THESE

EN VUE DE L'OBTENTION DU GRADE DE

DOCTEUR DE L'UNIVERSITE DE PAU ET DES PAYS DE L'ADOUR

DISCIPLINE

PHYSIOLOGIE et BIOLOGIE DES ORGANISMES-POPULATIONS-INTERACTIONS

PRESENTEE ET SOUTENUE PAR

Claire KERMORVANT

le 19 novembre 2019

Optimisation de procédures d'échantillonnage appliquées aux suivis de la biodiversité et des ressources

---

Jury :

Aurélien BESNARD, Maître de conférences, HDR, CEFE/CNRS	Rapporteur
Stéphane DRAY, Directeur de recherche, Université Lyon 1	Rapporteur
Nicolas BEZ, Directeur de recherche, IRD	Examinateur
Nathalie CAILL-MILLY, Cadre de recherche, Ifremer	Co-encadrante
Frank D'AMICO, Maître de conférences, HDR, UPPA	Directeur de Thèse
Benoît LIQUET, Professeur, UPPA	Examinateur

”Work it Harder, Make it Better
Do it Faster, Makes us Stronger
More than Ever, Hour After
Our Work is Never Over.”

Daft Punk

Remerciements :

Je tiens tout d'abord à exprimer ma profonde reconnaissance à mes deux directeurs de thèse Frank D'Amico et Noëlle Bru pour m'avoir accordé leur confiance, guidée et encouragée tout le long de ce projet. Merci pour cette grande autonomie que vous m'avez laissée dans ce travail mais aussi pour vous être toujours rendus disponibles quand j'en ai eu besoin.

Un grand merci à la Communauté d'Agglomération Pays Basque pour avoir appuyé et financé ce projet.

J'exprime aussi mes sincères reconnaissances à l'Université de Pau et des Pays de l'Adour, et notamment à l'équipe du Laboratoire des Mathématiques et de leurs Applications de Pau, pour m'avoir accueillie et accompagnée durant la thèse.

Je tiens aussi énormément à remercier l'équipe du Laboratoire Environnement Ressources d'Arcachon LERAR de l'Ifremer à Anglet. Merci de m'avoir offert ce stage de master 2 qui a débouché sur ma thèse. Un très grand merci à Nathalie, Florence, Muriel, Marie-Noëlle et Gilles pour m'avoir accueillie dans votre équipe et trainée tel un boulet à toutes les pauses café (voire plus).

Bien sûr, il me tient aussi à coeur de remercier tous mes copains du bureau : Alyssa, Laura, Carole, Aurore, Jonathan et Tiphaine ! Mais aussi le meilleur copain de café : Alexis. Merci à tous pour votre bonne humeur journalière et pour avoir fait passer ces trois années aussi rapidement.

Merci à toute l'équipe du Collège STEE d'Anglet pour les pauses, les ragots, les potins, les pique-niques dans le parc, les cafés, les sorties ski, les sorties fêtes de Bayonne… pour les bonjours et sourires journaliers qui font que l'on passe tous les jours une bonne journée.

Il me faut aussi remercier mon copain et ma famille, qui, surtout ces derniers mois, on su rester calmes et patients à ma place. Merci de m'avoir supportée et poussée tout au long de mes études pour me permettre d'aller toujours plus de l'avant.

J'adresse aussi mes remerciements à toutes les personnes qui m'ont aidée, de près ou de loin, durant cette thèse.

Enfin, mille mercis à Marcel et Jeannette, qui ont aidé à faire naître en moi la passion pour l'environnement au travers d'élevages, de cabanes et de jeux enfantins.

Résumé :

Cette thèse s'intègre dans un contexte où les méthodes utilisées pour la mise en place de suivis environnementaux sont souvent problématiques et peuvent mener à des résultats controversables. L'objectif est de proposer une méthodologie adaptable à la plupart des suivis environnementaux qui permettra aux utilisateurs de produire des suivis scientifiques efficaces ou d'optimiser des suivis déjà en place. Nous avons développé une méthodologie qui permet à l'utilisateur de fixer la précision qu'il veut sur ses résultats d'estimation et qui lui renvoie un protocole d'échantillonnage optimal associé à un nombre d'unités statistiques à échantillonner. Une fois le nombre de points connu, il est simple d'estimer le coût de mise en place de la procédure d'échantillonnage sélectionnée sur le terrain.

