École Centrale de Lyon
TFE 2015

INRA Toulouse, UMR AGIR
31326 Castanet-Tolosan

Rapport Final de Travail de Fin d'Études

Optimisation de stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes

NAVARRO Eloi

11 septembre 2015

Projet CASIMIR

Option : Mathématiques et Décision
Filière : Mathématiques et Ingénierie du Risque
Métier : Ingénieur Management des risques industriels et environnementaux

Tuteurs :	
ECL :	HELBERT Céline
INRA :	AUBERTOT Jean-Noël
	GOULARD Michel

Remerciements

Je tiens à remercier particulièrement Jean-Noël Aubertot et Michel Goulard qui m'ont encadré pendant ce stage pour leur accueil et leurs conseils précieux. Ce stage a été l'occasion pour moi de mettre en pratique mes connaissances en statistiques tout en apprenant beaucoup sur la protection des cultures et le monde de la recherche agronomique en général.
Je remercie Céline Helbert, ma tutrice pour l'École Centrale Lyon, pour avoir suivi le déroulement de mon stage pendant cinq mois.
Je remercie Morgane Froger, ingénieur en charge du projet CASIMIR pour nos échange fructueux sur mon travail.
Je remercie Aude Trichard, Stéphane Cordeau, Morgane Froger et Claire Lavigne pour m'avoir transmis des bases de données sur lesquelles j'ai pu travailler.
Je remercie François Brun pour m'avoir invité à une réunion d'échange autour de diverses questions d'échantillonnage.
Je remercie Caroline Colnenne, pour nos échanges sur les questions que se posent les expérimentateurs du programme Rés0pest concernant l'échantillonnage.
Je remercie Céline Colombet pour m'avoir présenté ses parcelles expérimentales et ses méthodes d'échantillonnage.
Je remercie Romain, Arthur et Olivia pour la bonne ambiance qui régnait dans notre bureau commun.
Je remercie enfin tous les stagiaires avec qui il a été très agréable de passer les pauses et les repas, et plus spécialement Antoine, Alexane et Aurélie qui m'ont invité à participer à leurs travaux expérimentaux.

Résumé du rapport

Alors que le plan ECOPHYTO du Ministère de l'Agriculture démontre depuis 2008 une certaine volonté politique de réduire l'usage des produits phytosanitaires en France, des progrès important restent à faire. Ils passent par une meilleure connaissances des agroécosystèmes et en particulier de leur composante biotique dont la caractérisation nécessite des protocoles adaptés, qui n'existent pas toujours.
Dans le cadre de l'élaboration de tels protocoles, ce rapport développe l'optimisation de stratégies d'échantillonnage dans le domaine de la protection des cultures, avec une approche essentiellement statistique. Dans un premier temps, le processus de conception d'une stratégie d'échantillonnage est explicitée dans le cadre très spécifique du stage, puis les notions de précision, de modèle et de plan d'échantillonnage sont approfondies. Ensuite, un outil d'aide à la conception de stratégies d'échantillonnage, qui repose sur les développements précédents, est proposé. Ils se compose de questions et de tâches à accomplir de façon à élaborer une stratégie d'échantillonnage de efficacement. Enfin, des cas particuliers d'échantillonnage sont présentés de façon à illustrer les notions évoquées.

Mots-clés libres : Échantillonnage ; Protection des cultures ; Agroécologie ; Optimisation ; Phytopathométrie ; Précision ; Coût ; INRA ; ECOPHYTO ;

Abstract

The ECOPHYTO plan from the French Ministry of Agriculture shows some sort of a political will to curb the use of phytosanitary treatments. The aim is hardly achieved and a better knowledge of agroecosystems appears to be necessary. It itself requires protocols, still not available, to characterise the biotic component of the agroecosystems.
In the context of the elaboration of such protocols, the report addresses the optimisation of sampling strategies in the field of crop protection with a mainly statistical approach. The process of designing a sampling strategy is first described. Then accuracy, modelling, and the spatial location of the observations are discussed more deeply. On that basis, a tool is proposed to support the design of sampling strategies. Last, the main notions of the report are exemplified through the analysis of data bases.

Keywords : Sampling ; Crop Protection ; Agroecology ; Optimisation ; Phytopatometry ; Accuracy ; Cost ; INRA ; ECOPHYTO

Table des matières

Table des figures

Liste des tableaux

1. Introduction

Mon stage se déroule au sein d'une Unité Mixte de Recherche de l'INRA [1] et de l'INPT [2], l'UMR AGIR (Agroécologie, Innovations, Territoires).

Suite au Grenelle de l'environnement, le Ministère de l'Agriculture et de la Pêche à lancé un plan baptisé ECOPHYTO dans l'objectif de réduire de moitié l'usage des produits phytosanitaires (intrants, pesticides, ...) en France sur une période de dix ans. Le plan ECOPHYTO n'a pas atteint cet objectif ambitieux puisque l'utilisation des produits phytosanitaires a augmenté depuis le lancement en 2008. Néanmoins des agriculteurs volontaires mettent en place des méthodes de gestion des bio-agressions utilisant moins de produits phytosanitaires dans 1900 fermes rassemblées au sein réseau FERMES du dispositif DEPHY. Un appel à projets Pour et Sur le Plan ECOPHYTO (PSPE) a été lancé, auquel a répondu le réseau PIC (Protection Intégrée de Cultures) dont font parti plusieurs membres de l'UMR AGIR.

Le projet piloté par le réseau PIC vise à proposer un soutien méthodologique quant à la caractérisation des évolutions des stress biotiques au sein du réseau FERMES ; il s'intitule CASIMIR pour *CAractérisation SIMplifiée des pressions bIotiques et des Régulation biologiques*. Plus précisément, les ingénieurs chargés du suivi des 1900 fermes doivent être pourvus de protocoles leur permettant, avec des ressources limitées, de rassembler des informations fiables sur les agroécosystèmes, et en particulier les stress biotiques à des fins de suivi et de recherche scientifique.

L'optimisation des stratégies d'échantillonnage participe de la bonne qualité des protocoles appliqués par les ingénieurs lors de leurs visites dans les parcelles agricoles. Après avoir explicité le contexte et la problématique agronomique abordée, ce rapport présente une synthèse sur les stratégies d'échantillonnage pour l'agronomie, il propose ensuite un outil d'optimisation des stratégies d'échantillonnage, enfin quelques exemples d'optimisation sont développés.

1.1. Contexte du stage

1.1.1. L'Unité Mixte de Recherche AGIR

L'UMR AGIR (Agroécologie [3], Innovations, Territoires) rassemble des chercheurs et enseignant-chercheurs en sciences biotechniques (agronomie, écophysiologie, écologie, épidémiologie végétale, modélisation) et en sciences sociales et humaines (sciences de gestion, économie, géographie sociale, sociologie). Ses travaux portent sur la compréhension et l'accompagnement de l'adaptation des agroécosystèmes, et des filières qui en dépendent, aux changements globaux de la

1. Institut National de la Recherche Agronomique
2. Institut National Polytechnique de Toulouse
3. D'une part, l'agroécologie est une science dédiée à l'étude des agroécosystèmes, c'est-à-dire des écosystèmes gérés par l'activité agricole dont on considère l'ensemble des composantes biotiques et non-biotiques, et au contexte social de l'activité agricole. D'autre part, elle décrit un mouvement social et un ensemble de pratiques agricoles

société et de l'environnement. Pour cela, des expérimentations en parcelle sont menées ainsi que des diagnostics régionaux en parcelles agricoles. Des enquêtes et des ateliers de co-conception participative avec les acteurs des filières agricoles sont aussi mobilisés. Enfin, plusieurs travaux de modélisation sont développés à la fois dans un objectif de synthèse des connaissances en agroécologie, et dans un objectif de test d'hypothèses et d'aide à la conception de systèmes de culture.
L'UMR AGIR est impliquée dans le projet CASIMIR, piloté par le réseau PIC dont Jean-Noël Aubertot est le coordinateur, par ce stage dédié aux aspects théoriques de l'échantillonnage mais aussi par des tests de protocoles d'observation menés par d'autres stagiaires.

1.1.2. Le plan ECOPHYTO

Depuis les année 1950, le secteur agricole français s'est orienté vers des systèmes de culture [Sebillotte, 1975] et d'élevage intensifs, recourant massivement à des produits phytosanitaires. Ce changement s'est fait en partie avec le double objectif de l'indépendance alimentaire et de la sécurité sanitaire des aliments, auxquels s'ajoutent des arguments économiques. Alors que les besoins en nourriture et en autres matières premières d'origine agricole continuent de croître, une nouvelle contrainte s'impose : réduire l'utilisation des pesticides pour limiter les risques sanitaires et l'impact négatif sur l'environnement et la biodiversité.
A la suite de Grenelle de l'environnement (2007), il a été décidé de réduire de moitié l'usage des pesticides en France sur une période de dix ans. Ce plan, dirigé par le Ministère de l'Agriculture et de la Pêche, a été baptisé ECOPHYTO. Il vise à promouvoir des méthodes de contrôle des stress biotiques subis par les cultures alternatives à l'utilisation des pesticides. Il s'appuie pour cela sur un réseau d'exploitations agricoles volontaires qui forment le réseau FERMES du dispositif DEPHY. Dans ce réseau de 1900 exploitations, des agriculteurs essaient de réduire leur consommation de pesticides. Pour cela, ils sont accompagnés par des ingénieurs qui doivent les conseiller. En complément de ce réseau, des expérimentations en vue de réduire fortement l'usage des produits phytosanitaires ont lieux sur 200 parcelles (réseau EXPE).
Le projet CASIMIR, financé par le plan ECOPHYTO, contribue à tirer profit au mieux de l'expérience acquise au sein des réseaux FERMES et EXPE.

1.1.3. Le projet CASIMIR

> *« Le projet CASIMIR (développements méthodologiques pour une CAractérisation SIMplifiée des pressIons biotiques et des Régulations biologiques) vise à étudier les régulations biologiques des bioagresseurs des cultures en grandes cultures, et à proposer, pour toutes les filières, une gamme de protocoles de niveaux de complexité variés, utilisables par les réseaux DEPHY-FERMES et DEPHY-EXPE, pour la caractérisation des états biotiques des parcelles DEPHY : niveau de pression des bioagresseurs, niveau de maîtrise, présence d'auxiliaires et niveau de régulation biologique. Ces protocoles doivent être testés à grande échelle en collaboration avec des groupes de fermes et des sites expérimentaux du réseau DEPHY, avant d'être proposés pour la définition d'une stratégie globale du réseau DEPHY sur ces questions. »*
>
> Ministère de l'Agricutlure et de la Pêche [2014]

En pratique, des protocoles d'observation pour les bioagresseurs, les auxiliaires et les régulations les plus importants sont testés sur diverses parcelles de façon à sélectionner les plus adaptés à chaque situation. Certains protocoles seront alors intégrés à des stratégies d'échantillonnage détaillées.
La conception de stratégies d'échantillonnage peut être optimisée de façon à en contrôler le coût et la précision, ce qui est l'objet du stage.

1.2. Problématique

L'objet du stage étant d'optimiser les stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystème, il est nécessaire d'expliciter la notion de composante biotique, qui rassemble ici les stress biotiques subis par les cultures, les organismes auxiliaires et les phénomènes de régulation biologique. Il faut ensuite expliciter l'enjeu de la caractérisation. Il faut enfin comprendre pourquoi les stratégies d'échantillonnage actuellement répandues dans le domaine de la protection des cultures ne sont pas, ou pas complètement, adaptées aux objectifs du projet CASIMIR.

1.2.1. Stress biotiques, auxiliaires et régulations biologiques

Types et mesures des stress biotiques Les stress biotiques subis par les cultures sont de trois types. Les plantes sont victimes de diverses maladies d'origine bactérienne, virale ou fongique. Elles sont aussi victimes d'attaques de ravageurs et parasites animaux (e.g. insectes, acariens, nématodes), elles sont enfin en concurrence avec des plantes non-cultivées qui sont qualifiées de plantes adventices.
Ces stress sont en général caractérisés par divers indicateurs d'intensité [Madden et al., 2006]. Tout d'abord, l'*incidence* est la proportion de plantes touchées par un stress ; cette définition de l'incidence est différente de celle utilisée en épidémiologie humaine (elle correspond à la prévalence en épidémiologie humaine).Ensuite, pour les maladies, la *sévérité* correspond à l'intensité de la maladie ; elle se mesure par exemple en pourcentage du feuillage atteint. Pour les ravageurs, on s'intéresse fréquemment à la *densité* et au nombre d'individus par plante, par branche ou par unité de surface. Enfin, pour les plantes adventices, on mesure principalement la densité ou la biomasse. Pour les comptages et la sévérité, on s'intéresse parfois à des valeurs conditionnelles, c'est-à-dire mesurées seulement sur les plantes touchées et pas sur les plantes intactes. Pour les différentes mesures, on utilise soit des échelles quantitatives, soit des échelles qualitatives.

Auxiliaires et régulations biologiques Les organismes auxiliaires sont des organismes qui contribuent à réduire un stress biotique ou plus généralement à améliorer le rendement. Leur présence peut être caractérisée de la même façon que celle des ravageurs. Les régulations biologiques considérées ici correspondent à une limitation des stress biotiques via les interactions entre auxiliaires et bioagresseurs.

Suivi et contrôle des stress Le suivi des stress biotiques est nécessaire soit pour l'aide à la décision de traitement, soit pour la recherche d'une meilleure connaissance des stress, de ses

causes et de ses conséquences, soit encore pour la mise au point de méthodes de contrôle culturales, chimiques, physiques, biologiques ou génétiques. Dans tous les cas, l'objectif final est de réduire l'écart entre le rendement accessible et le rendement réel : le *yield gap*. Ainsi, diagnostiquer un stress permet d'appliquer des traitements adaptés aux cultures au bon moment. Une quantification correcte des stress biotiques permet aussi d'évaluer l'efficacité d'un traitement ou d'une pratique culturale visant à le contrôler, elle participe à comprendre les causes et les interactions qui modulent les stress biotiques.
Pour un suivi efficace, on peut faire appel à différentes méthodes d'observation et d'échantillonnage. On trouve par exemple une documentation à ce sujet sur le site *Quantipest* [IPM Network, 2013], mis en place par un réseau de chercheurs dédié à la Protection Intégrée des Cultures (PIC) et diffusé au niveau européen via l'European Research Group ENDURE. Une approche statistique est alors intéressante pour disposer d'informations avec une indication de certitude et de fiabilité.

Quand les stress biotiques sont maintenus à un niveau raisonnable, c'est-à-dire qu'ils n'impliquent pas de pertes économiques notables, on dit que les cultures sont en situation de maîtrise phytosanitaire.

1.2.2. Stratégies d'échantillonnage

L'échantillonnage vise à obtenir une connaissance globale sur une population alors qu'il n'est pas envisageable d'en observer tous les éléments, une partie seulement des éléments est donc examinée de façon à en déduire du mieux possible l'information recherchée. Dans le cadre du projet CASIMIR, la population concernée est restreinte à une parcelle agricole, l'objectif de l'échantillonnage étant de caractériser des composantes biotiques à l'échelle de la parcelle.
Une stratégie d'échantillonnage se compose d'un protocole d'examen pour chaque élément de la population totale observée, et d'une procédure de choix d'éléments qui inclut leur nombre total (la taille d'échantillon) et leur répartition dans le temps et l'espace. La précision des résultats obtenus, ainsi que le coût de mise en œuvre de la stratégie, sont les deux critères selon lesquels les stratégies d'échantillonnage sont optimisées.

Outre les nombreux protocoles expérimentaux existants, diverses stratégies d'échantillonnage ont été développées pour caractériser les stress biotiques. Les protocoles utilisés pour produire les Bulletins de Santé du Végétal, dont les protocoles Vigicultures®, en sont des exemples intéressants.

1.2.3. Besoins et contraintes spécifiques du projet CASIMIR

Les stratégies d'échantillonnage développées pour le projet CASIMIR doivent permettre de faire un état des lieux des stress biotiques et des régulations biologiques sur une parcelle de façon à refléter l'effet de ces phénomènes sur le rendement. Cela implique d'avoir une estimation de la moyenne du phénomène sur la parcelle au moment où son effet est le plus visible. L'objectif est par conséquent différent de celui des Bulletins de Santé du Végétal qui sont plutôt destinés à la détection et à la prévision du risque.

D'autre part, beaucoup de protocoles existants sont utilisés et réutilisés sans que leur pertinence n'ait été validée, c'est-à-dire que la précision n'est pas contrôlée et que le coût n'est pas optimisé. On souhaiterait que les stratégies développées pour le projet CASIMIR évitent ces deux écueils, en particulier en écartant des protocoles dont la performance n'a jamais été formellement analysée.

Enfin, le suivi des parcelles des réseaux FERMES et EXPE est soumis à un certain nombre de contraintes en termes de compétences et de temps disponibles. Pour qu'elles ne deviennent pas un frein à une caractérisation efficace des stress biotiques et des régulations biologiques, les stratégies d'échantillonnage proposées doivent être suffisamment simples et flexibles.

1.3. Objectifs du stage

L'objectif initial du stage était la mise à disposition d'un ensemble d'outils permettant l'optimisation de stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes. Pour cela, après avoir élaboré une typologie de développements spatio-temporels et de variables descriptives pour les diverses composantes biotiques, il était prévu de recenser les outils mathématique concourant à la conception de stratégies d'échantillonnage efficaces. Ces outils auraient alors été complétés si nécessaire. La première phase d'étude bibliographique, et les premières synthèses sur la mise en place d'une stratégie d'échantillonnage, ont ensuite fait ressortir le besoin de rassembler et d'articuler toutes les connaissances recueillies dans un unique outil d'aide à la conception de stratégies d'échantillonnage.

Au fil des échanges avec les encadrants et les membres du projet CASIMIR, les contraintes suivantes sont ressorties concernant les stratégies d'échantillonnage proposées :

- L'objectif des caractérisations est plutôt d'établir une intensité moyenne des stress biotiques ou une densité moyenne d'auxiliaires, une étude spatialisée approfondie est exclue pour le réseau FERMES mais envisageable pour le réseau EXPE.
- Les optimisations concernent un échantillonnage à l'échelle de la parcelle.
- Les justifications théoriques des recommandations faites devront être à disposition, au moins en annexe du rapport de stage.
- Il est préférable de ne pas faire appel à des outils de traitement informatique pendant l'échantillonnage (sur le terrain).
- Les optimisations proposées doivent êtres adaptées aux utilisateurs en termes d'attentes et de complexité.

D'autre part, de façon à ce que l'outil d'aide à la conception de stratégies d'échantillonnage soit adapté au besoin, les orientations suivantes ont été établies :

- L'informatisation de l'outil doit être envisagée pour la suite, et donc facilitée autant que possible par le travail actuel.
- L'utilisation de l'outil doit permettre de préciser les choix concernant taille d'échantillon et la répartition des observations dans l'espace.
- L'outil doit être générique, mais accompagné d'illustration sur des cas pratiques pour mettre en valeur son utilité.

