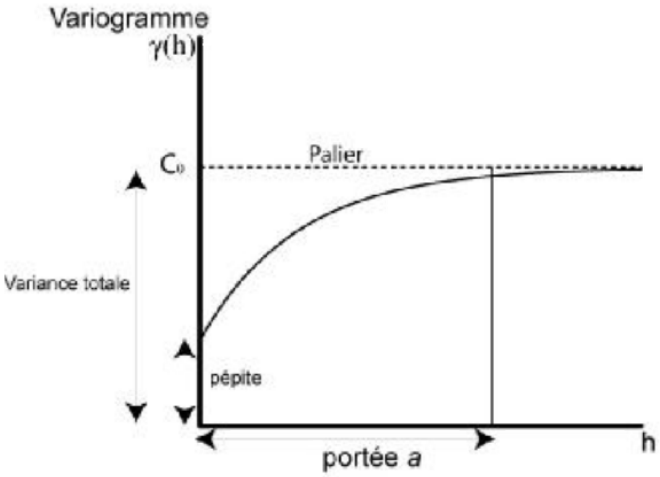

Figure 22: Variogramme

Un variogramme empirique peut être obtenu à partir des valeurs de la variable étudiée Z en un certain nombre de points de l'espace des coordonnées$x_1, \ldots, x_I$. Pour une distance h, le variogramme vaut la moitié de la moyenne des écarts au carré des valeurs en des points distants d'environ h. On a donc la formule suivante :

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{h-\delta|x_i-x_j|<h+\delta} \left(Z(x_i) - Z(x_j)\right)^2$$

où $n(h)$ est le nombre de paires de points dont la distance est comprise dans [$h - \delta$, $h + \delta$] et $\gamma(h)$ est la valeur du variogramme. Pour que le résultat soit utilisable il vaut mieux faire les calculs avec un intervalle de tolérance [$h - \delta$, $h + \delta$].

15.4.1 Obtenir un variogramme

Dans le cadre des études scientifiques présentées en exemple (13.1, 13.3), l'analyse des données a montré qu'il est difficile d'obtenir un variogramme de qualité sans un jeu de données très important.
Néanmoins, si un variogramme est obtenu pour une espèce donnée et dans un système de culture donné, il peut être utilisé indépendamment de la méthode d'observation pour choisir la distance entre deux observations ; si possible supérieure à la portée [Rew and Cousens, 2001]. De plus, Rew and Cousens [2001] signale, que la portée dépend fortement du système de culture qui influence la dispersion des différents organismes.

15.4.2 Estimation de variance

Si les observations sont réparties à une distance régulière les unes des autres, la corrélation entre deux observations successives peut être déduite du variogramme. Il est alors envisageable de déterminer la valeur de la variance pour la moyenne sur tout l'échantillon en compensant la sous-estimation due à la corrélation spatiale. Ainsi, s'il y a un coefficient de corrélation ρ entre deux observations successives, alors la variance de la moyenne sur N observations est :

$$s_{\bar{m}}^2 = \frac{s^2}{N'}$$

où s^2 est la variance estimée sur les N observations et N'est défini par :

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

$$N' = \frac{N}{1 + 2\frac{\rho}{1-\rho}\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2 \frac{1-\rho^{N-1}}{N}}$$

On obtient par exemple pour N = 10 et ρ = 0.26, N' = 6.2. En fonction de la précision voulue, et à condition d'avoir bien estimé le coefficient de corrélation, on peut calculer N' puis N. Le coefficient de corrélation peut être déduit du variogramme avec la formule :

$$\rho = 1 - \frac{\gamma(h)}{s_0^2}$$

où $\gamma(h)$ est la valeur du variogramme pour la distance h entre deux observations consécutives et s_0^2est la variance pour la variable observée, égale à la valeur du variogramme au palier. Ces deux valeurs ne sont *a priori* pas connues, mais leur rapport peut être déduit à partir de variogrammes ajustés sur des données du même type que celle recueillies.