Nous sommes partis de la définition même de la performance d'un protocole d'échantillonnage pour élaborer une méthodologie sous forme de procédure séquentielle qui permet de tester, puis de choisir, le protocole le plus performant pour chaque étude. Plus un protocole d'échantillonnage est performant, moins il nécessite d'unités statistiques pour atteindre une précision voulue. La méthodologie présentée permet donc, pour une (ou des) précision(s) désirée(s) sur les résultats, de déterminer puis de comparer le nombre optimal d'unités statistiques à échantillonner pour différents protocoles. La première étape de la procédure développée nécessite de recréer mathématiquement la population statistique la plus représentative possible de la population étudiée. Ensuite, les différentes combinaisons protocole d'échantillonnage / nombre d'unités statistiques sont simulées puis comparées. Cela permet d'obtenir le meilleur rapport coût-efficacité pour une étude nécessitant un échantillonnage dans un objectif d'inférence, autrement dit, de baisser son prix tout en garantissant une précision adéquate.

Les objectifs de cette thèse ont été atteints : la méthode a été développée puis testée sur trois cas d'études. Le premier est la mise en place d'un suivi efficace lorsqu'il n'existe pas de données disponibles. L'exemple utilisé est celui de la mise en place du suivi du moustique tigre sur l'agglomération de Bayonne-Anglet-Biarritz. L'espèce est en début d'invasion dans cette zone et il n'existe donc quasiment pas de données de suivi. Nous avons récupéré les données de détection dans des villes méditerranéennes, les avons modélisées et avons appliqué le modèle à l'agglomération d'intérêt pour ensuite y définir un protocole de suivi optimal.

Le second cas d'étude est l'optimisation d'un suivi lorsque des données sont disponibles. L'exemple est celui du suivi de la palourde dans le bassin d'Arcachon. Ce suivi est effectué tous les 2 ans depuis 2006, nous avons d'abord travaillé sur une seule année de données et prouvé qu'il était possible d'optimiser ce suivi. C'est-à-dire baisser son coût de 30% en gardant une précision assez bonne sur les résultats pour être en capacité de mettre en place des mesures de gestion adaptées. Nous avons ensuite travaillé sur toutes les données depuis 2006 pour proposer une optimisation de ce suivi pérenne dans le temps.

Abstract :

Developing robust, and reliable, environmental surveys can be a challenge because of the inherent variation in natural environmental systems. This variation, which creates uncertainty in the survey results, can lead to difficulties in interpretations. The objective of this thesis was to develop a general framework, adaptable to environmental surveys, to improve scientific survey-results. We have developed a method that allows the user, by defining their desired level of accuracy for the survey results, to develop an efficient sampling design with a minimal sample size.

Once the sample size is known, calculating total cost of the survey becomes straightforward. We start from the definition of sampling design performance and build a method for comparison and assessment of an optimal sampling design. As a rule of thumb, the more efficient a sampling design is, the fewer statistical units are needed to achieve the desired accuracy. With less sampling effort the sampling procedure becomes more cost effective. Our method assists in identifying cost-efficient sampling procedures.

In this PhD thesis a general methodology is developed, and it is assessed with three case studies. The first case study involved design of an efficient survey when no prior data are available. Here we used the example of tiger mosquito in the Bayonne-Anglet-Biarritz agglomeration (south-west France). This species has only now started invading this area and therefore there are no site-specific data available. We used data from other French Mediterranean's cities to model the probability mosquito are present in the Bayonne-Anglet-Biarritz area of interest. We used this modeled-population to assess, compare and select an effective sampling procedure.

The second case study was for survey optimization when only one season of data are available. The chosen example was from Arcachon bay's manila clam survey in western France. This clam-monitoring has been done biennially (i.e. once every two years) since 2006. We applied our general methodology on one-year data and demonstrated that survey costs can be reduced by 30% a year with no loss of accuracy or reduction in resource management information.