1.4. Démarche

Le stage a débuté par des recherches assez générales sur l'échantillonnage, notamment par le biais du site *Quantipest* [IPM Network, 2013]. Ces recherches ont permis de préciser le vocabulaire utilisé pour l'échantillonnage en agronomie et en écologie, ainsi que les types de questionnement habituels pour la mise en place de stratégies d'échantillonnage (taille d'échantillon, plan, usage de modèles). Cette première étape a abouti à la rédaction d'un rapport bibliographique.

Une synthèse de tous les déterminants d'une stratégie d'échantillonnage a ensuite été réalisée, et les nouvelles recherches bibliographiques ont été restreintes aux besoins précis du projet. Les déterminants qui peuvent être traités de manière généraliste (puisque le travail ne porte sur aucun phénomène à échantillonner en particulier mais sur un large gamme de situations possibles) ont fait l'objet d'une étude plus approfondie. Il a alors été possible d'expliquer comment prendre en compte les déterminants pour décider de la taille adaptée pour un échantillon et de la meilleure répartition spatiale des observations.

Les étapes de décision et les connaissances synthétisées ont ensuite été organisées en un outil d'aide à la conception de stratégies d'échantillonnage en s'inspirant des outils présents dans la littérature. Cet outil prend la forme d'un suite de questions et de tâches pour lesquelles il est parfois nécessaire de consulter les détails rédigés dans la partie théorique du rapport. Les étapes et questions sont incluses dans le rapport, de façon à ce que les chapitres 2 et 3 forment un outil d'aide à la conception de stratégies d'échantillonnage utilisable dès maintenant.

En parallèle du travail décrit, les pistes d'optimisation proposées ont été confrontées à des situations réelles par le biais d'études de bases de données issues d'échantillonnages avec un grand nombre d'observations.

2. État des lieux des stratégies d'échantillonnage pour la caractérisation de la composante biotique des agroécosystèmes

Ce chapitre fait la synthèse d'un large éventail de considérations sur l'échantillonnage en agronomie et en écologie issues de la littérature et complétées par des échanges avec des chercheurs. Elle est à la fois un manuel pour la mise en place de stratégies d'échantillonnage et la base théorique de l'outil d'aide à la conception de stratégies d'échantillonnage proposé au chapitre suivant. Pour certains points l'outil fait d'ailleurs directement appel aux passages rédigés ici. Ce chapitre présente quelques redondances de façon à ce qu'il puisse être lu dans le désordre ou partiellement, en fonction des besoins des utilisateurs de l'outil d'aide à la conception de stratégies d'échantillonnage.
Dans un premier temps, la démarche générale de mise en place d'une stratégie d'échantillonnage pour la caractérisation des composantes biotiques à l'échelle d'une parcelle est présentée. Plusieurs aspects de cette démarche sont ensuite approfondis en se restreignant à ce qui est nécessaire et utilisable pour la mise en place d'un outil d'aide à la conception de stratégies d'échantillonnage généraliste dans la partie suivante.

2.1. Élaboration d'une stratégie d'échantillonnage

Les informations concernant l'échantillonnage pour la caractérisation des stress biotiques à l'échelle d'une parcelle rencontrées dans la littérature sont nombreuses. Il a donc semblé pertinent de les synthétiser sous la forme d'un schéma (figure 2.1) qui montre l'établissement d'un protocole d'échantillonnage à partir des objectifs, du contexte et des connaissances disponibles, le tout dans le cadre fixé par le projet (stress biotiques et régulations biologiques à l'échelle d'une parcelle). Le schéma est accompagné de précisions destinées à lever toute ambiguïté sur les termes utilisés et à donner un aperçu des liens entre eux.

Le schéma pose un cadre de réflexion ; s'il met en évidence les dépendances qui peuvent exister, il n'apporte en revanche aucune indication sur le protocole d'échantillonnage adapté à une situation donnée. En effet, les relations présentées sur le schéma sont parfois complexes, elles nécessitent donc des explications complémentaires conséquentes pour qu'il deviennent un outil d'aide à la conception de stratégies d'échantillonnage. Comme l'explique de Gruijter et al. [2006, page 77], rassembler le choix d'une stratégie d'échantillonnage dans un seul schéma est bien trop complexe. Il propose lui même des arbres de décision simplifiés pour suggérer des plans d'échantillonnages adaptés en fonction du contexte et des priorités de l'observateur (coût ou précision).

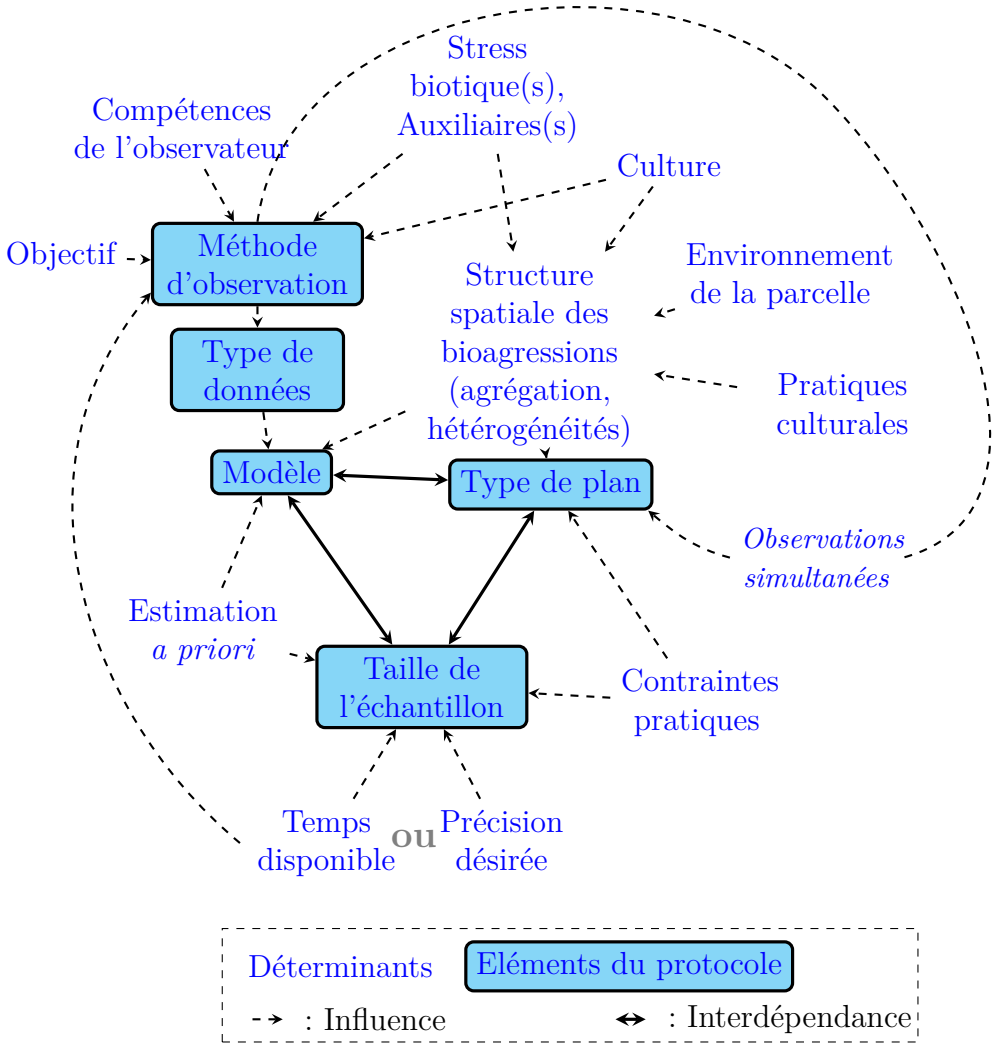

FIGURE 2.1.: Schéma synthétique des déterminants d'une stratégie d'échantillonnage pour la caractérisation de la composante biotique d'un agroécosystème

2.1.1. Définitions

On précise dans les paragraphes suivants les définitions utilisées car certains termes sont ambigus. Ces définitions sont cohérentes avec celles rencontrées dans la littérature et sont valables pour la suite de ce travail.

Stress biotiques et régulateurs Les stress biotiques résultent d'un ensemble de phénomènes ayant pour origine un organisme vivant et qui réduisent le rendement d'une culture ou nuisent à sa santé. Les auxiliaires sont des organismes qui limitent les stress biotiques.
Du fait de leur diversité (plantes adventices, insectes, agents pathogènes), les stress biotiques et les auxiliaires sont caractérisés de différentes manières et peuvent présenter des motifs spatiaux variés.

Culture Le type de culture (grande culture, verger, serres,...) impose une structuration naturelle à une parcelle (voir 2.7.1).

Pratiques culturales Ce sont l'ensemble des pratiques agricoles telles que le travail du sol, le choix de l'espèce cultivée et de la variété, le choix de la date et de la densité de semis, l'irrigation. Elles peuvent créer des hétérogénéités aux formes variées dans les parcelles (voir 2.7.3).

Environnement de la parcelle Les haies, les bordures, les forêts, les cours d'eau, les parcelles voisines forment un environnement qui influence les stress biotiques et les régulations dans la parcelle étudiée (voir 2.7.2).

Compétences de l'observateur Les connaissances et l'expérience de l'observateur influent sur la justesse, la fiabilité et la rapidité des observations, voire sur la capacité à mettre en œuvre un certain protocole.

Objectif On se restreint aux objectifs qui se ramènent à déterminer l'intensité d'un stress biotique ou d'un phénomène de régulation sur une parcelle (voir 2.2). L'intensité est mesurée par une variable cible à déterminer (incidence, sévérité, sévérité conditionnelle, densité, densité conditionnelle).
Selon que l'on veuille obtenir une appréciation, une appréciation ordonnée, une note, une mesure sur une échelle, une sévérité, une densité, les modèles utilisés et la façon d'appréhender les questions ne seront pas les mêmes (voir 2.4).

Observations simultanées Si plusieurs stress sont étudiés simultanément, il peut être préférable de réaliser les observations lors d'un unique parcours de la parcelle.

Méthode d'observation Des exemples de méthodes d'observations sont, la détermination du pourcentage du feuillage endommagé, le comptage de parasites, la classification des plantes saines et malades. Une variable cible ayant été déterminée, la méthode d'observation correspond au détail du protocole à l'échelle de l'unité d'observation (plante, feuille, rameau, unité de surface). La méthode d'observation va définir quel type de variable va être obtenu, quelle plage

de valeurs elle peut prendre. Les méthodes d'observation n'ont pas toutes la même fiabilité ; un biais ou une forte variabilité sur le résultat des mesures se traduira par un résultat final de qualité réduite quelle que soit la stratégie d'échantillonnage choisie (voir 2.3).

Type de données Les valeurs obtenues lors de chaque observation peuvent être de différent types. Elles peuvent être quantitatives (discrètes, continues, bornées ou pas) ou qualitatives (à deux modalités ou plus, ordinales ou pas).
Le type de données est à la base du choix d'un modèle (voir 2.4).

Estimation *a priori* Un fourchette de valeurs pour le paramètre à estimer est souvent nécessaire si on ne veut pas passer par un échantillonnage adaptatif. En effet, les modèles permettent en général de déduire une variance, et donc une précision attendue, à partir d'une moyenne.

Temps disponible Le temps disponible pour mener à bien les observations sur la parcelle est à associer à l'estimation du temps nécessaire par observation. On pose de cette façon une contrainte sur la taille maximale de l'échantillon et sur le trajet à parcourir dans la parcelle. On pourra faire appel à des modèles simples pour estimer le temps nécessaire pour un certain nombre d'observations (voir 2.5.1).

Précision désirée La précision reflètera la qualité des résultats obtenus, elle peut être exprimée en termes de coefficient de variation, de variance, d'intervalle de confiance, de taux d'erreurs de classification acceptable. On essaie d'en déduire une taille minimale pour l'échantillon (voir 2.4.2).

Contraintes pratiques Les contraintes pratiques seront par exemple la difficulté de parcours de la parcelle, l'impossibilité d'abimer les cultures, la forme de la parcelle qui complique le repérage [de Gruijter et al., 2006, page 36].

Structure (voir 2.7) Deux informations sont importantes quant à la structuration spatiale d'un stress biotique, d'une part les sources connues d'hétérogénéité, d'autre part le niveau d'agrégation ou de corrélation spatiale qui dépend à la fois du stress ou de l'auxiliaire étudié et de son environnement.

Hétérogénéités dues à l'environnement : Les propriétés de la parcelle étudiée en termes d'ensoleillement, de géologie, de proximité d'une haie/rivière/parcelle différente créent des gradients à l'échelle de la parcelle ou des zones de forme correspondant à celles des variations d'environnement présentes.

Hétérogénéités dues aux pratiques culturales : Les labours, semis, cultures antérieures, traitements phytosanitaires, créent des zones où l'intensité du stress biotique étudié est différente. Ces zones, souvent bien délimitées sont plutôt en forme de bandes, ou au moins bien délimitées (par exemple, un passage plus fertilisé qui serait envahi de pucerons).

Agrégation : Le mode de propagation d'une maladie (spores dispersées par le vent par exemple), la mobilité d'un insecte, la nature pluriannuelle ou pas des plantes adventices sont autant

de facteurs d'agrégation ou de corrélation. Ces différentes répartitions possibles dans l'espace peuvent être prises en compte dans le choix d'un modèle qui permettra ensuite de dimensionner au mieux l'échantillon.

Taille de l'échantillon (voir 2.5) La taille de l'échantillon est le nombre d'observations qui seront réalisées, cela peut correspondre à une surface observée, un nombre de plantes, un nombre de pièges, etc.
Elle peut être fixée dès le départ, ou bien au cours de l'échantillonnage (on parle d'échantillonnage adaptatif). Elle peut être fixée selon une contrainte de temps, selon une contrainte de précision ou comme un compromis des deux. On ne peut en général associer une précision à une taille d'échantillon que si un modèle suffisamment élaboré existe pour les données recueillies, toutefois on peut souvent indiquer de quelle façon le nombre et la disposition des observations influeront sur la précision.

Plan d'échantillonnage (voir 2.8) Il caractérise la répartition spatiale des observations (aléatoire, périodique, stratifiée, par grappes, sur des coupes,...). Il peut influencer la possibilité d'utiliser un modèle ou non.
Un choix pertinent permet d'augmenter la précision relativement au coût, pour cela il est important de bien prendre en comptes les disparités spatiales, connues ou supposées, de la variable observée. Pour être mis en place convenablement, il peut falloir un échantillon plus grand que prévu par modélisation.

Modèle (voir 2.6) Il peut être très simple et supposer l'existence d'une espérance et d'une variance pour la variable observée, ou plus élaboré et se présenter sous la forme d'une loi de probabilité ou d'une relation fonctionnelle entre plusieurs grandeurs statistiques (par exemple entre variance et moyenne).
L'objectif de la modélisation est soit d'établir une estimation de la variance qui va être observée, soit de mieux exploiter les données recueillies. On déduit de la variance la taille d'échantillon adaptée à la précision désirée, ou bien la précision qui peut être obtenue avec un effort d'échantillonnage donné.
L'usage d'un modèle restreint les plans d'échantillonnage possibles, de plus il peut nécessiter un nombre minimal d'observations pour être pertinent.

2.2. Influence de l'objectif de l'échantillonnage

La définition claire des objectifs d'un échantillonnage est importante pour éviter de récolter des données inutilisables. En plus de 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]. Il faut aussi choisir le niveau de précision ou de certitude souhaité pour ce paramètre.Il faut enfin établir quel coût est acceptable (en particulier en temps passé). Ainsi, la stratégie d'échantillonnage dépend des informations que l'on souhaite en tirer et des contraintes pratiques. Les trois situations suivantes en sont l'exemple.

Diagnostic et informations qualitatives Pour réaliser un diagnostic (d'infestation par exemple), 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 (échantillonnage simple [Moura et al., 2007] ou parcours d'une diagonales par exemple) 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]). Plus généralement, un diagnostic passera par la détermination de 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.
Il ne faut pas oublier de décider si on souhaite avoir une information sur la parcelle dans son environnement ou bien sur le système de culture indépendamment de l'environnement de la parcelle. Dans un cas les bords de la parcelle sont à surveiller spécialement, dans l'autre ils seraient plutôt à éviter.

Études de variabilités et corrélations Si on veut obtenir des informations sur la variabilité et les corrélations, que ce soit en espace ou en temps, on doit veiller à disposer d'une bonne quantité d'échantillons pour une certaine résolution spatiale ou temporelle. C'est un problème qui peut s'avérer difficile dans l'espace si il y a des contraintes sur la zone à considérer. Deux choix judicieux peuvent alors être des plans d'échantillonnage selon un quadrillage [Pulakkatu-Thodi, 2014] (figure 2.2) ou selon des coupes [Parker et al., 1997, Clark et al., 2007].

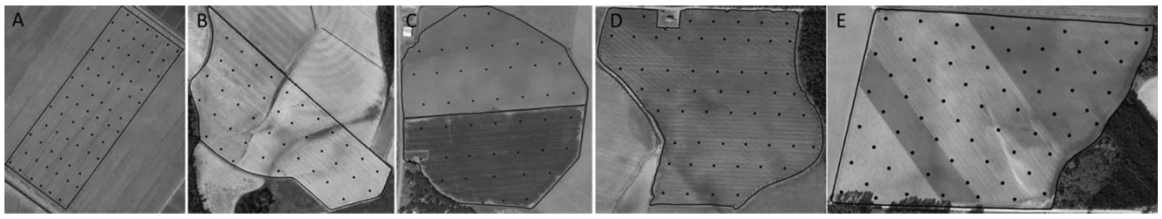

FIGURE 2.2.: Exemple de plans d'échantillonnage pour une étude de variabilité spatiale, l'échantillonnage est systématique et adapté à la forme des parcelles étudiées - *Source : Pulakkatu-Thodi [2014]*

Cartographie Pour réaliser une cartographie, dans l'objectif de traiter seulement les zones le nécessitant (agriculture de précision) ou de faire un suivi spatio-temporel, un échantillonnage régulier dans l'espace est préférable. Les prélèvements sont donc disposés sur un quadrillage (exemples : Aubertot et al. [2004], Barroso et al. [2005]). Il est important de souligner que l'échelle du quadrillage et celle du phénomène étudié ou du traitement appliqué doivent être compatibles.

Pour le projet CASIMIR Dans le cas qui nous intéresse, ce sont des informations de type *diagnostic* qui sont recherchées. Cela restreint automatiquement les méthodes d'observations, les types de variables, et les stratégies abordées dans la suite du travail.

2.3. Méthodes d'observation

La conception du plan d'échantillonnage ne doit pas être dissociée du protocole utilisé pour examiner chaque plante, chaque arbre ou chaque quadrat, il influence en effet la précision et peut rendre les observations incompatibles avec un modèle élaboré grâce à un autre protocole. De même, dans le cas d'une évaluation de sévérité, l'échelle choisie influe sur la précision et peut poser des problèmes de comparabilité avec d'autres études.