15.5 Remarque sur l'explication des hétérogénéités

Lorsque dans une parcelle on observe une hétérogénéité de la répartition de la population d'un organisme auxiliaire ou d'un bioagresseur, on peut difficilement décider si elle est due à un écart à la moyenne (une tendance) ou à un écart à l'indépendance (une corrélation spatiale qui fait apparaitre une structuration, mais avec une moyenne constante). Ces deux approches différentes de modélisation n'impliquent pas forcément les mêmes choix de stratégie d'échantillonnage.

L'objectif est de prévoir des hétérogénéités à partir de connaissances disponibles *a priori*, et non de les expliquer. Si plusieurs échantillonnages de la même parcelle ont lieu successivement, on peut se baser sur les résultats précédents pour ajuster les observations selon les hétérogénéités observées, sans avoir besoin de les expliquer (en considérant qu'elles seront similaires, ce qui dépend de la dynamique de ce qui est observé). Le phénomène d'agrégation sur la parcelle ne sera pas étudié pour ajuster un modèle car cela demande un effort d'échantillonnage trop important. En revanche, des modèles existants d'agrégation pourront être utilisés pour optimiser la taille de l'échantillon.

15.6 Test des modèles de distributions discrètes ajustés

Une fois un des modèles sous forme de distribution pour des variables à valeurs discrètes ajusté, sa pertinence peut être testée par un test du χ^2 :

- On fait l'hypothèse que les observations, qui sont réparties avec $N\hat{p}_j$ observations dans la classe j, suivent une loi multinomiale telle que la probabilité associée à la classe j est p_j. La statistique $T = \sum_{j=1}^{J} \frac{(N\hat{p}_j - Np_j)^2}{Np_j}$ suis alors une loi du χ^2 à $J-1$ degrés de liberté pour N assez grand.
- On peut alors construire un test de niveau α en rejetant l'hypothèse nulle lorsque la statistique de test est plus grande que le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $J-1$ degrés de libertés : $T \geq \chi^2_{J-1,1-\alpha}$.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

15.7 Quantiles de la loi de Student

Tableau 2: Quantiles de la loi de Student

LOI DE STUDENT AVEC k DEGRÉS DE LIBERTÉ
QUANTILES D'ORDRE $1 - \alpha$

	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

16 Code R : Tracé des diagrammes pour l'échantillonnage séquentiel

```
# n : nombres d'observations
# Tn : nombres d'observations positives limite
n=1:100
#-----------------------------------------------------------
# Pour une loi de puissance de Taylor
a<-1
b<-1.5
# minimum pour avoir un coefficient de variation inférieur à cv
cv<-0.25
Tn<-(cv^2*(n^(b-1))/a)^(1/(b-2))
# maximum pour avoir une demi-largeur d'intervalle de confiance inférieure à d
d<-1
Tn<-(d^2*(n^(b+1))/(4*a))^(1/b)
#-----------------------------------------------------------
# Pour une loi binomiale négative
a<-1.39
# minimum pour avoir un coefficient de variation inférieur à cv
# valide pour n>1/cv^2*k
cv<-0.25
Tn<-n*k/(cv^2*n*k-1)
# maximum pour avoir un écart-type inférieur à s
s<-1
Tn<-k*n*(sqrt(1+n*s^2/k)-1)/2
#-----------------------------------------------------------
# Tracé (executer la ligne 'polygon...' adaptée à la situation de
# façon à griser la zone où l'échantillonnage doit être poursuivi)
type<-"l"
xlim<-c(min(n),max(n))
ylim<-c(0,100)
titre<-"Diagramme pour échantillonnage séquentiel"
xlab<-"Nombre d'observations"
ylab<-"Total individus comptés"
plot(n,Tn,type,xlim,ylim,"",titre,"",xlab,ylab)#définition du diagramme
# ajout de la zone "non-valide" sous la courbe
polygon(c(n[1],n,n[length(n)]),c(0,Tn,0),col="gray70",border = NA)
# ajout de la zone "non-valide" au dessus de la courbe
polygon(c(n[1],n,n[length(n)]),c(10000,Tn,10000),col="gray70",border = NA)
lines(n,Tn,type="l")#repassage de la frontière
axis(1,10*(1:60))#marques axe x
axis(2,10*(1:20))#marques axe y
abline(h=10*(1:20),v=5*(1:40),lty=2)#grille
```


Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

17 Échanges sur l'échantillonnage pour Rés0pest

Lors du séminaire Rés0pest en 25 juin 2015, des questions sur les stratégies d'échantillonnage du réseau ont été discutées. Cela a permis de mesurer l'importance de proposer un outil d'ide au choix des stratégies d'échantillonnage (travail de stage d'Eloi Navarro sous la direction de Jean-Noël Aubertot (UMR AGIR INRA-Auzeville)).

Remarque : Temps de travail important lors des notations adventices si le salissement est important

Lorsque la densité des adventices est élevée, leur répartition est plus homogène (moins de chance de tomber sur une zone «vide»), ce qui permet de choisir des quadrats de taille plus petite. Par contre, réduire le nombre de quadrats est à éviter, de façon à conserver une bonne idée la précision (variance).

Question : comment faire les moyennes de classes ? Particulièrement quand les classes correspondent à une gamme de valeurs allant d'une valeur jusqu'à l'infini ?

Plusieurs approches sont possibles :

- un score peut être attribué à chaque classe (en fonction, par exemple de la perte de rendement attendue), et la moyenne est faite sur ces scores (idée utilisée par Jean-Noël et Vincent apparemment, ils ont développé un outil pour déterminer les meilleures tailles d'échantillon dans ce cas https:/www6.inra.fr/quantipest/Tools-and-methods-for-sampling/Samplingstrategies/ How-to-define-the-sample-size/N-Index-Calculation-of-minimumsample- size-for-an-index). Dans le cas des essais Rés0pest, si la classe supérieure va jusqu'à l'infini c'est peut-être parce que d'un certain point de vue toutes les valeurs se valent dans cette classe.
- Soit une valeur de la variable qui sert à définir les classes est choisie pour représenter chaque classe, ça peut par exemple être la médiane (valeur au milieu de la classe). Dans le cas où il n'y a pas de milieu parce que la classe est infinie, pourquoi ne pas utiliser la moyenne (ou la médiane) qui est habituellement observée dans chacune des classes (il vaut mieux choisir une valeur définitivement si on veut pouvoir comparer les résultats sur différentes parcelles ensuite) ?

Question : Les mesures adventices sur 0.36m² sont discutables en termes de "représentativité"?

1- Un malherbologue aura certainement un avis sur l'hétérogénéité de la présence des adventices sur une parcelle.

2- Il faut voir sur des données existantes (issues des tests du protocole par exemple) si la variance est élevée entre les quadrats. Prendre des quadrats plus grands permet de faire une moyenne sur une surface plus grande et donc de lisser les résultats, mais quitte à observer plus de surface il vaut peut-être mieux le faire sur des quadrats distincts pour avoir un aperçu de la diversité sur la parcelle (à condition que le déplacement et le repérage des quadrats ne prennent pas trop de temps). Le nombre de quadrats peut aussi ne pas être décidé à l'avance, et si une variance élevée est observée des observations peuvent être ajoutées (échantillonnage adaptatif) jusqu'à ce que la précision obtenue soit satisfaisante (dans la limite du temps disponible bien sûr, mais en partant de peu de quadrats on peut aussi gagner du temps dans certain cas).

Question : « Si les comptages de peuplement effectués sur stations ne sont pas représentatifs du reste de la parcelle, faire des comptages supplémentaires dans d'autres zones de la parcelle ». Comment apprécier les différences (à partir de quel seuil) ? Combien d'échantillons supplémentaires ? Comment faire au stade floraison avec des plantes de plus de 2 m pour identifier les hétérogénéités ?