The third case study was based on optimization of a survey when several seasons of data are available. We used the clam surveyed but here as a multi-year dataset. We proposed a long-term spatial and temporal sampling design for monitoring the clam resource.

Table des matières

Table des figures

INTRODUCTION GENERALE

Contexte général et problématique

Un problème récurrent survient généralement lors de la mise en place d'un suivi - que ce soit, par exemple, dans le domaine de la chimie pour étudier la diffusion d'un contaminant, dans le domaine de la sociologie pour étudier des comportements humains ou dans le domaine des sciences politiques pour étudier des intentions de votes - le recensement de toutes les unités statistiques sur lesquelles les données doivent être recueillies est impossible.

Ce recensement serait une procédure complexe à mettre en place (Chiarucci et al. 2003 ; MacKenzie 2006) pour des raisons de coûts (Theobald et al. 2007 ; Jackson et al. 2008 ; Lazarina et al. 2014) et de temps (Cox, Cox, et Ensor 1997). La pratique courante s'applique donc à sélectionner certaines unités statistiques dans la population formant ainsi un échantillon, y mesurer ou observer la (ou les) variable(s) d'intérêt, puis d'inférer ces valeurs à la population de départ sous certaines conditions. Un tel procédé entraîne inéluctablement une erreur d'échantillonnage, quantifiable ou non, liée à la façon dont ces unités statistiques ont été sélectionnées. C'est pourquoi la totalité de la procédure d'échantillonnage doit être pensée *a priori* pour que l'erreur d'échantillonnage *in fine* soit la plus faible possible. Il faut, tout au moins, garantir l'absence de biais et indiquer la précision des résultats vis-à-vis de la population d'origine afin que les décisions prises, qui se fondent sur les résultats fournis par le calcul statistique, puissent avoir du sens et soient robustes. L'absence de biais dans les résultats est apportée par le caractère aléatoire de la sélection de l'échantillon. La précision est amenée par une taille d'échantillons suffisamment grande (Hurlbert 1984) pour rendre acceptable l'erreur associée au résultat final. La somme du biais d'échantillonnage et de la précision de celui-ci est appelé justesse (MacKenzie 2006). La notion d'acceptabilité est en général du fait du commanditaire de l'étude.

La mise en place de la procédure expérimentale pour effectuer un suivi pourrait être apparentée à une science à part entière. Dans le domaine des sciences environnementales, les unités statistiques sont généralement dénommées "sites" ou "échantillons", dispersés dans l'espace et/ou le temps. Toute personne ayant déjà, au moins une fois, eu à produire une procédure d'échantillonnage, s'est potentiellement vue confrontée à un grand nombre d'interrogations : où faut-il placer les sites que l'on va échantillonner ? Combien d'unités statistiques faut-il récolter ? Faut-il visiter chaque saison les mêmes sites ? Quelle unité d'échantillonnage (quadrat, transect ou point) est la plus adaptée ?

En sciences environnementales, bon nombre de ces interrogations ne peuvent pas admettre une seule et même réponse pour tous les cas d'études puisque la technique de l'échantillonnage est utilisée dans le but d'atteindre un grand nombre d'objectifs différents. La plupart des études sont créées pour estimer des états et/ou détecter des changements de tendances (McDonald 2003). L'objectif d'utilisation d'un échantillonnage peut être une description temporelle ou spatiale, voire les deux à la fois. Duncan et Kalton (1987) ont listé les objectifs que pouvait avoir une étude temporelle :

(1) estimer des paramètres de population à des temps ponctuels ou sur des périodes de temps pendant lesquelles le changement est considéré comme négligeable,

(2) estimer des moyennes de paramètres de population sur une période de temps,

(3) mesurer un changement net,

(4) mesurer les changements qu'a subi un individu ou un groupe d'individus,

(5) rassembler des données individuelles sur le long terme,

(6) mesurer la fréquence, le timing, la durée des événements et

(7) accumuler des échantillons au cours du temps, surtout des échantillons rares.

De même, nous pourrions lister les objectifs potentiels d'une étude spatiale :

(1) estimer des paramètres de population sur une surface définie,

(2) estimer des moyennes de paramètres de population sur des aires définies,

(3) mesurer un changement net dans l'espace,

(4) mesurer les changements spatiaux de plusieurs variables,

(5) mesurer l'auto-corrélation spatiale d'une ou plusieurs variables,

(6) accumuler des données spatiales au cours du temps.