2.3.1. Caractériser la qualité des méthodes d'observations

Plusieurs critères décrivent la qualité d'une méthode d'observation. La répétabilité et la reproductibilité sont caractéristiques d'une méthode fiable. Pour certains stress biotiques, des études comparatives de différentes méthodes d'observation ont été réalisées [Nutter, Jr., 1993, Guan and Nutter, 2003, Aubertot et al., 2004].

Biais Le premier est le biais potentiel par rapport à une méthode d'observation de référence (réputée fiable), s'il est stable il pourra faire l'objet d'une correction, sinon il nuira fortement à la qualité des résultats.

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 des besoins de sa mise en œuvre.

Résolution La résolution est la l'écart limite en dessous duquel la méthode d'observation ne permet pas de différencier deux situations. Elle est importante si on étudie un dépassement de seuil.

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 pour des méthodes simples ou automatisées.

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

Au contraire du biais, la répétabilité et la reproductibilité sont indispensables si on souhaite modéliser sans trop de difficultés les données. D'autre part, il ne semble pas pertinent d'augmenter la taille de l'échantillon dans l'objectif d'obtenir une erreur faible si cette erreur est plus petite que la variation entre des répétitions des mêmes mesures (par un même observateur ou par des observateurs différents).

2.3.2. Influence du temps disponible

Le temps disponible, en plus de jouer sur la taille de l'échantillon, a une forte influence sur le choix du protocole d'observation. Les protocoles ne demandant pas beaucoup de temps seront favorisés. Il faut bien faire attention cependant à la précision perdue par des protocoles potentiellement moins fiables.
Par exemple, si le choix du protocole 2 par rapport au protocole 1 multiplie la variance sur les observations par 2 (si tant est qu'on puisse comparer les valeurs fournies par les deux protocoles) alors la taille d'échantillon nécessaire pour obtenir la même précision (en terme d'écart-type) qu'avec le protocole 1 est $\sqrt{2}$ fois plus importante (en faisant l'hypothèse d'observations indépendantes). Pour que le protocole 2 soit vraiment plus efficace, il faut donc qu'il divise le temps moyen par observation par au moins $\sqrt{2}$.
Ce problème illustre l'avantage d'utiliser des protocoles d'observation bien documentés, c'est à dire des protocoles dont on connait la fiabilité et le temps de mise en œuvre. En cas d'hésitation entre deux protocoles, des stratégies d'échantillonnages adaptées à chacun pourront être conçues, puis comparées en fonction du coût et de la précision prévus.

2.3.3. Positionnement dans le temps

On sait que les ingénieurs du dispositif FERME n'auront pas les moyens de surveiller semaine par semaine les cultures pour faire les mesures clés au bon moment. La meilleure option pour le diagnostic est alors d'échantillonner au moment ou le stress biotique, ou la présence d'auxiliaires sont les plus visibles. Pour cela, la dynamique connue des phénomènes observés, la surveillance sommaire par les agriculteurs ou toute autre source d'informations pourront être mises à profit.

2.4. Types de données et précision associée

La variable d'intérêt et la méthode d'observation associée déterminent le type de données obtenu. Pour chaque type de données, divers modèles de distribution, modèles de variance (voir 2.6) et quantificateurs de précision (voir 2.4.2) existent. Les modèles proposés ne sont pas exhaustifs, ce sont ceux rencontrés dans la littérature dans des études en agronomie ou en écologie. La traduction formelle de l'objectif de l'échantillonnage en une question concernant les données est déterminante. En effet, on n'a pas besoin du même nombre d'observations pour comparer une variable à un seuil que pour déterminer sa valeur à $\pm 25\%$.

2.4.1. Formalisme

On considère qu'une stratégie d'échantillonnage se compose de N observations, ou grappes d'observations, qui aboutissent à N valeurs $X_1, \ldots, X_N$. La suite de cette partie détaille le genre de valeurs sur lesquelles un travail a été réalisé. L'utilisation éventuelle de modèles, ainsi que le choix d'un indicateur de précision, dépend du paramètre qui sera calculé à partir de ces valeurs.

Paramètre d'intérêt L'objectif de l'échantillonnage se traduit par une question sur un paramètre des données recueillies (par exemple : *Quelle est l'incidence moyenne de tel bioagresseur*

sur la parcelle ?). Pour un diagnostic, les données sont essentiellement traitées pour déterminer une moyenne, une répartition entre classes, ou si un seuil est dépassé. Les modèles et les indicateurs de précision proposés dans la suite sont donc restreints à ces trois cas.
Si on souhaite estimer la moyenne m de la variable mesurée sur la parcelle, (*l'estimateur*, voir A.2) est

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

Si on souhaite estimer la proportion p_i d'individus dans la i-ème classe d'un ensemble de k classes, l'estimateur est

$$\hat{p_i} = \frac{N_i}{N} \qquad \text{où } N_i \text{ est le nombre d'observations dans la classe } i$$

2.4.2. Indicateurs de précision

Si une stratégie d'échantillonnage garantit une bonne répartition des observations dans la parcelle, les résultats obtenus seront *non-biaisés*, c'est-à-dire justes en moyenne (sur des répétitions de la stratégie). Il reste alors à 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 observé et des objectifs de l'échantillonnage.

Variance Lorsque sommer des valeurs observées a un sens, la variance sur l'échantillon observé est donnée par la formule

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

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

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

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 équivalentes.
L'estimation précise de la variance demande une grande taille d'échantillon, on se satisfait donc en général d'une estimation de la variance médiocre (voir A.3).

Écart-type Un indicateur classique de qualité pour une moyenne sur l'échantillon est l'écart-type, c'est-à-dire la racine carrée de la variance qui vaut $\frac{s}{\sqrt{N}}$ si s est l'écart type observé sur les éléments échantillonnés et N est la taille de l'échantillon.
La valeur de l'écart-type, dans la même unité que la variable étudiée, donne un ordre de grandeur de l'incertitude associée au paramètre calculé. Des intervalles de confiance peuvent en être déduits.

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

Coefficient de variation Le coefficient de variation (noté c_v dans la suite) pour un estimateur est son écart-type relatif, soit en pratique la valeur de l'écart-type pour l'estimation divisé par l'estimation : $\frac{s}{\hat{m}\sqrt{N}}$. Pour des comptages d'insectes, l'objectif en terme de précision est souvent un coefficient de variation de 25% ou 30% (par exemple dans Navarro-Campos et al. [2012]). Le coefficient de variation a deux avantages par rapport à l'écart-type : il est adimensionné ce qui permet de comparer des mesures dans différentes grandeurs physiques, et il explicite bien la dispersion des données par rapport à la moyenne. En revanche il perds de son utilité pour une moyenne proche de zéro, il prend alors des valeurs très grandes qui portent à confusion.

Probabilité d'erreur de classification Dans le cas où 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 2.3 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.

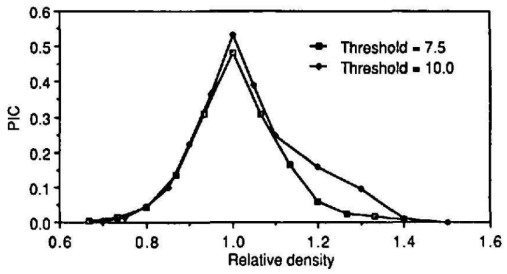

FIGURE 2.3.: Probabilité d'erreur de classification (PIC pour Probability of Incorrect Classification) en fonction du rapport entre la valeur réelle d'une variable et le seuil définissant la classification - *Source : Nyrop et al. [1989]*

2.4.3. Types de données retenus

Au vu des publications rencontrées et des protocoles existants (Rés0Pest, Vigicultures®), quatre types de données ont été étudiés. Ils sont sensés englober tous les choix de méthodes d'observation.

2.4.3.1. Valeurs continues bornées ou non

Des valeurs continues peuvent par exemple être obtenues par des mesures de sévérité sur feuille, de taille, de biomasse. Elles peuvent aussi provenir de mesures d'incidence exprimées en

pourcentage, même si dans ce cas on préfère en général modéliser les résultats de comptage et non le pourcentage obtenu [Madden and Hughes, 1999]. On s'intéresse ici à la moyenne sur la parcelle de la variable observée.

Modèles Sous certaines hypothèses, l'estimateur de la moyenne $\hat{m}$ suit une loi normale ou une loi de Student. Si c'est le cas, on peut construire des intervalles de confiance pour la moyenne. La corrélation spatiale de la variable observée peut fausser l'estimation de l'écart-type, un écart-type corrigé peut alors être envisagé.

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), des tests basés sur la loi normale ou la loi de Student permettent de calculer une probabilité d'erreur.

Remarque Les méthodes évoquées (sauf celles impliquant le loi de Student) peuvent aussi être appliquées à des variables numériques à valeurs discrètes, toutefois il faudra plus d'observations pour que $\hat{m}$ suive la loi normale (voir 2.6).

2.4.3.2. Valeurs entières bornées

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.
Il peut être intéressant d'adapter un plan d'échantillonnage pour pouvoir utiliser les modèles disponibles. En effet, ils permettent de rendre compte de la sur-dispersion causée par les phénomènes d'agrégation et de corrélation spatiales.

Modèles Dans le cas où les observations à l'intérieur de chaque grappe sont indépendantes les unes des autres, la loi binomiale devrait bien modéliser les valeurs observées. Dans les autres cas, une sur-dispersion est en général observée. Elle peut être modélisée par la loi bêta-binomiale ou une loi de puissance sur la variance [Madden and Hughes, 1999]. Ces modèles permettent de mieux évaluer la variance de $\hat{m}$, voire de la prévoir.

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 approximative avec la loi normale ou de manière plus fine avec un des modèles évoqués.

Remarque Pour les mesures d'incidence, l'écart-type en pourcentage 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.

2.4.3.3. Valeurs entières non-bornées

Des valeurs entières non-bornées 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. Ces valeurs ne sont pas bornées en théorie seulement, en pratique elles sont bornées par des contraintes physiques diverses. L'agrégation spatiale observable sur ce type de variable a été largement étudiée en écologie, elle se traduit par une sur-dispersion sur les moyennes estimées qui est prise en compte par certains modèles.

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

Précision selon la question Les indicateurs de précision et le calcul ces intervalles de confiance sont calculés comme pour les variables à valeurs entières bornées.

2.4.3.4. Classes

Ici le résultat de chaque observation est une classe (déterminée parmi un ensemble fini de k classes). Ce peut être par exemple des notations de sévérité sur une échelle ordonnée (cas de Aubertot et al. [2004]). Comme le signale [Madden et al., 2006, page 20], une variance (ou une moyenne) est parfois calculée abusivement sur ce type de données prises comme numériques. Pour éviter cette confusion entre variables ordinales et variables numériques, il vaut mieux faire ces calculs après une conversion pertinente (exemple : on attribue à chaque classe un score qui peut être la valeur moyenne qui y est attendue).

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

Précision selon la question Dans le cas où on souhaite estimer la proportion d'observations dans chaque classe, on peut souhaiter une précision sous forme d'intervalles de confiance.
Si, comme dans l'article de Aubertot et al. [2004], un score est affecté à chaque classe, on retrouve une variable du style « à valeurs continue »mais sans que les modèles précédents ne soient appropriés : on peut toutefois déduire un coefficient de variation ou un intervalle de confiance à partir des propriétés de la loi multinomiale (qui repose quand même sur l'indépendance des observations, qui n'est pas toujours garantie).

Remarque Si on fait directement une notation en classe sur toute la parcelle (comme sur les protocoles Vigicultures® et Rés0Pest parfois) alors on ne peut pas vraiment parler d'échantillonnage et la fiabilité du résultat dépend surtout de la compétence de l'observateur.

2.5. Taille de l'échantillon

La taille de l'échantillon est déterminée en fonction de l'objectif qui a été choisi en terme de précision et des différentes contraintes de coût (temps, matériel). Des information disponibles avant l'échantillonnage peuvent être mises à profit pour que la taille de l'échantillon soit un bon compromis entre coût et précision.
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.

2.5.1. Détermination par le coût d'échantillonnage

Le temps disponible pour examiner une parcelle et diverses autres contraintes matérielles, limitent directement le nombre d'observations possibles. On pourra parfois prévoir la précision obtenue grâce à ce nombre maximal d'observations, et ainsi juger de la qualité de la stratégie d'échantillonnage. Le temps nécessaire pour se déplacer dans les parcelles (fonction du type de culture, du matériel à transporter) est aussi important, il est pris en compte pour déterminer la répartition optimale des observations dans la parcelle. Des modèles pour faire le lien entre un plan d'échantillonnage, une taille d'échantillon et le coût associé sont peu génériques, le temps passé pour chaque observation dépend en effet de la compétence de l'observateur, des conditions météo, de l'état de la culture, etc. Il sera donc plus raisonnable de laisser au concepteur de la stratégie d'échantillonnage le soin de déterminer quels déplacements et quel nombre d'observations sont acceptables en terme de coût.

2.5.2. Détermination par la contrainte de précision

La taille de l'échantillon a une très forte influence sur la précision obtenue. Pour obtenir la précision voulue, on peut soit se reposer sur une connaissance *a priori* de la variable à observer et en déduire avant l'échantillonnage une taille optimale, soit opter pour un échantillonnage adaptatif, c'est-à-dire adapter la taille finale de l'échantillon en fonction des information recueillies pendant l'échantillonnage. Dans la plupart des cas, la prévision de la précision passe par la détermination de la variance pour la variable étudiée. Cette variance peut être connue à l'avance (grâce à l'expérience acquise lors d'échantillonnage précédents), elle peut être déduite d'un modèle et d'hypothèses sur la moyenne qui sera observée, elle peut enfin être déterminée en cours d'échantillonnage, via une méthode adaptative.
La distribution des observations dans la parcelle ayant aussi une forte influence sur la précision, une distribution judicieuse permettra parfois de réduire la taille de l'échantillon en conservant une précision acceptable.

2.5.2.1. Taille d'échantillon fixée à l'avance

Les différents indicateurs de précision peuvent être exprimés en fonction de la taille de l'échantillon et de la variance pour la variable étudiée ou de paramètres d'un modèle. Par conséquent, si on dispose d'une prévision de variance ou des paramètres adéquats pour un modèle fiable (voir 2.6), on peut déduire un nombre minimal d'observations à réaliser pour que la précision

soit satisfaisante (voir les formules associées à chaque modèle).

2.5.2.2. Échantillonnage adaptatif

Les méthodes d'échantillonnage adaptatif s'imposent quand il est nécessaire de contrôler la précision de l'échantillonnage mais que celle-ci ne peut pas être estimée avant le début des observations. On présente ici une méthode par phases et une méthode séquentielle. Ces méthodes peuvent nécessiter de réaliser quelques calculs au cours de l'échantillonnage.
On notera bien que les méthodes d'échantillonnage adaptatives ne sont pas appropriées si les observations durent dans le temps, comme pour le piégeage.

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 observations adaptés selon les informations obtenues dans la première phase. Les résultats des observations de la première phase ne sont perdus, ils sont simplement complétés par ceux de la deuxième phase.
La première phase peut permettre d'estimer la variance, même si cette estimation est très grossière de fait du petit nombre d'observations (voir 2.4.2). Elle peut aussi servir à déterminer une moyenne approximative, dont on déduit la précision atteignable à l'aide d'un modèle.
Enfin, la première phase d'échantillonnage peut révéler des hétérogénéités au sein de la parcelle auxquelles le plan d'échantillonnage peut être adapté (voir 2.7).

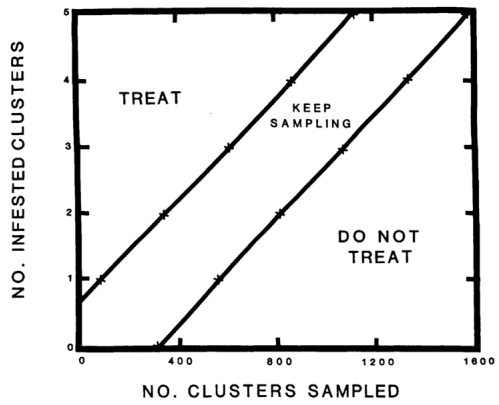

FIGURE 2.4.: Diagramme pour simplifier l'échantillonnage séquentiel - *Source : Ring et al. [1989]*

Échantillonnage séquentiel Les procédures d'échantillonnage séquentiel sont en général proposées pour des estimations d'incidence ou pour des comptages, leur principe est que l'échantillonnage se poursuit jusqu'à ce que la précision soit satisfaisante. De telles procédures permettent donc, 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. A partir du modèle en question, 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). 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. On peut représenter sur un diagramme, avec N en abscisses et T_N en ordonnées, les zones ou la précision est satisfaisante.
Par exemple, Ring et al. [1989] propose un tel diagramme (figure 2.4) dans le cas où l'objectif est de savoir avec une certitude donnée si la densité de certains ravageurs dépasse un seuil. On pourra aussi se reporter aux deux diagrammes de l'exemple sur l'étude des carpocapses (4.2). Enfin, le code R pour tracer de tels diagrammes est disponible en annexe (B.1).

2.6. Modèles

En faisant appel à des modèles, il est possible de mieux évaluer la précision obtenue lors d'un échantillonnage, il est surtout possible de prévoir la précision obtenue pour une certaine taille d'échantillon. Pour cela, les informations disponibles *a priori* à propos de ce qui est observé sont mises à profit. La pertinence des modèles proposés sera caractérisée par leur bon ajustement aux données existantes d'échantillonnage.
La disponibilité de données *a priori* est cruciale. Il peut s'agir d'une estimation de la moyenne, d'une estimation de la variance, d'une estimation de la répartition entre classes des observations, de paramètres d'une loi empirique ou d'une distribution, d'un variogramme empirique, ...

2.6.1. Écarts-types et coefficients de variation

Dès que la variable mesurée est numérique, on peut calculer après l'échantillonnage une estimation s de l'écart-type et une estimation $\hat{m}$ de la moyenne pour la variable étudiée. Alors, si les observations sont indépendantes (*i.e.* non-corrélées), l'écart-type qui serait 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 $c_v = \frac{s}{m\sqrt{N}}$. Dans le cas où les résultats des observations sont corrélés, 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 (voir 2.7.5).
Si, grâce à une modélisation, un échantillonnage adaptatif, ou tout autre méthode, une estimation de l'écart-type est disponible avant la fin de l'échantillonnage, la taille de l'échantillon N pourra être adaptée pour obtenir une précision voulue. Ainsi, pour obtenir un écart-type sur la moyenne inférieur 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

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

2.6.2. Modèles pour la moyenne calculée

Pour N grand, la moyenne calculée après l'échantillonnage $\hat{m}$ suit de plus en plus précisément une loi de probabilité connue. Cette loi permet d'établir des intervalles de confiance et de donner

une certitude pour les tests de dépassement de seuil. Ces modèles utilisent des approximations de la moyenne et de la variance qui lui est associée (estimée par $\frac{s^2}{N}$). Comme précédemment, ces approximations peuvent être celles obtenues à la fin de l'échantillonnage, ou bien elles peuvent être disponibles plus tôt ce qui permet de prévoir la précision qui sera obtenue pour une taille d'échantillon donnée.