Au stade de la floraison, l'hétérogénéité ne pouvant pas être évaluée directement, on peut se baser sur les états antérieurs de la parcelle (quand elle était encore observable facilement) et sur une expertise des propriétés de ce qui est observé (type d'évolution, de dispersion,…) On peut aussi calculer en direct la variance de ce qui a été observé, en déduire un intervalle de confiance ou un coefficient de variation et juger si les valeurs obtenues sont satisfaisantes ou s'il faudrait ajouter des observations (cf. deux questions plus loin).

Question: pourquoi 10 plantes consécutives pour noter les maladies à floraison ?

Parce qu'il est souvent plus pratique de choisir un nombre de plantes consécutives que un nombre de plantes pris au hasard sur le rang. Ça peut aussi être utile pour avoir plusieurs valeurs de comptage/pourcentages avec lesquelles faire un traitement statistique. Dans l'absolu, il vaut mieux espacer les observations pour que l'agrégation spatiale de la maladie ne réduise pas trop la précision de la moyenne obtenue. Si la maladie diagnostiquée ne présente pas de phénomènes d'agrégation, alors on peut légitimement privilégier un protocole pratique avec des plantes consécutives. Le nombre de 10 a certainement été déterminé par expérience comme étant un bon compromis entre le temps nécessaire et la précision voulue.

Remarque :

Parfois, les protocoles proposés demandent d'enregistrer une note sous forme de classe ou de moyenne, mais sans conserver nécessairement le détail des comptages. Conserver les données issues de comptages permettrait de calculer un indicateur de précision à la fin des mesures et ainsi avoir une idée de la précision.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto

Question : 10 plantes prélevées pour réaliser la détermination du stade épi 1 cm sont-elles suffisantes en termes de représentativité ?

Il faudrait calculer l'écart-type observée en pratique sur la distance mesurée afin d'en déduire un intervalle de confiance à 95% approximatif pour la moyenne (IC =m+/-(2 x écart-type/racine (nb plantes)) (en 5 min avec une petite calculatrice). Ainsi, il est possible de décider d'examiner plus de plantes pour réduire l'intervalle de confiance (pas forcément besoin de recalculer l'écart-type sur les données, il ne devrait pas trop changer). Enfin, il est également possible de faire un test statistique pour savoir si on dépasse le seuil de 0,8 cm, mais cela est un peu plus compliqué.

Question : mesures maladies en foyers, combien de plantes ? Difficulté de cerner le périmètre ?

Au milieu du foyer, on peut supposer que les comptages de plantes malades seront assez homogènes. On ne peut pas tellement estimer la qualité statistique des comptages sur 3 répétitions donc il faut faire confiance à la capacité de l'observateur de choisir des rangs représentatifs du foyer.

Pour déterminer le périmètre il n'existe pas vraiment de méthode rapide. L'observateur peut, par exemple, parcourir la parcelle en essayant de déterminer la proportion de la distance (ou du temps) pendant laquelle il est dans des foyers. L'aspect subjectif de cette notation ne peut pas garantir la précision souhaitée.

Question : quelle est la représentativité de la valeur unique de la récolte ?

Pendant la récolte les grains sont mélangés donc ce qu'on observe est déjà une sorte de moyenne sur la parcelle. Si on pense que les grains ne sont pas bien mélangés, on peut toujours essayer d'en prendre à plusieurs moments/endroits dans le tas/silo. On pourrait tester la représentativité en répétant les mesures quelques fois pour une des parcelles, à conditions que le coût ne soit pas trop élevé.

Question : Lorqu'il s'agit de mesures des maladies et des ravageurs, est-il préférable de prendre un seul échantillon / station pour tous les bioagresseurs ou de prendre x échantillons pour x bioagresseurs ?

Il n'existe pas de raison statistique de prendre des échantillons différents pour observer les différents bioagresseurs, sauf si des observations sont destructrices et/ou incompatibles.

Action pilotée par le ministère chargé de l'agriculture, avec l'appui financier de l'Office national de l'eau et des milieux aquatiques, par les crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Ecophyto