À ces objectifs peuvent s'ajouter ceux d'une étude qui serait à la fois spatiale et temporelle. Malgré cette hétérogénéité d'objectifs, une seule méthode est employée : l'échantillonnage. Cela suffit à comprendre pourquoi il n'existe aucune procédure d'échantillonnage type (un même nombre d'unités statistiques, un même protocole de sélection des sites, un même ordre de récolte…) ; qui conviendrait à toutes les problématiques.

L'inexistence d'une procédure type et la complexité de la théorie de l'échantillonnage conduisent souvent les utilisateurs non avertis vers des pratiques non scientifiques. Smith, Anderson, et Pawley (2017) rapportent que trop d'écologues (78% des papiers publiés dans des journaux scientifiques) n'utilisent pas de protocoles d'échantillonnage aléatoires pour leurs suivis, seuls outils garantissant des résultats non biaisés. À cela s'ajoute que les suivis environnementaux souffrent d'un manque de formulation des objectifs et problématiques en amont de la phase de récolte des données (Legg et Nagy 2006). Les résultats issus de tels suivis n'apportent généralement pas la réponse à la problématique puisqu'ils échouent à détecter un changement (Fournier, White, et Heard 2019). Ils sont de fait, non seulement trompeurs, mais aussi dangereux car ils créent l'illusion que quelque chose d'utile a été fait (car considérés comme corrects), alors que ce n'est pas le cas (Peterman 1990). La non utilisation de procédure d'échantillonnage adaptée peut mener à un échec total de l'échantillonnage (Legg et Nagy 2006) dans le sens où les résultats seront controversables (Hayward et al. 2015). Par exemple, si l'objectif de l'échantillonnage était de déterminer l'effet de travaux de gestion sur une population, les résultats issus de ces procédures non scientifiques compromettent cette évaluation et donc la future mise en place de mesures adaptées (Vos, Meelis, et Ter Keurs 2000a).

La planification de la collecte des données est une phase essentielle à la conception d'une procédure d'échantillonnage. Martin, Kitchens, et Hines (2007) montrent que cette étape est indispensable pour que les résultats puissent répondre à la problématique d'une étude. Ils découvrent qu'une diminution dramatique du nombre d'individus d'une population n'a pas

été détectée faute de procédure non planifiée, entrainant de mauvaises mesures de gestion pour l'espèce considérée. Cet exemple n'est pas une exception. Trop souvent la problématique et la planification du suivi ne sont énoncées qu'*a postériori* de la récolte des données (Roberts 1991 ; Nichols et Williams 2006 ; Goldsmith 2012) ; menant à une incapacité de répondre à cette problématique. En ce sens, une tribune (Hayward et al. 2015) met en avant que les méthodes utilisées pour construire les suivis environnementaux sont souvent problématiques, et mènent à des résultats controversés. Ils soulignent que des méthodes robustes et des aides à la formulation de procédures d'échantillonnage doivent être développées et utilisées par les environnementalistes. L'objectif étant d'enrayer des résultats trop souvent sujets à controverses. Une procédure d'échantillonnage doit forcément contenir un processus statistique (Cunningham et Lindenmayer 2017), faire appel à un statisticien professionnel tout au long des étapes de la procédure d'échantillonnage est par ailleurs conseillé (Likens et Lindenmayer 2018). Les différentes étapes d'une procédure d'échantillonnage type, qui permettent d'atteindre des résultats au moins non controversés, sont connues. Elles seront répertoriées et développées dans la suite de ce chapitre introductif. Cependant, en plus de ne plus être sujets à controverse, la nécessité pour les scientifiques est maintenant que des résultats précis soient produits et cela pour un coût minimal.