2.6.2.1. Loi Normale

La modélisation par la loi normale est appropriée pour toutes les moyennes de variables numériques à condition que le nombre d'observations soit grand. Pour un nombre d'observations intermédiaire (de 10 à 30), 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.

Intervalle de confiance Selon ce modèle, l'intervalle de confiance à $100(1-\alpha)\%$ pour la moyenne est

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

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite donné par le tableau 2.1

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

TABLE 2.1.: Quantiles de la loi normale centrée réduite

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 de d_0, il faut prendre

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

Dépassement de seuil Si la moyenne calculée $\hat{m}$ est supérieure à un seuil m_0, alors la moyenne réelle est supérieure à m_0 avec un niveau de confiance

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

où Φ est la fonction de répartition de la loi normale.
De même, la moyenne est entre les valeurs seuil m_1 et m_2 avec un niveau de confiance

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

2.6.2.2. Loi de Student

Dans le cas où le nombre d'observations est faible, l'estimation de la variance est imprécise (voir A.3), ce qui perturbe l'estimation de la précision pour la moyenne. L'utilisation de le loi de Student pour modéliser la moyenne obtenue permet de prendre en compte cette imprécision. En revanche, elle n'est pertinente que si la variable X suit elle même approximativement une répartition normale, ce qui peut être vérifié par des tests statistiques (Kolmogorov-Smirnov), ou plus approximativement sur un histogramme. Si la variable est discrète mais avec suffisamment de valeurs différentes, et ayant un histogramme symétrique, on pourra aussi utiliser la loi de Student.
Les formules pour les intervalles de confiance, et la confiance associée au dépassement de seuil sont les mêmes que pour 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 (voir annexe A.6).

2.6.3. Modèles sous forme de lois de probabilité pour la variable mesurée

Dans certains cas, des modèles plus élaborés permettent des prévisions plus efficaces de la précision, et donc un meilleur choix des tailles d'échantillon. Un complément sur l'ajustement des modèles est disponible en annexe (A.5). 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*).

2.6.3.1. Loi de Poisson

La loi de Poisson modélise en théorie le résultat de comptage, dans des zones de même dimensions, d'individus répartis aléatoirement dans l'espace. Elle a la particularité d'avoir une variance égale à sa moyenne. On l'évoque ici à titre d'exemple bien qu'elle soit rarement utilisée dans le domaine de la protection des cultures.

2.6.3.2. Loi binomiale

La loi binomiale de paramètres n, 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 $(1-p)$). Son espérance et sa variance sont :

$$m = np \qquad \sigma^2 = np(1-p)$$

Elle est naturellement adaptée pour modéliser le résultat de comptages des plantes subissant un stress biotique (voir typologie de variables 2.4.3.2) parmi un ensemble de plantes de taille fixée connue n. p est alors l'incidence moyenne du stress biotique étudié. En pratique, les observations peuvent ne pas être indépendantes, dans ce cas la précision de l'estimation de p sera sur-estimée (voir paragraphe 2.6.3.2).

Estimation et précision pour l'incidence Ici n correspondra au nombre d'observations N. Alors l'estimation $\hat{p}$ de l'incidence 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

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

On en déduit l'intervalle un confiance valide sous conditions :

$$I_{cp} = [p_i, p_s] = [\hat{p} - z_{1-\alpha/2} s_{\hat{p}}, \hat{p} + z_{1-\alpha/2} s_{\hat{p}}]$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite donné par le tableau 2.1. Cet intervalle de confiance est valable à condition que Np_i, Np_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 êtres 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

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

Intérêt en l'absence d'information *a priori* Si le but de l'échantillonnage est de déterminer une incidence dont seulement un vague estimation est disponible (mais aucun paramètre de modélisation), cette première estimation permet tout de même de se faire une idée de la variance minimale qui pourrait être observée. Si plusieurs valeurs de p sont envisagées, celle la plus proche de 0.5 donnera la variance la plus élevée.

Corrélation et sur-dispersion En cas de corrélation spatiale (voir 2.7.5), si les observations sont trop proches les unes des autres, alors elle ne seront pas indépendantes. Par conséquent la variance estimée comme précédemment pour $\hat{p}$ sera inférieure à celle observée en répétant l'échantillonnage.
Ce phénomène peut être quantifié lors d'un unique échantillonnage si les observations sont rassemblées en grappes (voir échantillonnage par grappes, 2.8.4.2). 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'ensemble $p_1, ..., p_N$ 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 phénomène d'agrégation peut être présent. Un modèle de loi bêta-binomiale pourrait alors être pertinent pour les résultats des comptages dans les grappes.

2.6.3.3. Loi bêta-binomiale

La loi bêta-binomiale modélise le nombre de 1 obtenus pour n tirages d'une variable de Bernoulli dont le paramètre, aléatoire et qui change 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, 2.8.4.2 ; voir aussi typologie de variables 2.4.3.2). 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

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

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 :

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

La loi bêta-binomiale peut aussi être paramétrée par n, p et un coefficient ρ qui reflète la sur-dispersion liée aux corrélations spatiales (exemples figure 2.5). C'est la paramétrisation par défaut proposée par le package R `VGAM` [Yee, 2015].

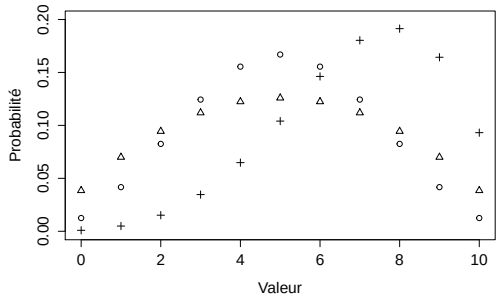

FIGURE 2.5.: Densités de la loi bêta-binomiale pour $n = 10$, $p = 0.5$, $\rho = 0.1(\circ)$; pour $n = 10$, $p = 0.5$, $\rho = 0.2(\triangle)$; pour $n = 10$, $p = 0.7$, $\rho = 0.1(+)$

Intérêt 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 données, ce qui permet de l'utiliser pour prévoir la précision d'échantillonnages à venir (ou mettre en place un échantillonnage adaptatif, 2.5.2.2). L'expression de la variance entre réalisations de la variable modélisée X est alors :

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

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

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

On en déduit aussi une méthode d'estimation de ρ (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], on en donne une exemple sur de données de prédation de graines (voir 4.1).

Objectif en terme d'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 :

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

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 :

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

Si l'intervalle de confiance de niveau $(1-\alpha)$ doit avoir une demi largeur inférieure à d, il faut donc prendre

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

Objectif en terme de 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

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

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 d'une difficulté pour mener à bien les 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 $\frac{n_{souhaite}}{n_{reel}}$.

2.6.3.4. Loi binomiale négative

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. Son espérance et sa variance sont :

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

En écologie le paramètre p n'est pas utilisé, de même que le sens initial de la définition, car ils ont peu de sens de ce point de vue, on préfère prendre directement μ comme paramètre. Alors $p = \frac{\mu}{k+\mu}$, $\sigma^2 = \frac{\mu(k+\mu)}{k}$.
Grâce à sa forme asymétrique, qui attribue une probabilité non nulle à toutes les valeurs positives, la loi négative binomiale s'ajuste bien à de nombreuses données de comptage non-bornées (voir typologie de variables 2.4.3.3), 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] souligne 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 toutefois pas un modèle très robuste, elle doit donc 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) mais représentatives de la diversité de situations d'échantillonnage pouvant exister.

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. Dans ce cas, un groupe de parcelles de la même région sont échantillonnées, et la loi explique bien la relation observée entre la moyenne et la variance des comptages.
Mukhopadhyay and Banerjee [2015] présente de manière approfondie l'utilisation de la loi binomiale négative pour mettre en place un échantillonnage adaptatif, il propose un exemple sur le comptage de doryphores.

Objectif en terme de 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 :

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

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 2.6 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%.

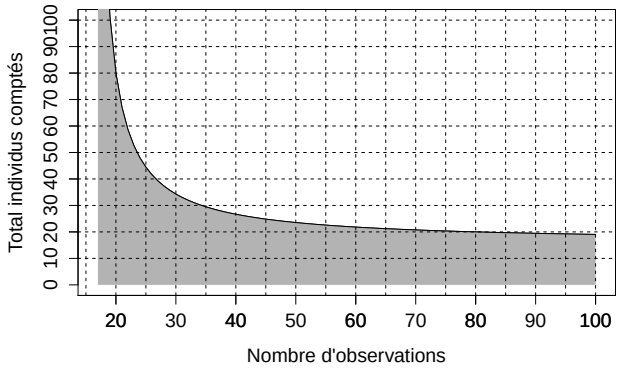

FIGURE 2.6.: Diagramme pour simplifier l'échantillonnage séquentiel avec pour modèle une loi négative binomiale. Il faut poursuivre l'échantillonnage tant qu'on se trouve dans la zone en gris pour obtenir un coefficient de variation de 25%.

Objectif en terme d'écart-type Pour un objectif de précision en terme 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 de 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 2.7 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.

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]. L'estimation de k connaissant la moyenne et la variance est :

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

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] (voir 2.6.4.3). Il y a un lien direct entre k et β, le paramètre de la relation entre m et $\overset{\star}{m}$: $\beta = 1 + \frac{1}{k}$.

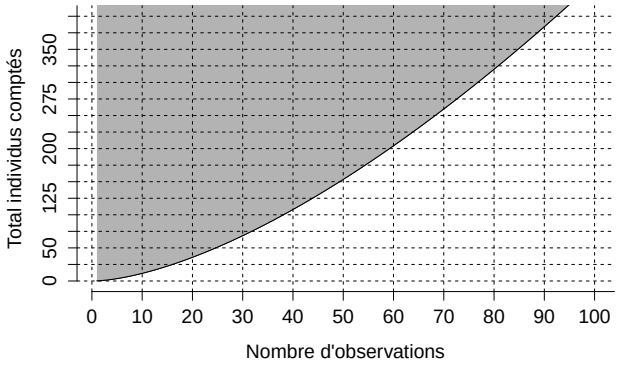

FIGURE 2.7.: Diagramme pour simplifier l'échantillonnage séquentiel avec pour modèle une loi négative binomiale. Il faut poursuivre l'échantillonnage tant qu'on se trouve dans la zone en gris pour obtenir un écart-type de 1.

2.6.3.5. 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 (voir typologie de variables 2.4.3.4) 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 $(v_1, v_2, \ldots, v_k)$ et $p_1, p_2, \ldots, p_k$ 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 tiré suit une loi binomiale de paramètres n, p_i. Si la variable aléatoire tirée est quantitative, alors sa moyenne sur N tirages indépendants a l'espérance et la variance suivantes :

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

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 v_i a été observé, 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 :

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

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

Coefficients sur les classes Si comme Aubertot et al. [2004], on attribue des coefficients 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.

Observations corrélées Si les observations ne sont pas indépendantes, la précision exprimée par les indicateurs précédents sera certainement sur-estimée.

2.6.4. Modèles empiriques de la variance

2.6.4.1. Loi de puissance de Taylor

Il s'agit d'un 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 suivante entre variance et moyennede résultats de comptages pour une population donnée (voir typologie de variables 2.4.3.3) :

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

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).

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, toutes les deux obtenues par le même protocole d'observation, et avec le même type de répartition spatiale (pour avoir une même influence des éventuelles corrélations spatiales), mais pas forcément avec la même taille d'échantillon (voir exemple pour le piégeage de carpocapses 4.2).

Objectif en terme de 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 ou 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.

Objectif en terme d'écart-type ou d'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 estimé inférieur à s 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

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

Pour une contrainte sur l'intervalle de confiance, on peut utiliser les formules pour la loi normale et en déduire la contrainte correspondante en terme d'écart-type.

2.6.4.2. Loi binomiale de puissance

Une telle loi est l'équivalent de la précédente dans le "cas binomial", c'est à dire pour l'étude du nombre d'individus dans un certain état dans des groupes de taille fixe. Elle fait en fait un lien entre la variance théorique du cas binomial (indépendance) et la variance observée qui est plus grande (sur-dispersion). Elle s'utilise dans le même contexte que la loi bêta-binomiale [Madden and Hughes, 1999] (voir typologie de variables 2.4.3.2).
On a donc N observations aboutissant à des valeurs $X_1, \ldots, X_N$ 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 subissant une bioagression dans des grappes de n plantes.
On s'intéresse alors aux proportions $x_i = \frac{X_i}{n}$ (qui sont les incidences dans chaque grappe). Dans

le cas ou l'état de chaque plante au sein d'une grappe serait indépendant des autres, la variance entre les valeurs $x_1, \ldots, x_N$ serait donnée par la formule :

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

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

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

où A et b sont des paramètres plus ou moins caractéristiques d'un espèce et d'une méthode d'observation données. Pour simplifier les expression, 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

$$s^2_{\hat{p}} = \frac{s^2_x}{N} = \frac{a}{N}(p(1-p))^b$$

Taille d'échantillon selon le coefficient de variation souhaité Si on s'attend 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 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].

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

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

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

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

2.6.4.3. Indice de Lloyd et régression d'Iwao

Ces modélisations s'appliquent aux mêmes situations que la loi de puissance de Taylor (2.6.4.1). Elles sont, elles aussi, basées sur la moyenne et la variance de résultats de comptage.

Deux indices distincts ont été introduits par Lloyd [1967] :

Index of mean crowding : $\overset{\star}{m} = m + \frac{s^2}{m} - 1$

Index of patchiness : $P = \frac{\overset{\star}{m}}{m} = 1 + \frac{s^2 - m}{m^2}$

$\overset{\star}{m}$ peut être interprété comme le nombre moyen de voisins d'un individu compté. P peut être interprété comme le 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, il dépendent par contre du plan d'échantillonnage et de la taille des quadrats.

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

$$\overset{\star}{m} = \alpha + \beta m$$

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 $\alpha = 0$, alors $\beta = P$. 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 individu. Waters et al. [2014] indique aussi que ce choix conduit à prendre des échantillons plus petits, et donc à réduire le coût, ce qui ne s'est pas traduit par une perte de précision pour les quelques tests qui ont été menés.

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

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

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

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

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

2.7. Structuration spatiale

Trois types de structuration au niveau de la parcelle sont à prendre en compte pour l'établissement d'une stratégie d'échantillonnage. Premièrement, la disposition des cultures influence le

type de plan d'échantillonnage. Deuxièmement, les valeurs de la variable étudiée peuvent être hétérogènes à cause de divers facteurs. Troisièmement, les valeurs peuvent être hétérogènes à cause de phénomènes d'agrégation ou de corrélation spatiale.
Le stress biotique ou l'auxiliaire étudiés peuvent avoir différents niveaux et causes de structuration spatiale qui influencent des aspects différents de la stratégie d'échantillonnage : soit le plan, soit les modèles utilisés. Si plusieurs sources d'hétérogénéité sont présentes, il peut être judicieux de ne prendre en compte que la principale.
Les modélisations ou répartitions des observations sont suggérées dans l'optique d'une augmentation de précision. Cela peut se traduire indirectement par une réduction de coût si on peut réduire le nombre d'observation et conserver une précision satisfaisante.

2.7.1. Influence du type de culture sur l'organisation de la parcelle

Le type de culture structure les parcelles étudiées (rang, arbre, ...) et donc influence la stratégie d'échantillonnage. Les champs et les vergers sont les cas les plus fréquents, mais il faut aussi considérer les serres, les vignobles,... Dans tous les cas, l'organisation spatiale des cultures impose des procédures de choix des unités statistiques à observer. En particulier, il faut prendre en compte ses motifs spatiaux réguliers pour ne pas biaiser les résultats (notamment dans le cas de cultures associées en bandes).

Cultures en champs Dans le cas de cultures en champs, on peut considérer que le champ est continu, on prélève ou examine alors toutes les plantes dans des quadrats (délimitation virtuelle) de forme prédéterminée. Les quadrats peuvent être repérés le long de rangs, placés totalement aléatoirement ou encore positionnés sur un quadrillage.
On peut aussi considérer chaque plante individuellement. Alors, certaines peuvent être prélevées le long de rangs, par grappe ou non. Le champ peut aussi être parcouru selon un parcours bien choisi et des observation faites aléatoirement ou régulièrement le long du parcours (parcours aléatoire, en X [Burts and Brunner, 1981], en zigzag [Workneh et al., 1999], en losange).
Enfin un champ se prête bien à un examen continu, par exemple dans Barroso et al. [2005] le champ est parcouru dans son intégralité en enregistrant en permanence s'il y a des plantes adventices ou non.

Vergers Dans un verger, les rangs, les arbres, branches, bourgeon ou fruit, forment naturellement une structure qui s'impose à la façon de répartir les observations. On obtient un *échantillonnage par degré* en choisissant les arbres, puis les branches dans les arbres, etc (par exemple Brown et al. [1993] et Navarro-Campos et al. [2012], Ojiambo and Scherm [2006] pour les myrtilles).
A chaque niveau, les unités sont faciles à distinguer et numéroter, ce qui peut favoriser un tirage aléatoire. Il est aussi envisageable que la sélection soit aléatoire à certains niveaux et pas à d'autres. Le choix du nombre d'unités à sélectionner à chaque niveau en fonction des différentes variances observées peut être optimisé, toutefois de telles subtilités (voir Ojiambo and Scherm [2006]) ne seront pas développées.

Parcelle vraisemblablement homogène En cas d'absence de structuration soupçonnée de la variable observée (moyenne identique sur toutes les zones de la parcelle, variance stable et faibles

corrélations spatiales) , le plan d'échantillonnage sera de préférence un *plan aléatoire simple* (garanties théoriques de fiabilité, 2.8.3.1), un *plan systématique* (pratique et adapté en cas de corrélation spatiale, 2.8.3.2), ou un autre plan pouvant être considéré comme équivalent sous certaines conditions (losange, Z, W, X, U). Un *échantillonnage par grappes*(2.8.4.2) peut aussi être envisagé pour les gains de temps qu'il permet, comme dans les protocoles Vigicultures®.
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.

2.7.2. Hétérogénéités dues à l'environnement

Les diverses propriétés de la parcelle étudiée, comme l'ensoleillement, la texture et la structure du sol, la proximité d'une haie ou d'une friche, peuvent créer des hétérogénéités du stress biotique ou de la présence d'auxiliaires. Elles peuvent se présenter sous forme de gradients à l'échelle de la parcelle ou de zones bien délimitées. Dans les deux cas, les hétérogénéités vont induire une sur-estimation de la variance. Ainsi une moyenne calculée avec précision pourrait sembler peu fiable à cause de la variance élevée parmi les observations. Si le plan d'échantillonnage n'est pas bien conçu, les hétérogénéités peuvent aussi se traduire par un biais important.

Repérer La présence d'un auxiliaire, d'une maladie ou d'un ravageur étudié peut avoir une relation connue avec certain facteurs environnementaux. Si ces facteurs varient au sein de la parcelle on s'attendra aux mêmes variations pour l'organisme ou le phénomène étudié.
Les résultats d'un échantillonnage passé, ainsi qu'une évaluation visuelle rapide, peuvent aussi révéler des hétérogénéités à prendre en compte.
Si une variable fortement corrélée à la variable mesurée est connue avec précision sur la parcelle, elle peut être utilisée pour mettre en place un *échantillonnage à probabilités inégales*, qui ne sera pas développé car il semble trop rarement applicable, mais qui peut être retrouvé dans les ouvrages de Gregoire and Valentine [2007] et de Gruijter et al. [2006].

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 *coupes* (2.8.4.3) dans la direction principale du gradient (de Gruijter et al. [2006, page 78]) ; un *échantillonnage systématique*(2.8.3.2) qui garanti la représentativité peut aussi être mis en place.
Surtout, le découpage de la parcelle en *strates*(2.8.4.1) où la variable étudiée est plutôt homogène, permet à la fois d'améliorer la précision et de mieux l'estimer *a posteriori*.

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

2.7.3. Hétérogénéités dues aux pratiques culturales

Les diverses pratiques culturales (labour, semis, espèces antérieures, traitements phytosanitaires) sont parfois source d'hétérogénéités marquées pour le stress biotique ou la présence

d'auxiliaires, dont certaines sont périodiques. Comme celles liées à l'environnement, ces hétérogénéités sont sources de biais et d'une surestimation de la variance si le plan d'échantillonnage est mal choisi.

Repérer Si l'itinéraire technique n'est pas le même sur l'ensemble de la parcelle étudiée, des hétérogénéités correspondantes pourraient être observées. Elles sont parfois visibles si elle impliquent un stade de culture différent ou une forte concentration de plantes adventices par exemple.

Adapter le plan d'échantillonnage Dès que la parcelle peut être découpée en zones (représentant une proportion connue de la surface totale), et au sein desquelles la variable étudiée est plus homogène que dans l'ensemble de la parcelle, alors une *stratification*(2.8.4.1) peut être mise en place (de Gruijter et al. [2006, page 78], Clostre et al. [2014]). Si la variable observée risque de présenter des variations périodiques, il faut veiller à ce qu'elles ne soient pas le source d'un biais. Pour cela, les observations ne doivent pas être disposées avec une périodicité spatiale multiple de celle de la variable (cas de *l'échantillonnage systématiques*, 2.8.3.2).
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 (voir 2.2).

Exemples dans la littérature : Clostre et al. [2014](pollution des sols)

2.7.4. Agrégation spatiale

On parle d'agrégation quand la population d'individus étudiée n'est pas répartie aléatoirement dans l'espace mais forme des agrégats. Elle a été largement étudiée en écologie, notamment pour les populations animales [Taylor, 1984]. L'agrégation se manifeste quand un comptage d'individus à lieu dans des quadrats ou sur des organes de plantes, plus elle est importante à l'échelle des unités d'observations, plus la variance entre les comptage sera élevée. En revanche, un phénomène d'agrégation à des distances plus grandes ou plus petites que celle des unités d'observation n'augmentera pas forcément la variance. Si l'agrégation à grande distance a pour conséquence des valeurs proches pour des comptages dans des unités d'observations proches, on parlera plutôt de corrélation spatiale (voir 2.7.5) ; dans ce cas la variance sur l'ensemble de la parcelle risque d'être sous-estimée.
Dans certain cas, il est possible d'ajuster un modèle d'agrégation à des données issues d'un protocole de comptage fixé, avec une taille de quadrat, ou un organe donné. Le modèle possède alors un paramètre relié au niveau d'agrégation de la population ; si ce paramètre est stable pour une espèce donnée ou un type d'observations donné, il pourra être utilisé pour prévoir la variance entre les comptages.

Sources d'agrégation L'agrégation spatiale des individus peut avoir des origines diverses, liées à la biologie de l'espèce considérée. Les sources d'alimentation des insectes et leurs capacités de déplacement, le mode de reproduction d'une plante adventice, et bien d'autres aspects influent

sur leur agrégation spatiale.

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

Modélisation Pour des résultats de comptage dans des quadrats, si les individus sont répartis aléatoirement dans l'espace, alors les valeurs suivent une *loi de Poisson*(2.6.3.1), qui a une variance égale à sa moyenne.
Si il y a de l'agrégation, la *loi binomiale négative*(2.6.3.4) est souvent un bon modèle pour la distribution des résultats de comptages. Elle peut être paramétrée par sa moyenne et un paramètre k qui est petit pour une population très agrégée. Quand k est très grand, cette loi approche le loi de Poisson. Le paramètre k est parfois caractéristique d'une espèce est peut être utilisé pour faire des prévisions mais ce n'est pas toujours le cas. Il est important de vérifier la robustesse du modèle, c'est à dire la stabilité du paramètre entre plusieurs échantillonnages passés.
La *loi de puissance de Taylor*(2.6.4.1), qui établit une relation entre la moyenne et la variance des résultats de comptages (pour une espèce et une zone de comptage données), est souvent citée en écologie. Elle est plus robuste que d'autres modélisations et permet par conséquent de bien prévoir la variance.
Enfin, divers indices d'agrégations ont été proposés (2.6.4.3). Quand ils se révèlent stables pour un certain type d'échantillonnage, ils peuvent aussi être mis à profit pour prévoir la variance. Chacun 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.

Ce large panel de modèles présentés augmente les chance qu'un des modèles ait déjà été ajusté pour le type d'échantillonnage auquel on s'intéresse. Des correspondances existent entre certains paramètres et indices, elles pourraient être utilisées pour rassembler plus d'informations utilisables concernant la même espèce.

Si ces modèles ont initialement été prévus pour des comptages sur des zones, il peuvent très bien s'adapter à d'autres situations, comme des comptages sur des organes, ou encore des piégeages.

Lien avec le plan d'échantillonnage Si la taille des unités d'observation (quadrats ou organes ou autre) varie, alors les paramètres des différents modèles d'agrégations risquent de varier aussi. En effet, plus les quadrats sont grands, plus les résultats de comptage deviennent proche 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 examinée, il vaut mieux examiner de nombreux petit quadrats plutôt que peu de grands. De cette façon, l'agrégation peut, en plus, être mieux quantifiée.

Remarque sur la répulsion Les phénomènes de répulsion entre individus d'une même espèce, inverse des phénomènes d'agrégation, sont parfois présents mais n'ont pas d'impact du point de vue de l'échantillonnage.

2.7.5. Corrélation spatiale

La corrélation spatiale entre les observations, c'est-à-dire la corrélation entre observations proches, est souvent observée. Elle peut conduire à sous-estimer la variance de la variable observée, et donc à sur-estimer la précision obtenue. En effet, la variance estimée de manière classique sur un échantillon avec des observations corrélées (parce que trop proches) sera plus faible que celle estimée sur un échantillon sans observations corrélées (parce que suffisamment éloignées).
En fonction de l'objectif, de la variable étudiée et des informations disponibles sur la corrélation (modèles), plusieurs approches pourront être adoptées.

Corrélation et incidence Sans corrélations, la loi binomiale est un bon modèle pour étudier l'incidence. Par contre en présence de corrélation spatiale, elle conduit à une sous-estimation de la variance de l'estimateur. On peut modéliser ce phénomène par une expression corrigée de la variance (avec un coefficient en plus qui traduit la sur-dispersion). Le choix conservé ici est celui de la loi bêta-binomiale.
La loi bêta-binomiale et d'autres modèles [Collett, 1991, page 194],[Madden and Hughes, 1999] permettent d'obtenir un expression de la variance proche de celle pour la loi binomiale avec un paramètre ρ supplémentaire qui est positif pour indiquer de la sur-dispersion et nul dans le cas où tous les individus observés sont indépendants.

Adaptation du plan Pour une étude d'incidence, 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]. La modélisation permet aussi de prévoir la précision à partir des paramètres ajustés du modèle (issus de la bibliographie ou d'échantillonnages passés) et d'une estimation grossière de l'incidence réelle (qui peut être obtenue par expérience ou par un échantillonnage adaptatif, voir 2.5.2.2). Toutefois, un échantillonnage par grappes aura un coût plus élevé qu'un échantillonnage simple ou systématique, pour la même précision, en présence de corrélation spatiale [Vaillant, 1996, page 35].
Pour une étude de sévérité ou autre, la démarche précédente n'est pas pertinente, il vaudra donc mieux espacer les observations autant que possible par un échantillonnage systématique. L'espacement régulier des observation est alors utile pour estimer la variance réelle (qui est supérieure à la variance observée) à l'aide d'un variogramme.

Le variogramme Un variogramme est un graphique qui représente l'écart au carré moyen des valeur 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). Un variogramme présente en général une partie croissante, partant ou non de l'origine puis un palier qui correspond à la variance de la variable

concernée (figure 2.8). La partie du variogramme avec des valeurs inférieure 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 intervalles de $\pm 5\%$ autour du palier est appelée portée, et on considère qu'il n'y a plus de corrélation significative à partir de cette distance.

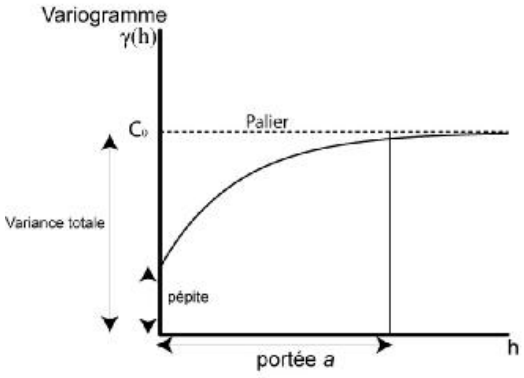

FIGURE 2.8.: Variogramme

Un variogramme empirique peut être obtenu à partir de valeurs de la variable Z étudiée en un certain nombre de points de l'espace des coordonnées $x_1, ..., x_I$. Il vaut, pour une distance h, la moitié de la moyenne des écarts au carré des valeurs en des points distants d'environ h (pour que le résultat soit utilisable il vaut mieux faire les calculs avec un intervalle de tolérance $[h - \delta, h + \delta]$). On a donc la formule suivante :

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

où $n(h)$ est le nombre de paires de points dont la distance est comprise dans $[h - \delta, h + \delta]$. Le travail sur les données disponibles a montré qu'il est difficile d'obtenir un variogramme de qualité, sauf avec un jeu de données très important (voir 4.1, 4.3). Si un variogramme peut néanmoins être obtenu pour une espèce donnée dans un système de culture donné, il pourra 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].
Rew and Cousens [2001] signale, en outre, que la portée dépend fortement du système de culture qui influence la dispersion des différents organismes.

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.
Cressie [1993, page 14] donne un exemple de cette démarche en dimension 1 (observations sur une ligne). S'il y a un coefficient de corrélation ρ entre deux observations successives, alors la

variance de la moyenne sur N observations est

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

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

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

On obtient par exemple pour $N = 10$ et $\rho = 0.26$, $N' = 6.2$. 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_0^2 est 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.

2.7.6. Adaptation à des contraintes pratiques

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

Stratification La parcelle peut être découpée selon la difficulté d'échantillonnage dans chaque zone, l'échantillonnage peut ainsi être intensifié dans les zones plus faciles d'accès sans qu'un biais ne soit créé (voir 2.8.4.1).

Observations en grappes Dans la mesure où la corrélation spatiale n'est pas trop importante, un échantillonnage par grappes (voir 2.8.4.2) peut être adopté de façon à limiter les déplacements nécessaires.

2.8. Plans d'échantillonnage

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, on néglige parfois ce problème en faisant l'hypothèse (raisonnable, voir annexe A.1) que la façon de choisir les observations donne un résultat représentatif de la parcelle.

Les plans d'échantillonnage sont généralement distingués selon qu'ils sont déterministes ou non et selon leur complexité. Toutefois, la classification proposé ainsi que les noms peuvent se recouper et ne pas être partagés partout. Chaque type de plan a des avantages et des inconvénients relativement au coût et à la précision, le choix d'un plan est donc souvent fondé sur un compromis.

2.8.1. La population statistique

La population statistique est l'ensemble formé de ce qui va être tiré aléatoirement au moment de l'échantillonnage. Ce peut donc 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 d'observation (sous-entendu : d'une unité statistique).

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 des protocoles déjà existants qui ont pu être documentés ou selon les spécificités biologique de ce qui est étudié. Pour mettre en balance la taille et le nombre de quadrats à examiner, la réflexion à avoir est la même que pour un échantillonnage par grappes (voir suite, 2.8.4.2) : Avoir plus de quadrats, plus petits, permet un gain de précision si la variable étudiée est corrélée spatialement, mais en contre-partie cela augmente le coût de déplacement.

2.8.2. 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 de prendre en compte simplement une expertise dans le domaine étudié, néanmoins il y a un risque de biais du à la subjectivité du concepteur du plan d'échantillonnage. En effet, contrairement aux échantillonnages aléatoires, on ne peut pas prouver théoriquement qu'il n'est pas biaisé.
En pratique, une version raisonnée des plans d'échantillonnage proposés est parfois mise en place, surtout pour les plans systématiques pour lesquels le biais est moins important.

2.8.3. Échantillonnage aléatoire à un niveau

Pour pouvoir exploiter les résultats des mesures statistiquement, le choix aléatoire des observations fournit des garanties théoriques. C'est lui qui permet d'utiliser les résultats sur la précision de la moyenne par exemple. Les mêmes résultats sont parfois utilisés pour des échantillonnages raisonnés, mais cela présente un certain risque d'erreur.
On ne peut assurer qu'une moyenne estimée par un échantillonnage est non-biaisée (voir 2.4.2) que si chaque unité statistique a une probabilité non-nulle d'être observée (c'est une des définitions d'un échantillon représentatif), ou si on fait d'autres hypothèses fortes sur la structure de la population observée.

2.8.3.1. Échantillonnage aléatoire simple

Un plan d'échantillonnage aléatoire simple est le choix aléatoire et indépendant d'un certain nombre d'éléments de la population statistique cible (figure 2.9).

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, en effet l'échantillonnage peut être

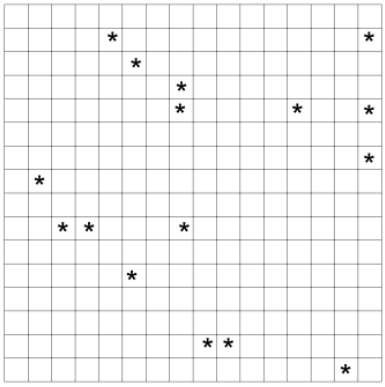

FIGURE 2.9.: Échantillonnage simple de 16 unités de surfaces sur 256 - *Source : Vaillant [1996]*

interrompu ou poursuivi sans que cela cause un biais (à condition d'examiner les éléments dans l'ordre de tirage).

Inconvénients L'échantillonnage simple offre une précision minimale qui se traduit par une variance élevée des estimateurs (par rapport à un plan systématique par exemple). De plus il peut être difficile à réaliser car il nécessite une procédure de tirage aléatoire qui puisse inclure n'importe quel élément de la population, il faut donc qu'ils soient tous repérés individuellement. Enfin, pour un échantillonnage adaptatif, le coût de déplacement et de repérage est élevé puisqu'on ne peut pas faire toutes les observations proches en même temps.

Variantes approximatives Si l'observateur choisit à peu près aléatoirement les emplacements de ses observations dans la parcelle, on considérera que les résultats théoriques sont toujours valides. Les observations peuvent par exemples être prises le long d'un parcours (comme ceux évoqués pour les échantillonnages systématiques).

Choix Finalement on choisira principalement ce plan pour pouvoir réaliser un échantillonnage adaptatif, à condition que les déplacements soient peu coûteux. Par exemple les déplacements seront courts dans des strates de petite taille. Le problème de coût peut être réduit avec un plan aléatoire à deux niveaux (voir 2.8.4.3).

2.8.3.2. Échantillonnage systématique

Un plan d'échantillonnage systématique repose sur un seul choix aléatoire, celui d'un élément de départ (figure 2.10). A partir de cet élément, des éléments régulièrement espacés sont examinés, selon un quadrillage ou un parcours (en U, en diagonales,...) par exemple. L'élément de départ est parfois choisi sans tirage aléatoire pour des raisons pratiques. Si les observations restent raisonnablement représentatives de la parcelle on considère tout de même que le résultat ne sera pas biaisé.

Avantages Le premier avantage est la simplicité de repérage des observations dans la parcelle. Les observations bien espacées et bien reparties dans la parcelles garantissent une bonne

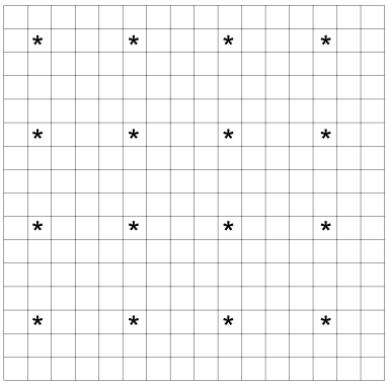

FIGURE 2.10.: Échantillonnage systématique de 16 unités de surfaces sur 256 - *Source : Vaillant [1996]*

précision, même en cas de forte corrélation spatiale ou d'hétérogénéités à l'échelle de la parcelle [Gregoire and Valentine, 2007, de Gruijter et al., 2006, pages 49 à 56 et page 103 resp.]. Cet avantage est d'autant plus important si on prend des éléments régulièrement espacés dans une liste de tous les éléments classés selon une variable connue fortement corrélée à la variable étudiée. Cette variable peut par exemple être la sévérité observée l'année précédente sur les arbres d'un verger.

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. De plus, un échantillonnage systématique peut introduire un biais important si la variable étudiée est périodique avec une période en rapport avec l'espacement entre observations.

Grille Échantillonner selon une grille (comme sur la figure 2.10) a le double avantage de bien espacer les observations et de garantir qu'elles soient réparties uniformément sur toute la parcelle. Une grille en quinconce maximise la distance entre les observations, et peut donc légèrement augmenter la précision

Parcours Échantillonner selon un parcours dans la parcelle, avec des observations espacées régulièrement, permet d'obtenir à un faible coût un aperçu assez large de la parcelle. 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 Vigicultures®)
- Deux diagonale, qui représente plus le centre de la parcelle
- Un losange, qui représente bien la parcelle et est pratique
- Un U pour pouvoir se déplacer le long des rangs (protocoles Rés0Pest)
- ...

Estimation de la variance En cas de forte corrélation spatiale de la variable étudiée, la variance calculée sur les données conduira à sous-estimer la variance pour la moyenne. Pour des observations régulièrement espacées, l'ordre de grandeur de cette erreur peut être estimé à partir d'un *variogramme* (voir 2.7.5) reflétant les corrélations spatiales de la variable étudiée.