La mise en place de suivis environnementaux efficaces et leurs optimisations sont les sujets d'une littérature déjà existante. Mais il apparaît que les méthodologies publiées sont généralement développées pour traiter un exemple particulier, et donc transférables à un nombre limité d'études. Ainsi, il existe par exemple une méthode qui permet d'optimiser les suivis d'espèces invasives dans le cas du changement climatique (Vicente et al. 2016). Elle est basée sur des modèles de distributions d'espèces. Une autre méthode (Carvalho et al. 2016) qui a pour objectif d'optimiser les suivis de populations multi-espèces permet à l'utilisateur de définir un nombre de sites et de réduire le coût total du suivi. Rudders (2011) montre que la performance des protocoles d'échantillonnage peut être évaluée en fixant au préalable le nombre d'unités statistiques nécessaires (trois niveaux différents pour estimer l'abondance de pétoncles dans ce cas-là). Ces méthodes restent très pertinentes et peuvent être utilisées, mais elles restent appropriées à certains types de suivis. Par exemple, la dernière méthodologie présentée ci-dessus est très utile si l'utilisateur veut savoir quelle précision il va pouvoir atteindre sur ses résultats de campagne avec un nombre fixé d'unités statistiques dans l'échantillon. Mais, nous pensons qu'il est préférable de calculer le nombre d'unités statistiques nécessaires pour atteindre un seuil de précision voulu. Cela permet, par exemple dans le cas d'un échantillonnage stratifié (*cf.* partie "Récolter les données" de cette introduction), de pouvoir atteindre la même précision dans toutes les strates. Ce qui n'est pas forcément le cas avec un effort d'échantillonnage fixe (cf. Kermorvant et al. (2017) pour exemple). Mais, il nous semble qu'il est essentiel d'avoir une même précision et donc une bonne vue d'ensemble de la variable étudiée dans la population plutôt que d'avoir certaines strates avec une estimation très précise et d'autres avec une estimation très peu précise. Il est donc nécessaire de développer une méthode utilisable sur le plus grand nombre de populations, quand l'objectif est l'interpolation spatiale ou le calcul d'estimateur(s). La méthode doit pouvoir prendre en compte des données s'il en existe déjà. Elle doit renvoyer un nombre d'unités statistiques permettant d'atteindre des résultats d'estimations précis et le protocole de sélection des unités statistiques le plus optimal. C'est-à-dire le protocole d'échantillonnage

qui permet d'atteindre un même niveau de précision dans les résultats que les autres, mais avec un nombre d'unités statistiques moindre. À partir de ces résultats, un coût total du suivi envisagé est facilement disponible.

Déroulé d'une procédure d'échantillonnage

On appelle procédure d'échantillonnage le processus complet qui permet de répondre à une problématique par un échantillonnage. Généralement, l'objectif est d'évaluer un paramètre noté Θ de la variable d'intérêt Y sur une population statistique Ω. La population statistique est composée de N unités élémentaires ω, aussi appelées unités statistiques ou unités d'échantillonnage. Y_{ω_j} indique la valeur prise par la variable Y pour l'unité statistique ω_j. Lors d'un échantillonnage sans remise, on sélectionne un ensemble de n unités statistiques ω différentes formant un échantillon noté S. La variable d'intérêt Y est ensuite étudiée sur cet échantillon pour pouvoir estimer le paramètre $\hat{\Theta}$ sur l'ensemble de la population statistique Ω.

Dans la littérature, lorsque l'objectif est l'inférence statistique des échantillons à la population, deux grandes méthodes d'échantillonnage sont souvent confrontées : les procédures basées sur un protocole d'échantillonnage probabiliste (*design-based*) et celles basées sur de la modélisation (*model-based*) (Gregoire 1998). La différenciation principale de ces deux méthodes vient des hypothèses sur le modèle statistique utilisé pour l'inférence (Särndal et al. 1978). La méthode *model-based* peut être définie comme une approche où le modèle statistique qui régit la distribution de la variable Y sur la population statistique Ω est inconnu. Il s'agit donc de reconstruire ce modèle à partir des mesures disponibles sur les unités statistiques échantillonnées pour ensuite l'utiliser et prédire $\hat{Y}$ sur Ω par inférence statistique et enfin calculer le paramètre $\hat{\Theta}$. *A contrario*, la méthode *design-based* pose l'hypothèse que le modèle de distribution de Y est connu sur Ω. L'utilisation d'un protocole d'échantillonnage probabiliste permet de sélectionner aléatoirement les unités statistiques ω_j qui doivent être recueillies. Un estimateur est ensuite utilisé pour estimer $\hat{\Theta}$ sur Ω par inférence des données récoltées et répondre à la problématique initiale.