Taille de l'échantillon Pour un échantillonnage systématique, le motif de répartition des observations ne change pas, c'est son échelle qui change de façon à toujours représenter l'ensemble de la parcelle. S'il y a beaucoup d'observation elles seront plus proches, s'il y en a moins elle seront plus dispersées.
Ce type de plan est peu adapté à l'échantillonnage adaptatif, on peut néanmoins échantillonner en plusieurs phases ; par exemple en ajoutant des observations à mi-distance entre les premières pour la deuxième phase. On peut aussi compléter un parcours avec un autre parcours ($\diamond$ et $+$, deuxième diagonale).

Choix L'échantillonnage systématique est un bon choix par défaut sauf si on souhaite mettre en place une démarche adaptative. Dans ce cas, il faut vérifier si un plan systématique laisse assez de libertés (voir paragraphe précédent).

2.8.4. Échantillonnage aléatoire à plusieurs niveaux

L'échantillonnage à plusieurs niveaux consiste dans un premier temps à découper la population totale en sous-populations (disjointes), puis à choisir des éléments 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.

2.8.4.1. Échantillonnage stratifié

Un plan d'échantillonnage stratifié est basé sur un découpage de la population totale. Il se compose de plusieurs autres échantillonnages effectués dans chacune des sous-populations (strates) obtenues (figure 2.11). Le nombre d'éléments choisi, et la façon de les choisir dans les strates n'est pas forcément le même, ce qui permet de prendre en compte des hétérogénéités de moyenne ou de variance dans la population totale. La moyenne de la variable étudiée est calculée dans chaque strate avant d'être rassemblée en une moyenne pour toute la parcelle. La stratification ne concerne qu'un degré de l'échantillonnage, chaque strate peut être échantillonnée de n'importe quelle manière.

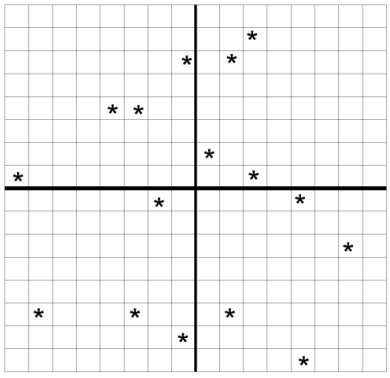

FIGURE 2.11.: Échantillonnage stratifié, 4 observations par strates - *Source : Vaillant [1996]*

Choix des strates La stratification se justifie dans plusieurs cas, et chacun implique une stratification différente. Ainsi les strates peuvent ...

... présenter des valeurs de la variable étudiée plus homogènes que sur l'ensemble de la parcelle.

... présenter une variance pour la variable étudiée plus homogènes que sur l'ensemble de la parcelle.

... être définies par leur coût d'échantillonnage (accès difficile en milieu de parcelle,...).

Pour pouvoir exploiter les résultats, il est impératif de connaître la proportion en surface (ou en nombre de plants/arbres) de chaque strate.

Avantages

- Des strates avec des observations aux valeurs proches permettent d'avoir dans chaque strate une estimation précise de la moyenne. Quand elles sont rassemblées en une moyenne sur toute la parcelle, on obtient une variance plus faible que si on avait échantillonné sur toute la parcelle d'un coup, qui reflète mieux la précision réelle. Alors, idéalement, les strates sont homogènes et présentent de fortes différences entre elles [Vaillant, 1996, page 28].
- Des strates avec une variance homogène permettent d'adapter le nombre d'observations dans chaque strate, pour avoir un précision comparable dans chacune, et une bonne précision pour la moyenne totale. Le tout en économisant des observations dans les strates à faible variance.
- Un découpage selon le cout d'échantillonnage permet de favoriser les observations dans les zones plus accessibles sans 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].

Inconvénients Le principal inconvénient est la difficulté de bien déterminer l'importance relative des strates, en surface ou en nombre total d'éléments. En cas d'erreur le biais créé peut être important.

D'autre part, Gregoire and Valentine [2007, pages 128] explique qu'il faut faire attention à ce que le coût de mise en place d'une stratégie élaborée ne soit pas supérieur aux économies réalisées via le gain de précision .

Enfin, si le nombre total d'observations est trop faible, il ne permettra pas de bien évaluer la variance, et donc la précision, dans chaque strate (voir 2.4.2, A.3).

Taille d'échantillon Si la taille de l'échantillon est fixée et qu'il faut répartir les observations entre les strates, les deux choix les plus pertinents sont [Gregoire and Valentine, 2007, page 142]

- Attribuer les observations proportionnellement à la taille de strates.
- 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 strates $s_1, \ldots, s_I$, 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 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 strates selon leur importance. Par exemple si on a I strates d'aires (ou d'importance) $a_1, \ldots, a_I$, avec les moyennes par strates $m_1, \ldots, m_I$, 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, 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

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

en considérant que les moyennes sur les différentes strates ne sont pas corrélées. La variance obtenue, qui reflète bien la précision de la moyenne, est plus faible que la variance qui aurait été estimée directement sur toute la parcelle.

Si il y a peu d'observations dans chaque strate, l'estimation des variances peut être imprécise. Gregoire and Valentine [2007] indique d'ailleurs que la loi de Student doit être préférée à la loi Normale pour établir des intervalles de confiance pour la moyenne. Dans le cas où il y a assez d'observations dans chaque strates, 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

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

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.

2.8.4.2. Échantillonnage en grappes

Pour un plan d'échantillonnage en grappes, les unités 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 de ces grappes est sélectionnée par un échantillonnage simple, systématique, ou autre. Enfin tous les éléments des grappes sélectionnées sont examinés (figure 2.12).

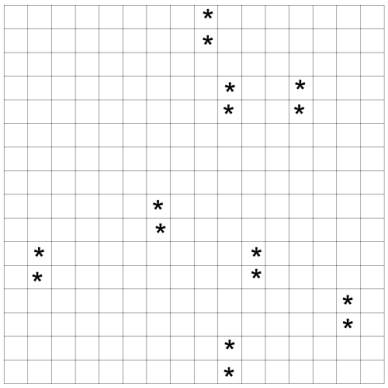

FIGURE 2.12.: Échantillonnage en grappes, 8 grappes de 2 grains - *Source : Vaillant [1996]*

Avantages Cet échantillonnage est privilégié pour des raisons de mise en pratique moins coûteuse et car il met en évidence l'hétérogénéité locale de la variable étudiée. Plus précisément, des observations proches les unes des autres sont plus pratiques à réaliser (repérage, transport du matériel).
Par ailleurs, dans le cas d'une étude d'incidence, l'incidence par grappe peut être calculée ce qui permet une meilleure estimation de la précision que sans grappes (la précision 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].

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

Lien avec la taille d'échantillon La taille des grappes et le nombre de grappes peuvent être déterminés selon un nombre total d'observations à atteindre (quand le plan est choisi pour des raisons de coût), ou séparément (quand les grappes servent à étudier les corrélations spatiales). Dans ce cas, plusieurs modèles permettent de travailler sur le nombre de grappes optimal, et la taille des grappes est fixée comme un compromis entre le coût et la représentativité de la grappe (souvent 5 ou 10 plantes).

Choix L'échantillonnage par grappe sera finalement à privilégier quand il permet de limiter le coût et que la corrélation spatiale est négligeable. Les grappes idéales du point de vue de l'information gagnée par observation seraient semblables avec une forte variance interne [Vaillant, 1996, page 33].

Remarques pratique : 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.

Ce peut être une façon de limiter l'influence de l'observateur qui pourrait introduire un biais s'il fallait sélectionner une seule plante à un point donné, il pourrait en effet choisir inconsciemment une plante plutôt que sa voisine sur un critère arbitraire.

2.8.4.3. Échantillonnage par degrés

L'échantillonnage par degrés est une notion très générale qui englobe tout les cas ou 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, un nouvel échantillonnage a ensuite lieu dans chaque élément sélectionné à la première étape et ainsi de suite (exemple figure 2.13). Ainsi un échantillonnage stratifié est un échantillonnage à deux degrés, exhaustif pour le premier degré.
Les plans d'échantillonnage par degrés permettent une adaptation fine à chaque situation, au détriment de la simplicité d'organisation.

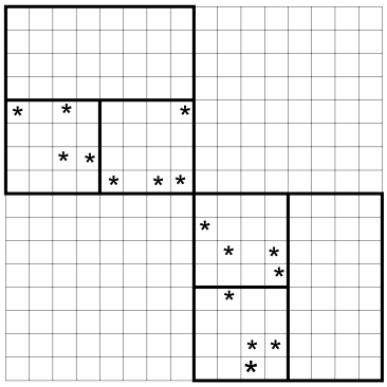

FIGURE 2.13.: Échantillonnage à 3 degrés - *Source : Vaillant [1996]*

Exemple : 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 degré. 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 degré 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.

Exemple : 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.
Des échantillonnages selon des coupes sont présentés par Barroso et al. [2005] et Clark et al. [2007].

2.8.5. 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 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 empêche de bien 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]. Par exemple, 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.

3. Développement d'un outil d'aide à la conception de stratégies d'échantillonnage

Un outil a été conçu de manière à articuler entre elles toutes les informations et recommandations sur la conception d'une stratégique d'échantillonnage (chapitre 2), ainsi ces information seront utilisables pour des utilisateurs ne maîtrisant pas forcément l'ensemble des aspects de l'élaboration d'une stratégie d'échantillonnage. L'outil est inspiré de diverses descriptions du processus de choix d'une stratégie d'échantillonnage qui sont présents dans la littérature (voir figure 3.1) [de Gruijter et al. [2006, page 29], Vaillant [1996, page 15], Domburg et al. [1994, page 154]].
Après avoir précisé les questions auxquelles il sera plus ou moins possible d'apporter une réponse grâce à l'outil, le schéma de principe est présenté, puis l'outil est exposé tel qu'il sera livré. Enfin, la question du test et de l'évaluation de l'outil est abordée.

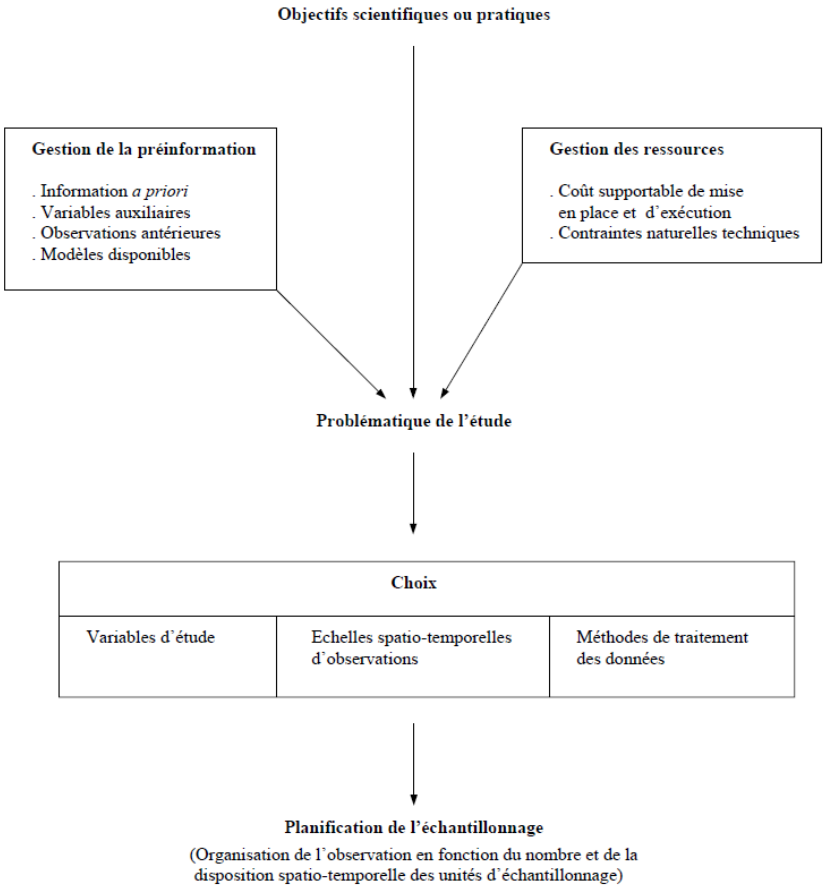

FIGURE 3.1.: Schéma d'un processus de décision pour l'échantillonnage - *Source : Vaillant [1996]*

3.1. Périmètre de l'outil

L'outil est source de conseils sur la répartition spatiale de observations à réaliser, il suggère aussi l'utilisation de modèles et des méthodes pour calculer une taille raisonnable d'échantillon. Pour cela, un maximum d'informations disponibles avant l'échantillonnage sont mises à profit. Toutefois, l'outil d'aide à la conception de stratégies d'échantillonnage produit ne peut pas contenir l'ensemble des stratégies d'échantillonnage possibles dans les domaines concernés. En effet, au cours des recherches effectuées, de nombreux cas particuliers d'optimisation d'échantillonnage ont été rencontrés, ces publications ne fournissant même pas toujours toutes les informations sur la stratégie d'échantillonnage à mettre en place. A chaque fois des modèles et des raisonnement spécifiques sont présentés, une telle diversité ne peut pas être englobée avec précision dans un outil qui doit rester générique.
En particulier, l'outil ne peut pas servir à traiter toutes les données utiles à la mise en place de l'échantillonnage. Par exemple, il n'est pas envisageable d'ajuster des distributions ou des lois à partir de données de mesures, d'estimer des moyennes, de fournir la meilleure stratification à partir d'un itinéraire technique,... Toutes ces opérations peuvent apporter des informations précieuses, mais elles doivent êtres laissées à la charge de l'utilisateur de l'outil. Celui-ci pourra éventuellement prendre des résultats plus travaillés comme entrées pour l'outil.
Enfin, l'outil proposé peut ne pas être la seule source d'optimisations. Il est complémentaire et non remplaçant des améliorations empiriques qui ont pu être développées.

3.2. Schéma de principe de l'outil

L'outil d'aide à la conception de stratégies d'échantillonnage est structuré de manière analogue à ceux qui peuvent être rencontré dans la littérature, car la démarche globale est la même quel que soit l'échantillonnage. Toutefois les questions et le détail des différentes étapes sont spécifiques aux problématiques de caractérisation de la composante biotique d'une parcelle agricole.

Dans un premier temps, les objectifs et les méthodes d'observation sont à préciser, et il faut rassembler les information potentiellement utiles à l'échantillonnage. Puis le choix cohérent d'un formalisme et d'un plan d'échantillonnage permettent de prévoir, dans une certaine mesure, la qualité des résultats et le coût de la réalisation de l'échantillonnage. Ces éléments sont enfin confrontés pour décider si la stratégie obtenue est satisfaisante ou si elle doit être retravaillée, ce qui implique de faire des compromis et de revoir les premières étapes de la réflexion.
L'ensemble de la démarche est représenté par le schéma 3.2. Elle se compose des étapes suivantes :

- La préparation consiste à préciser l'objectif, les méthodes d'observation, et les contraintes (voir 3.3.1).
- Des recherches sur la biologie du stress ou de l'auxiliaire étudiée, sur des données d'échantillonnage existantes, sur des modèles utilisés, permettent de recueillir des informations utiles pour l'optimisation de la stratégie d'échantillonnage (voir 3.3.2).
- La formalisation mathématique, qui peut aboutir à l'utilisation d'un modèle plus ou moins élaboré, permet d'adapter la taille d'échantillon aux contraintes de précision, mais aussi de coût (voir 3.3.3).

- Le choix judicieux d'un plan d'échantillonnage (compatible avec la formalisation mathématique) permet d'améliorer la précision et/ou de réduire le coût (voir 3.3.4).
- Les étapes précédentes aboutissent à la caractérisation de l'échantillonnage en associant une précision et un coût à la taille d'échantillon (voir 3.3.5).
- Enfin, la stratégie d'échantillonnage obtenue est validé et appliquée, ou invalidée et retravaillée (voir 3.3.6).

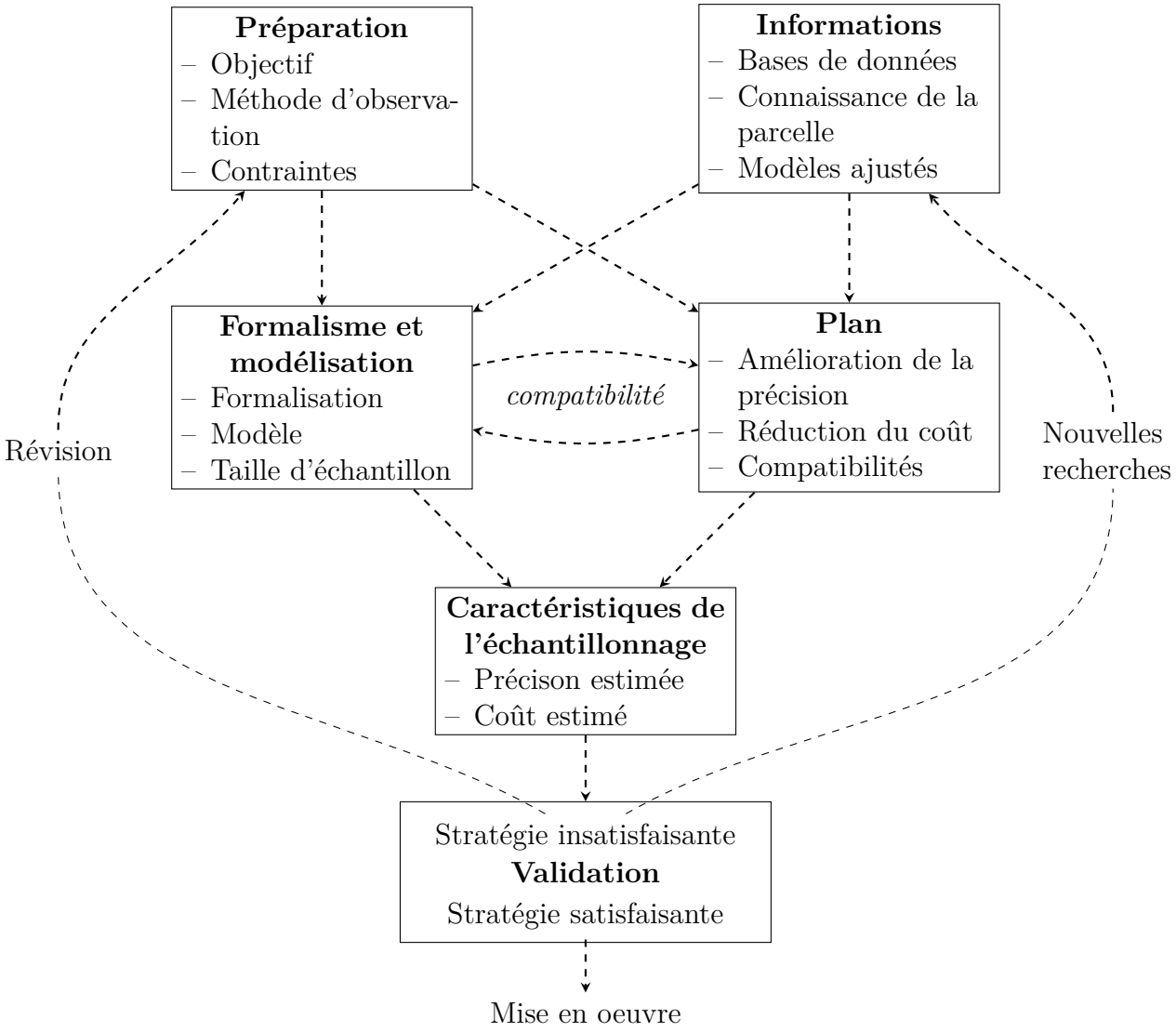

FIGURE 3.2.: Fonctionnement schématisé de l'outil d'aide à la conception de stratégies d'échantillonnage proposé

3.3. L'outil d'aide à la conception de stratégies d'échantillonnage

En répondant aux questions suivantes, et en effectuant les différentes tâches avec l'aide des explications détaillées sur les stratégies d'échantillonnage, on pourra établir une stratégie d'échantillonnage pour la caractérisation d'un stress biotique ou d'une régulation biologique. Chaque groupe de questions et tâches correspond à un item du schéma de principe de l'outil (figure 3.2). Une réflexion préparatoire est d'abord nécessaire, puis les étapes intermédiaires peuvent être menées en parallèle, tout en veillant à la cohérence. Enfin, après une voire plusieurs itérations des étapes précédentes, une stratégie d'échantillonnage est obtenue.
Si une stratégie d'échantillonnage éprouvée existe déjà pour la caractérisation qui doit être réalisé, elle peut être analysée au vu de l'outil proposé. La précision des résultats obtenus peut aussi être calculée de manière à savoir si le coût pourrait être réduit ou si au contraire plus d'efforts doivent être fournis.