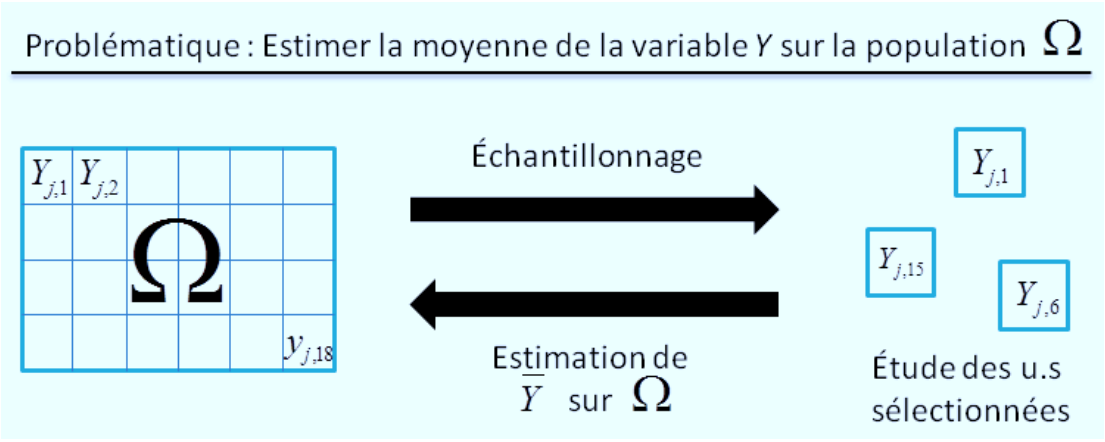


FIG. 1 : Exemple d'échantillonnage pour l'estimation de la moyenne d'une variable $\hat{Y}$: 18 unités statistiques dans la population, 6 dans l'échantillon sélectionné

La procédure d'échantillonnage nécessite plusieurs étapes indispensables : 1/ la définition d'une problématique, 2/ la définition du contexte de l'étude (population statistique, unités statistiques, variable et paramètre étudiés), 3/ la phase de récolte des données sur le terrain (combien d'unités statistiques ? Quand ? Où ?), 4/ le traitement des données récoltées puis 5/ la réponse à la problématique posée en 1/.

Définir une problématique

La première étape à effectuer lorsque l'on veut mettre en place un suivi scientifique est de bien cibler et formuler la problématique. Cette étape est le coeur de toute recherche qualitative ou quantitative, elle servira de guide à toutes les étapes suivantes du travail de recherche. Elle doit être formulée sous forme d'une question, et prendre en compte le type d'audience visé.

À une question particulière convient une procédure d'échantillonnage particulière. Une seule procédure d'échantillonnage peut rarement répondre à plus d'une problématique.

Définir le contexte de l'étude

Il est, dans un premier temps, important de bien définir cette population statistique et ses unités statistiques. La population statistique ne doit pas être confondue avec la population écologique elle-même. La population statistique est généralement une partie de la population écologique sur laquelle l'étude va être effectuée. Elle est donc composée d'un nombre fini d'unités statistiques. Lors d'un suivi purement spatial, la population statistique est représentée par une aire et une unité statistique peut être un point, une ligne ou un polygone. Si l'on se place dans le cas d'un suivi strictement temporel ; la population statistique devient une période de temps et l'unité, une date ponctuelle ou un intervalle de temps. Maintenant, si le problème est à la fois spatial et temporel ; la population statistique est une aire sur une période donnée et une unité statistique est représentée par un point, une ligne ou un polygone à un temps ou sur un intervalle de temps donné. En sciences environnementales où l'étude des phénomènes est souvent spatialisée, une unité statistique peut être un polygone (quadrat), une ligne (transect) ou un point, en fonction de la variable étudiée et de la taille de la population statistique. Il faut aussi définir la variable à étudier sur chaque unité statistique (abondance par exemple) et le paramètre à estimer (moyenne d'abondance par exemple).