3.3.1. Préparation

Poser les bases de la réflexion sur l'échantillonnage.

3.3.1.1. Objectif

- ▶ Est-ce que la question posée rentre dans le cadre de cet outil ?
  C'est à dire est-ce qu'elle revient à déterminer l'intensité globale sur une parcelle d'un stress biotique ou la présence d'un auxiliaire ?
  Non Le reste de la démarche de conception d'une stratégie d'échantillonnage peut tout de même être suivi pour obtenir des pistes de réflexion.
  Oui Il faut maintenant préciser l'objectif.
- ▶ **Tâche :** Définir un objectif précis de l'échantillonnage (voir 2.2). S'il y a plusieurs objectifs, la conception de stratégie d'échantillonnage peut être poursuivie séparément pour chaque objectif. Ensuite l'organisation pourra être adaptée pour réaliser les échantillonnages plus ou moins simultanément.

3.3.1.2. Méthode d'observation

- ▶ Est-ce qu'*une* méthode d'observation est fixée ?
  Aucune méthode n'est envisagée Consulter un expert ou la littérature sur les méthodes courantes. Les plus courantes seront mieux documentées ce qui simplifiera les autres choix.
  Il y a plusieurs méthodes envisagées La conception de stratégies d'échantillonnage peut être poursuivie séparément pour chaque méthode, ensuite elles pourront être comparées au vu des conclusions obtenues.
  Oui étape suivante
- ▶ Est-ce que la méthode d'observation est fiable (voir 2.3) ?
  Non L'utilisation de modèles sera hasardeuse à cause des perturbations générées.
  Oui L'utilisation de modèles sera possible.

3.3.1.3. Coût acceptable

- **Tâche :** Estimer le temps nécessaire par observation pour la méthode d'observation choisie. En déduire un nombre maximal d'observations envisageables.
- **Tâche :** Estimer si le temps de déplacement au sein de la parcelle sera important ou non comparé au temps nécessaire pour une observation.

3.3.2. Informations

- Est-ce que ce type de caractérisation a déjà été mené avec des résultats et un coût satisfaisants ?
  Oui On peut quand même étudier la stratégie utilisée pour voir de potentielles améliorations.
  Non, le coût était trop élevé Voir avec l'outil comment réduire le coût sans que la précision devienne problématique.
  Non, la précision était insuffisante Voir avec l'outil comment améliorer la précision sans que le coût ne devienne problématique.
- Est-ce que des études de l'agrégation ou de la corrélation spatiale du stress biotique ou de la présence d'auxiliaire existent ?
  Oui Ces informations permettront d'utiliser des modèles spécifiques ou d'adapter l'espacement des observations dans l'espace.
  Non Des connaissances sur la biologie du phénomène étudié peuvent être mises à profit pour évaluer grossièrement l'agrégation ou la corrélation.
- Est-ce que des modèles existent pour le stress biotique ou la présence d'auxiliaire ?
  Oui Ils pourront guider les choix de modélisation, voire être réutilisés avec les mêmes paramètres dans certain cas.
  Non Des bases de données issues d'échantillonnages similaires peuvent être recherchées pour tester ou ajuster un éventuel modèle.
- **Tâche :** Rassembler des informations sur les conditions environnementales et les pratiques culturales qui pourraient influencer le phénomène observé.
- **Tâche :** Essayer d'obtenir une pré-estimation de résultats qui vont être obtenus grâce à l'échantillonnage, elle permettra souvent d'optimiser la taille de l'échantillon. Des mesures passées, des mesures sur d'autres parcelles ou encore des modèles épidémiologiques peuvent être utilisés.

3.3.3. Formalisme, 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.

3.3.3.1. Formalisation

- **Tâche :** Formaliser l'objectif selon la typologie de variables proposée (2.4) et choisir un indicateur de précision.

Le formalisme proposé ne convient pas à la situation Il faudra se passer de cette partie et optimiser la répartition des observations dans la parcelle.

Un des type de données et un des indicateur de précision conviennent Les modélisations possibles peuvent maintenant être étudiées.

3.3.3.2. Modèle

▶ **Tâche :** Choisir un ou des modèles (voir 2.6) en fonction des informations disponibles pour les paramétrer et les valoriser et des connaissances sur l'agrégation spatiale qui ont pu être rassemblées.

▶ Le modèle et le plan 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.

3.3.3.3. Taille d'échantillon

▶ 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 (voir 2.5).

Oui En déduire la plus petite taille d'échantillon permettant d'obtenir la précision désirée.

Non Choisir entre :

- Échantillonner sans contrôler la précision, avec une taille d'échantillon choisie pour sa faisabilité (voir 2.5.1).
- Mettre en place un échantillonnage adaptatif (voir 2.5.2.2) pour ajuster le nombre d'observations au cours de l'échantillonnage.

▶ Est-ce que la taille d'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.

3.3.4. Plan

L'objectif ici est de bien prendre en compte, pour le choix des unités statistiques à examiner, les différentes hétérogénéités spatiales que l'on peut supposer exister pour la variable, ainsi que le coût d'échantillonnage.

3.3.4.1. Amélioration de la précision

▶ Est-ce que des informations sont disponibles sur d'éventuelles hétérogénéités spatiales de la variable observée (voir 2.7) ?

Non Privilégier un échantillonnage systématique (voir 2.8.3.2), sauf si on souhaite mettre en place un échantillonnage séquentiel.

Oui – Si des zones plutôt homogènes peuvent être repérées, s'intéresser à l'échantillonnage stratifié (voir 2.8.4.1).
– En cas de corrélation spatiale notable, espacer les observations par un échantillonnage systématique (voir 2.8.3.2).
– Consulter les explication sur les hétérogénéités spatiales (voir 2.7) pour plus de détails et des conseils supplémentaires.

3.3.4.2. Réduction du coût

▶ Est-ce que le coût des déplacement dans la parcelle est élevé par rapport au coût des observations
Oui S'intéresser aux échantillonnages par grappes (voir 2.8.4.2), selon un parcours (voir 2.8.3.2) et à plusieurs degrés (voir 2.8.4.3).
Non En profiter pour bien espacer les observations, par exemple avec un échantillonnage systématique (2.8.3.2).
▶ **Tâche :** Si chaque observation demande une analyse coûteuse (temps ou matériel), évaluer la possibilité d'analyser des échantillons composites (voir 2.8.5), éviter l'échantillonnage par grappes.
▶ **Tâche :** Si le coût d'échantillonnage est très variable au sein de la parcelle, évaluer la possibilité de stratifier (2.8.4.1) la parcelle selon le coût.

3.3.4.3. Compatibilités

▶ Est-ce qu'un modèle imposant un échantillonnage par grappe a été choisi ?
Oui L'échantillonnage par grappes simple peut être choisi, les grappes peuvent aussi être disposées de manière stratifiée ou systématique dans la parcelle.
Non Le modèle s'appliquera indépendamment du plan choisi.
▶ Est-ce qu'une démarche adaptative est envisagée ?
Oui Éviter l'échantillonnage systématique selon une grille (voir 2.8.3.2), sauf cas particuliers.
Non Le plan aléatoire simple n'aura aucun intérêt.

3.3.5. Caractéristiques de l'échantillonnage

▶ **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, calculer si nécessaire :
– La précision compte tenu du coût maximal acceptable
– Le coût compte tenu de la précision minimale acceptable
– Les diagrammes pour l'approche séquentielle
▶ **Tâche :** Choisir une taille d'échantillon ou une approche adaptative.

3.3.6. Validation

▶ Est-ce que le coût et la précision prévus semblent raisonnables ?
Oui La stratégie élaborée est prête à être mise en pratique.

Le coût est trop élevé Revoir les optimisations possibles, les objectifs de précision puis le choix de la méthode d'observation.

La précision est trop faible Revoir les optimisations possibles, les limitations de coût puis le choix de la méthode d'observation.

3.4. Tests

Si l'outil avait été un outil d'aide à la décision plus automatisé, qui fournit des consignes précises à partir de données d'entrée, sans pour autant expliciter la réflexion sous-jacente, alors des tests représentatifs de tout les cas pouvant être rencontrés auraient été nécessaires. Toutefois, l'outil proposé est plus proche d'un guide de réflexion alimenté de conseils issus d'une large synthèse bibliographique. La qualité des informations est donc garantie par les sources, et des tests doivent plutôt servir à vérifier l'adéquation des situations couvertes par l'outil avec celles rencontrées en pratique. Pour cela, quelques exemples sont traités dans le chapitre 4.
Les quelques jeux de données analysés permettent de montrer comment certains modèles proposés peuvent être appliqués. Ils permettent aussi de mettre en évidence des cas où la répartition spatiale des observations est intéressante.

4. Exemples

4.1. Taux de prédation de graines

Dans le cadre de l'étude de la prédation de graines par des carabes, des cartes de prédation ont été disposées sur deux parcelles, avec deux types de graines à chaque fois [Trichard et al., 2014]. Ce travail a produit une base de donnée des taux de prédations observés en 33 stations géo-référencés de chaque parcelle repartis comme sur la figure 4.1. Huit sessions d'observations ont eu lieu.

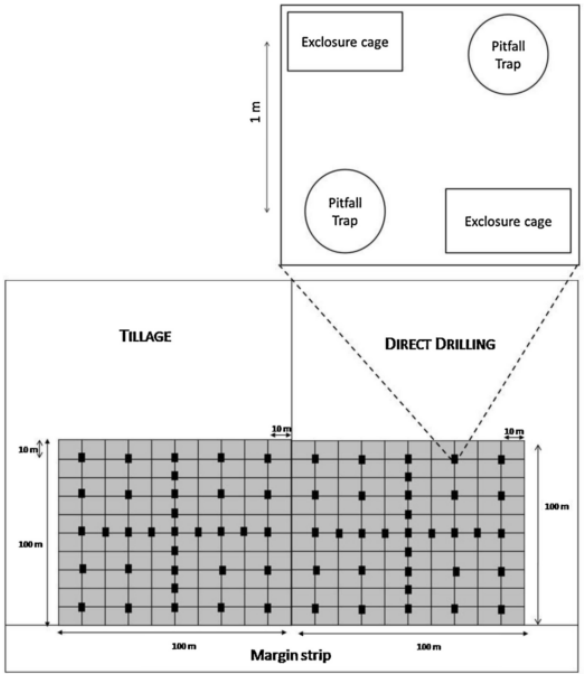

FIGURE 4.1.: Stations d'observation pour l'étude de la prédation de graines

A partir de ces données, des pistes peuvent être proposées concernant la taille et la répartition des observations qui seraient nécessaires pour bien estimer le taux de prédation moyen p de graines sur une parcelle.

Des variogrammes (tracés avec le package `geoR`, Jr and Diggle [2015]) ne montrent pas de corrélation spatiale pour des distances allant de $10\,m$ à $100\,m$ (figure 4.2). On peut faire deux hypothèses à partir de ce résultat. Premièrement, on aurait eu un résultat comparable avec des observations placées aléatoirement (mais en essayant de garantir une distance minimale de l'ordre de $10m$ entre deux observation). Deuxièmement, l'estimation s^2 faite de la variance n'est

pas biaisée, et le taux moyen calculé a une précision que l'on peut caractériser par la variance $\frac{s^2}{N}$.

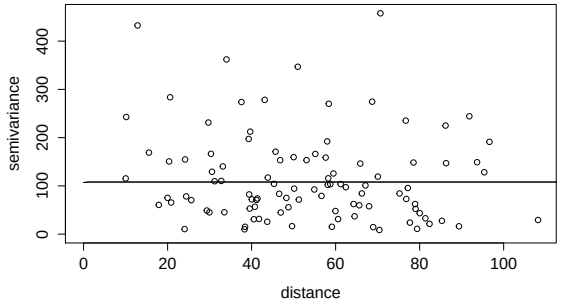

FIGURE 4.2.: Variogramme pour un groupe de 33 observations simultanées

Bien que les données soient des taux moyens de prédation sur deux cartes de la même station, on peut considérer, du fait du protocole, qu'elle représentent des nombre de graines consommés parmi 100 graines. On peut donc utiliser les modèles pour la variables entières à valeurs bornées (voir 2.4.3.2).
En représentant sur un graphe les variance observées et les variances théoriques selon la loi binomiale, on constate que ni la loi binomiale, ni une loi de puissance entre les deux variance ne serait pertinente. En effet, la loi binomiale 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 4.4) 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 le résultats des huit sessions (figure 4.3), de cette façon on obtient un ordre d'idée de la qualité de modélisation qui peut être obtenue.

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 (voir 2.6.3.3). Au vu des graphiques 4.5, la variance estimée à partir du taux de prédation et du paramètre ρ, pour chaque session et chaque modalité, est 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 car chaque observation demande la pose d'une carte de prédation et son relevé sept jours plus tard. La taille d'échantillon pourrait plutôt 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,...).

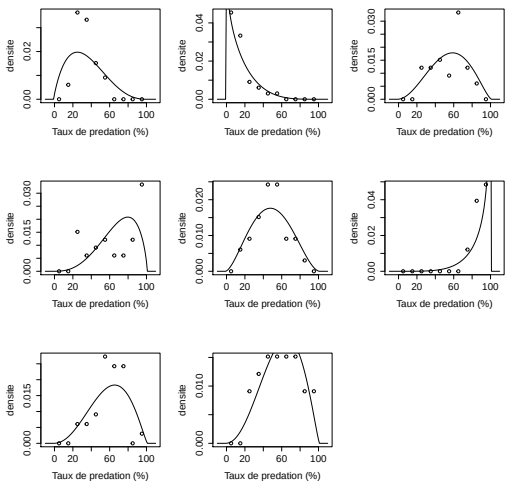

FIGURE 4.3.: Loi bêta-binomiale ajustée aux 8 sessions de mesure pour une des modalités, avec le même paramètre de sur-dispersion ρ

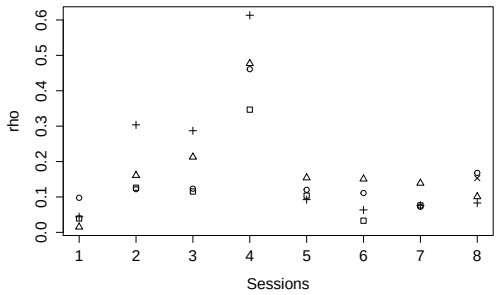

FIGURE 4.4.: Valeurs estimées du paramètre de sur-dispersion de la loi bêta-binomiale pour les quatre modalités (symboles distincts) et pour les 8 sessions de mesures

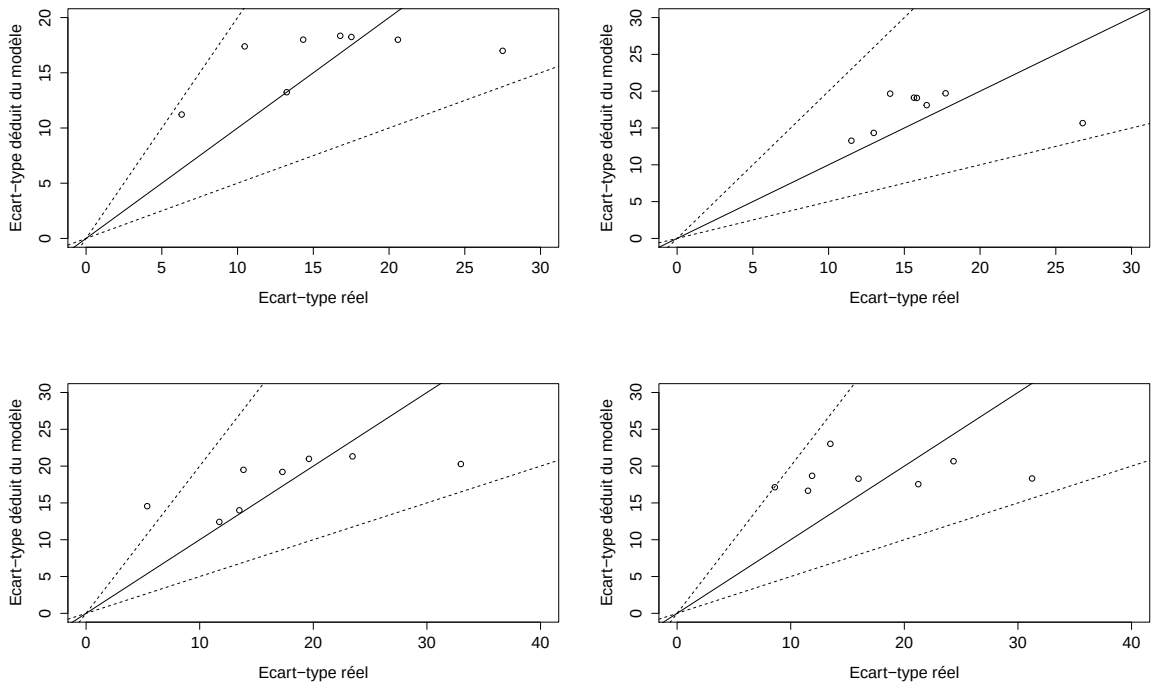

FIGURE 4.5.: Ecart-types (en %) réels et déduits à partir du modèle pour les taux de prédation selon quatre modalités expérimentales, pour 8 sessions de mesure par modalité. Les trois droites, de pentes 0.5, 1 et 2 indiquent l'erreur réalisée.