Récolter les données

Le nombre d'unités statistiques dans l'échantillon a un effet direct sur la précision, définie comme l'opposé de la variance (Walther et Moore 2005), des résultats : plus ils sont nombreux, plus l'inférence produira des résultats de grande précision (avec, par opposition, une variance faible). Il est donc recommandé de prévoir le plus d'unités statistiques possibles dans l'échantillon. Mais, en sciences environnementales, et contrairement aux études faites en sociologie par exemple, où les sondages peuvent être effectués par téléphone ou mail, le coût des suivis varie plus fortement en fonction du nombre d'unités statistiques récoltées sur le terrain qui nécessitent déplacement et matériel de mesure spécifique pour ne citer que

cela… Il existe donc un fort compromis entre la précision de l'estimation finale, amenée par un grand nombre d'unités statistiques dans l'échantillon, et le coût total de l'étude, qui est souvent un facteur limitant.

En plus d'amener de la précision dans l'inférence finale, le nombre d'unités statistiques détermine quels outils statistiques pourront, ou ne pourront pas, être utilisés lors de l'inférence. Tous ne nécessitent pas les mêmes conditions d'application et le nombre d'unités statistiques doit être adapté pour l'inférence des données. Par exemple, si l'on veut utiliser un krigeage ordinaire pour interpoler la variable Y à toutes les unités statistiques de la population Ω les conditions d'application sont déterminées par le modèle de distribution paramétrique qui est utilisé lors de la construction du semi-variogramme. Une règle généralement recommandée est d'avoir au moins un échantillon de 30 unités statistiques dans chaque lag (distance entre les paires sur laquelle le semi-variogramme est calculé) du semi-variogramme (Cressie 1985). Il faut donc vérifier les conditions d'application des outils statistiques qui vont être utilisées pour l'inférence des données à la population, pour pouvoir les respecter et ne pas compromettre l'interprétation des résultats finaux.

Si le nombre d'unités statistiques à récolter ne dépend pas de la méthode d'échantillonnage choisie, c'est à dire *design-based* ou *model-based*, le choix des unités à échantillonner, lui, en dépend fortement. Pour l'utilisation d'une méthode *model-based*, les unités statistiques ne doivent pas crucialement être sélectionnées selon un protocole d'échantillonnage probabiliste (Gregoire 1998). Par contre, la méthode *design-based* nécessite que les unités statistiques de l'échantillon soient réparties aléatoirement sur la population étudiée pour garantir une inférence non biaisée. Pour cela, il faut faire appel à un protocole d'échantillonnage. Les protocoles d'échantillonnage sont des outils servant à sélectionner les unités statistiques et à définir leur emplacement et leur ordre de prélèvement. Indépendamment du nombre d'unités statistiques sélectionnées, le protocole d'échantillonnage utilisé a un impact sur l'étendue de l'incertitude, mais aussi sur la validité statistique des résultats récoltés (Müller, Rodríguez-Díaz, et Rivas López 2012). Il en existe un grand nombre, chacun avec ses avantages et ses inconvénients. En revanche, quelle que soit la ressource échantillonnée, il est conseillé d'utiliser un protocole d'échantillonnage probabiliste. Ce sont les protocoles d'échantillonnage qui ont un algorithme utilisant de l'aléatoire dans la sélection des unitées statistiques. Cette composante aléatoire est la seule méthode qui permet aux résultats d'inférence de ne pas être biaisés. Un grand nombre d'unités statistiques couplé à un protocole d'échantillonnage probabiliste assure donc des résultats précis et non biaisés. La justesse d'un estimateur du paramètre $\hat{\Theta}$, ou moyenne des erreurs au carré (MSE) est donnée par la formule (MacKenzie 2006) :

$$MSE(\hat{\Theta}) = Var(\hat{\Theta}) + [Biais(\hat{\Theta})]^2$$

Dans cette thèse, nous définirons la performance d'un protocole d'échantillonnage comme sa capacité à rendre une inférence précise avec un nombre d'unités statistiques minimum dans l'échantillon. Cette définition nous servira à comparer les protocoles entre eux. Pour une même précision fixée, un protocole qui nécessite moins d'unités statistiques que les autres sera considéré comme le plus performant. Malheureusement, il n'existe pas un protocole qui soit le meilleur dans tous les cas de figures. Certains sont plus performants quand il existe une tendance spatiale, d'autres quand il n'y en a pas, d'autres encore pour étudier des espèces