4.2. Étude des données de piégeage de carpocapses

Cet exemple est basé sur des données fournies par Claire Lavigne, de l'unité PSH (Plantes et Systèmes de culture Horticoles) de l'INRA Avignon. Ce sont les résultats de piégeage de carpocapses par des pièges répartis de manières systématiques dans quarante huit vergers au sud d'Avignon (figure 4.6). Les résultats de comptage, qui indiquent pour chaque verger et

FIGURE 4.6.: Disposition des pièges à insectes dans un verger

chaque piège le nombre d'insectes capturés, se caractérisent principalement par le très grand nombre de zéros. On peut ajuster une loi binomiale négative si on rassemble les données de tous les vergers mais ça n'a qu'un sens limité quand on sait qu'il sont très hétérogènes. D'autant plus que 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 4.7). Il s'agit de faire le lien entre la moyenne et la variance observées dans chaque verger avec une relation de la forme $s^2 = a \cdot m^b$. On obtient ici les approximations $a = 2.55$ et $b = 1.379$ par régression linéaire sur le logarithme des données.

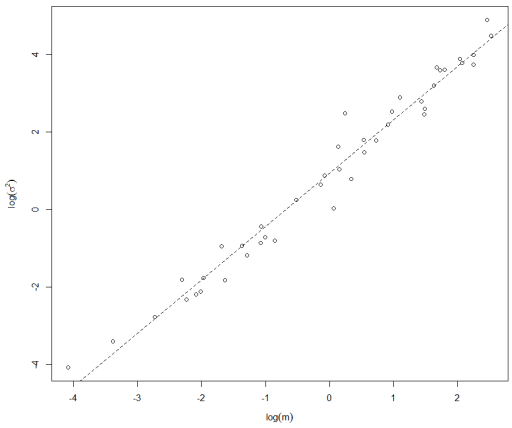

FIGURE 4.7.: Ajustement d'une loi de puissance de Taylor entre la moyenne et la variance des résultats de piégeage dans 48 vergers

La relation ajustée permet, à partir d'une estimation grossière de la moyenne du nombre d'insectes piégés sur une parcelle, de déduire la variance de ce nombre. 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. Un nombre minimal d'observations $N_{\min}$ à réaliser peut être calculé en fonction de l'objectif de précision souhaité. 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 qu'il y aura assez d'observations pour que la moyenne suive une loi normale, alors il faudrait prendre

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

A propos de piégeage, parler du nombre total d'insectes sur la parcelle n'est pas pertinent. En effet, les données récupérées sont une combinaison de la densité d'insectes et de l'efficacité des pièges sans qu'il soit possible de distinguer l'une de l'autre. L'objectif sera donc naturellement de bien 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 strates de petite taille avec peu d'observation par strates (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 toutes façon inadapté à un protocole avec des pièges qui doivent rester en place longtemps avant de donner un résultat.

4.3. Sévérité du Phoma du Colza

Dans le cadre d'une étude sur la répétabilité des mesures de sévérité pour le phoma du colza [Aubertot et al., 2004], la sévérité de la maladie sur toutes les plantes de trois zones de $9m^2$ à été relevée. Elle est notée selon six classes de sévérités, auxquelles sont 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 jusqu'à $3m$, comme on le voit sur la figure 4.8. Ce résultat corrobore ceux obtenu pour des distances plus importantes par Aubertot et al. [2004].

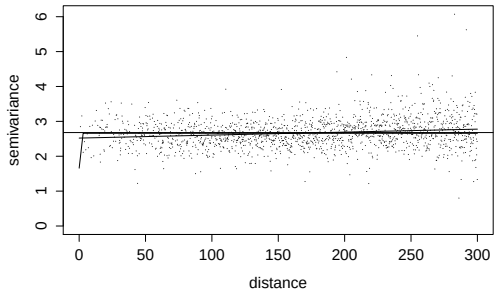

FIGURE 4.8.: Nuée variographique et variogrammes ajustés pour le score de sévérité du phoma du colza pour des distances de $20cm$ à $300cm$

La qualité de l'échantillonnage ne serait donc pas dégradée par des prélèvements de plantes en grappes. De plus, la précision associée à la proportion de chaque classe sera bien estimée par la loi multinomiale, de même que la précision pour un score moyen. Ce ne serait pas le cas s'il y avait une corrélation spatiale. Par exemple, le coefficient de variation 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 serait

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

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

Pour obtenir un précision donnée par $c_v = 0.05$, il faudrait donc 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 qui correspond à ce cas.

Finalement, on peut faire trois recommandations pour obtenir le score moyen de sévérité du phoma sur une parcelle (respectant l'absence de corrélations spatiales) tout en vérifiant l'absence de motifs à plus grande échelle.

- Les observations peuvent êtres placées aléatoirement où selon n'importe quel motif couvrant plutôt bien la parcelle (en U par exemple).
- Si les plantes sont prélevées par grappes (8 grappes de 8 par exemple), 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.

5. Discussion

Répondre à un besoin réel Alors que le terme agroécologie est de plus en plus utilisé, et qu'il se charge d'un sens politique, un travail important reste à faire pour comprendre et caractériser les phénomènes biologiques à la base d'une pratique plus écologique de l'agriculture. En particulier, le besoin de mieux caractériser la composante biotique des agroécosystème persiste, et il y a un enjeux important d'amélioration du recueil de données exploitables scientifiquement. Le projet CASIMIR contribue aux améliorations et aux innovations nécessaires, avec la préoccupation, entre autres, d'obtenir une vision globale de la composante biotique des parcelles agricoles. Pour cela, le développement de stratégies d'échantillonnage les moins coûteuses possibles est décisif, et un échantillonnage simultané pour les principaux bioagresseurs sur une parcelle est à envisager.
L'attention particulière dédiée à l'optimisation des stratégies d'échantillonnage a vu sa pertinence confirmée au cours du stage. Premièrement, les divers échanges au sujet du stage ont suscité un intérêt important de la part des interlocuteurs, et ils ont fait ressortir le besoin d'informations (basiques surtout) concernant l'échantillonnage et la précision des résultats obtenus. Deuxièmement, des questions sur l'échantillonnage pour la caractérisation de la composante biotique d'agroécosystèmes, transmises par des chercheurs du réseau expérimental Rés0Pest (voir annexe C), ont été traitées en marge du stage ; une bonne partie de problématiques rencontrées sont traitées dans ce rapport. Enfin, le constat que des stratégies d'échantillonnages sont souvent reprise d'une année sur l'autre, ou de publications passées, sans que la précision et le coût ne soient évalués, est récurrent.

Approfondissement et ressources complémentaires Malgré sa longueur déjà importante, ce rapport 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.

Limites du stage Au cours du stage il a semblé que la conception de stratégies d'échantillonnage pourrait être plus automatisée ; un programme informatique aurait alors pu prendre en entrée une série d'informations sur les déterminants d'une stratégie, et aurait renvoyé en sortie une stratégie détaillée (taille d'échantillon et localisation des observations au sein de la parcelle). Le besoin de compromis et de subtilités, ainsi que le nombre de déterminants ont rendu cette tâche inabordable. Le résultat obtenu finalement est plus proche d'un guide, intégralement contenu dans ce rapport.
Pour s'approcher un peu plus de l'objectif final du projet CASIMIR, il aurait été intéressant de développer le lien entre le choix des protocoles d'observation, la taille d'échantillon et le plan d'échantillonnage. Toutefois, il était difficile de travailler sur ce sujet du fait du peu d'information disponible sur les protocoles dont certains sont encore de cours de test. Une convergence

reste à faire entre les différentes composantes du projet. Notamment, le résultat des tests des différents protocoles d'observation seront une bonne source de données par évaluer la pertinences de modèles, et éventuellement en ajuster.

Conclusion

Réduire l'usage des produits phytosanitaires pour l'agriculture est aujourd'hui un enjeux clé pour des raisons sanitaires et environnementales. Cela passe, entre autres, par une meilleure connaissances des agroécosystèmes. Dans ce cadre, le projet CASIMIR développe 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 méthodologiques ; ce stage avait pour but de faciliter la mise au point des stratégies en proposant des outils adéquats.
Une large synthèse bibliographique a permis de dégager les déterminants d'une stratégie d'échantillonnage puis de décrire étape par étape la façon de les prendre en compte. Ce travail a aboutit à la création d'un *outil d'aide à la conception de stratégies d'échantillonnage* adapté aux contexte de la caractérisation de la composante biotique des agroécosystèmes. L'outil se compose de questions et tâches qui, en s'appuyant sur les explications rédigées, guident l'élaboration d'une stratégie d'échantillonnage. Il est amené à évoluer après le stage de façon à être utilisé et diffusé le plus largement possible. Il pourra être mis en ligne, sous sa forme actuelle et sous une forme plus informatisée, sur le site *Quantipest* pour compléter les ressources qui s'y trouvent déjà. Il pourrait aussi être édité sous forme de guide.

Bibliographie

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. 2.2, 2.3.1, 2.4.3.4, 2.4.3.4, 2.6.3.5, 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. 2.2, 2.7.1, 2.8.4.3

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. 2.2, 2.7.1, 2.7.2

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. 2.6.4.3, 2.7.1

S. J. Clark, P. Rothery, J. N. Perry, and M. S. Heard. Farm Scale Evaluations of herbicide-tolerant 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. 2.2, 2.8.4.3

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. 2.7.3, 2.7.3

David Collett. *Modelling Binary Data*. Chapman and Hall, 1991. ISBN 0-412-38790-5. 2.7.5

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. 2.7.5, 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. 2.1, 2.1.1, 2.2, 2.7.2, 2.7.2, 2.7.3, 2.8.3.2, 2.8.4.1, 2.8.4.2, 2.8.4.3, 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. 3

Sundar Dorai-Raj. *binom : Binomial Confidence Intervals For Several Parameterizations*, 2014. URL http://cran.r-project.org/package=binom. 2.6.3.2

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. 2.7.2

Timothhy G. Gregoire and Harry T. Valentine. *Sampling Strategies for Natural Resources and the Environment*. Chapman and Hall/CRC, 2007. ISBN 9781584883708. 2.7.2, 2.8.3.2, 2.8.4.1, 2.8.4.1, 2.8.4.1

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. 2.3.1

IPM Network. Quantipest Website, 2013. URL http://www6.inra.fr/quantipest. 1.2.1, 1.4, 2.6.3.5

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. 2.6.3.4, 2.6.4.3

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. 2.6.3.4, 2.6.4.3

L V Madden and Gareth Hughes. Sampling for Plant Disease Incidence. *Phytopathology*, 89 : 1088–1103, 1999. 2.4.3.1, 2.4.3.2, 2.6.3.3, 2.6.4.2, 2.6.4.2, 2.7.5, 2.7.5, 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/. 1.2.1, 2.4.3.4, 5

Ministère de l'Agricutlure et de la Pêche. Réseau DEPHY-FERME Synthèse des premiers résultats à l'échelle nationale. Technical report, 2014. 1.1.3

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. 2.2, 2.6.3.4

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. 2.6.3.4

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. 2.4.2, 2.6.4.1, 2.6.4.3, 2.7.1

F. W. Nutter, Jr. Assessing the Accuracy, Intra-rater Repeatability, and Inter-rater Reliability of Disease Assessment Systems, 1993. ISSN 0031949X. 2.3.1

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. 2.3, 2.6.4.1

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. 2.7.1

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. 2.2, 2.6.3.4

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. 2.2, 2.2

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/. 2.6.3.2

L. J. Rew and R. D. Cousens. Spatial distribution of weeds in arable crops : Are current sampling and analytical methods appropriate ?, 2001. ISSN 00431737. 2.6.3.4, 2.7.5

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. 2.4, 2.5.2.2

M. Sebillotte. Système de culture, un concept opératoire pour les agronomes. In *Les systèmes de cultures*, pages 165–196. INRA, 1975. 1.1.2

Lionel Roy Taylor. Assessing and Interpreting the Spatial Distributions of Insect Populations, 1984. ISSN 00664170. 2.6.4.1, 2.7.4, 2.7.4, 5

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

Todd a. Ugine, John P. Sanderson, Stephen P. Wraight, Les Shipp, K. Wang, and Jan P. Nyrop. Binomial Sampling of Western Flower Thrips Infesting Flowering Greenhouse Crops using Incidence-Mean Models, 2011. ISSN 0046-225X. URL http://ee.oxfordjournals.org/content/40/2/381.

Jean Vaillant. Notions de base en échantillonage, 1996. URL http://www7.inra.fr/mia/ftp/V/FPstat/module6/version1/Vaillant.pdf. 2.7.5, 2.9, 2.10, 2.11, 2.8.4.1, 2.12, 2.8.4.2, 2.13, 3, 3.1, 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. 2.6.3.4

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. 2.8.5

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. 2.6.3.5

Hsiuying Wang. Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, 99(5) :896–911, May 2008. ISSN 0047259X. doi : 10.1016/j.jmva.2007.05.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X07000784. 2.6.3.5

Edward Kyle Waters, 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. 2.6.4.3

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. 2.7.1

Thomas W Yee. *VGAM : Vector Generalized Linear and Additive Models*, 2015. URL http://cran.r-project.org/package=VGAM. 2.6.3.3, 4.1

Annexes

A. Statistiques

A.1. Source d'aléa pour l'échantillonnage

Une approche statistiques nécessite que les données obtenues soient théoriquement aléatoires, c'est-à-dire qu'elles sont composées de variables aléatoires. On peut soit considérer que l'aspect aléatoire provient de la procédure d'échantillonnage, soit considérer que le phénomène observé suit un modèle aléatoire. Cette distinction est importante pour justifier théoriquement l'utilisation de certain modèles.

A.2. Estimateurs

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).

A.3. Estimation de la variance

L'estimation de la variance d'une variable échantillonnée est en général très imprécise car celle-ci requiert une taille d'échantillon importante.

Variable suivant une loi normale Dans le cas où la variable étudiée suivrait 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. On peut voir sur deux exemples l'ordre de grandeurs de la taille d'échantillon nécessaire pour avoir une bonne estimation 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 ne sera entre 0.8 et 1.3 que dans 75% des cas (sur des répétitions nombreuses de l'échantillonnage aléatoire). Pour 1000, l'estimation sera entre 0.93 et 1.07 dans 90% des cas.

Variable ne suivant pas une loi normale Pour une variable de loi continue mais asymétrique, un transformation logarithmique peut permettre de ce ramener au cas précédent.

A.4. Remarque sur l'explication des hétérogénéités

Quand en pratique on observe une hétérogénéité, de la présence d'un organisme auxiliaire ou d'une pression biotique, dans une parcelle, 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 de modélisation différentes n'impliquent pas forcément les mêmes choix de stratégie d'échantillonnage.
Ce n'est pas vraiment un problème ici car l'objectif est de prévoir des hétérogénéités (à partir de connaissances disponibles *a priori*), pas 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). On n'ira pas jusqu'à étudier un phénomène d'agrégation sur la parcelle pour y ajuster un modèle car cela demande un effort d'échantillonnage trop important, on pourra par contre se baser sur des modèles existants d'agrégation pour optimiser la taille de l'échantillon.

A.5. Test des modèles de distributions discrètes ajustés

Un 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 un 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 : $T \geq \chi^2_{J-1,1-\alpha}$.

A.6. 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

TABLE A.1.: Quantiles de la loi de Student

B. Code R

B.1. 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
```

C. Échanges sur l'échantillonnage pour Rés0pest

Voilà quelques réponses à des questions soulevées lors du séminaire Rés0pest. Elles reflètent le travail de mise en place d'un outil d'aide à la décision pour les stratégies d'échantillonnage des stress biotiques, que je développe pendant mon stage sous la direction de Jean-Noël Aubertot (pour le projet CASIMIR).

Eloi Navarro - eloi.navarro@toulouse.inra.fr

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

Il est possible que la répartition des plantes adventices soit plus homogène quand leur densité est élevée (puisqu'on aurait moins de chance de tomber sur une zone «vide» par exemple). Si c'est bien le cas, des quadrats plus petits peuvent être choisis quand la densité est plus élevée.

Par contre il vaut mieux éviter de réduire le nombre de quadrats , de façon à conserver une bonne idée de la variance et donc de la précision.

Q : 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.

Soit un score peut être attribué à chaque classe en fonction par exemple de la perte de rendement attendue, et la moyenne est faite sur le score (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/Sampling-strategies/How-to-define-the-sample-size/N-Index-Calculation-of-minimum-sample-size-for-an-index). C'est peut-être la meilleure solution ici, 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 valeur au milieu de la classe. Dans le cas ou on n'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) ?

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

1- Un malherbologue aura certainement un avis à donner 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 grand 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).

Protocole chanvre : 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).

Q : Protocole chanvre : pourquoi 10 plantes consécutives pour noter les maladies à floraison ?

Parce que c'est plus pratique à choisir que 10 plantes au hasard dans le rang. Ça peut aussi être utile pour avoir plusieurs valeurs de comptage/pourcentages sur lesquelles faire un traitement statistique. Dans l'absolu il vaut mieux espacer les observations pour que l'agrégation spatiale de la maladie de 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.
+voir remarque à la fin

Q : Protocole céréales à paille : 10 plantes prélevées pour réaliser la détermination du stade épi 1 cm sont-elles suffisantes en termes de représentativité ?

Il faudrait voir l'écart-type observée en pratique sur la distance mesurée. On peut en déduire un intervalle de confiance à 95% approximatif pour la moyenne (IC =m+/-(2 x etype/racine(nbplantes)), au pire il faut 5 min pour faire le calcul avec une petite calculatrice) et voir en gros ci ça convient. A ce moment on peut 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). On pourrait aussi faire un test statistique pour savoir si on est bien sûr de dépasser le seuil de 0,8cm mais c'est un peu plus compliqué.

Q : protocole Céréales à paille : mesures maladies en foyers, combien de plantes ? Difficulté de cerner le périmètre atteint ?

Si on est placé à peu près au milieu du foyer, on peut supposer que les comptages de plantes malades seront assez homogènes, c'est pour ça que seulement 3x10 plantes sont conseillées ici. 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 atteint il n'y a pas de méthode vraiment rapide, l'observateur pourrait peut-être parcourir la parcelle en essayant de déterminer la proportion de la distance (ou du temps) pendant laquelle il est dans des foyers. Comme c'est assez subjectif on ne peut pas s'attendre à une bonne précision.

Q : tous protocoles : quelle 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 du 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.

Q : protocole Céréales à paille : mesures maladies et ravageurs ? Prendre un seul échantillon / station pour tous les bioagresseurs ou X échantillons pour X bioagresseurs ?

Pas de raison (statistiquement) de prendre des échantillons différents pour les différents bioagresseurs, sauf si des examens sont destructifs et incompatibles.

Remarque : Comme c'est souvent le cas, les protocoles proposés demandent d'enregistrer une note sous forme de classe ou de moyenne, mais sans conserver le détail des comptages. On perd au passage toute idée de la précision avec laquelle la note (ou moyenne) est estimée, alors que cette précision pourrait être utile pour l'étude des résultats. Vous pourriez donc envisager, en fonction de ce que vous comptez faire des données, de calculer un indicateur de précision à la fin des mesures (c'est par exemple très facile dans le cas des classes 3 et 4 du protocole pour le chanvre, page 7/15).