rares (McDonald 2014). Certains sont capables de fournir une liste d'unités statistiques à échantillonner supplémentaire si celles de la liste principale ne peuvent pas être récoltées, d'autres ne le permettent pas. Certains proposent des probabilités d'inclusion différentes en fonction des unités d'échantillonnage, rendant de meilleurs résultats si la distribution de la variable d'intérêt est hétérogène. Avant de choisir le protocole d'échantillonnage, il faut donc avoir une bonne connaissance de la ressource que l'on veut étudier.

Il existe deux grands types de protocoles d'échantillonnage : les non-probabilistes et les probabilistes. Nous passerons rapidement sur les protocoles non probabilistes. Dans ce cas-là, la sélection des unités statistiques constituant l'échantillon n'est pas soumise à un processus aléatoire (McDonald 2003). Il existe plusieurs types de protocoles d'échantillonnage non probabilistes (par commodité, à l'aveuglette, par quotas, par choix raisonné) mais leur utilisation n'est pas reconnue scientifiquement puisqu'elle entraîne un biais presque certain dans les estimateurs, qui est souvent impossible à calculer. La non randomisation implique que les résultats obtenus par ces procédures ne suivent pas de loi de probabilité et ne peuvent donc pas être bornés par un intervalle de confiance. Ils peuvent donc difficilement être traités statistiquement et utilisés pour de l'inférence. Même si leur utilisation est simple et ne nécessite pas de travail en amont de la phase terrain, ils ne garantissent pas d'avoir des résultats fiables. Il est donc fortement déconseillé d'utiliser ces protocoles lors d'un échantillonnage.

En utilisant un protocole probabiliste sur une population statistique finie, chaque unité incluse dans l'échantillon a une probabilité d'inclusion associée connue. Les résultats suivent des lois de probabilité qui sont elles aussi connues, on peut donc calculer une variance des estimateurs suivis mais aussi des intervalles de confiance. Ces protocoles permettent d'avoir les pré-requis pour l'inférence statistique (Stehman et Overton 1996). Les ordinateurs sont aujourd'hui seulement capables de générer des nombres pseudo-aléatoires (un biais dans la qualité de l'aléatoire peut en effet être observé). Mais les études sur la capacité du cerveau humain à générer une suite de nombres aléatoires se contredisent (voir, par exemple Persaud (2005) et Figurska, Stańczyk, et Kulesza (2008)). Bien souvent, les protocoles d'échantillonnage suivent un algorithme plus complexe qu'un simple tirage aléatoire des unités d'échantillonnage, et le cerveau humain n'est pas capable de les mettre en pratique. C'est pourquoi il est conseillé de toujours utiliser un ordinateur pour la sélection des unités d'échantillonnage. Les outils informatiques qui permettent d'utiliser les protocoles d'échantillonnage probabilistes renvoient les coordonnées des unités statistiques à échantillonner, ainsi que l'ordre dans lequel elles doivent être récoltées.

Le protocole d'échantillonnage probabiliste le plus simple, et certainement le plus utilisé, est appelé aléatoire simple (EAS ou SRS). Comme son nom l'indique, les unités statistiques sont simplement tirées aléatoirement. L'utilisateur choisit le nombre d'unités statistiques qu'il veut échantillonner et l'ordinateur les sélectionne au hasard sur la population statistique ; toutes les unités statistiques ont donc la même probabilité d'être sélectionnées. La distribution complètement aléatoire des unités à échantillonner peut être problématique sur de grandes populations ayant une forte dispersion : en effet, il peut arriver que certaines parties du site d'étude soient sur-représentées ou sous-représentées. Les parties sur-échantillonnées sont donc bien documentées mais celles non échantillonnées ne le sont pas du tout. Il existe une famille de protocoles d'échantillonnage qui permet de pallier ce problème récurrent du

SRS. Les protocoles d'échantillonnage spatialement équilibrés sont des protocoles d'échantillonnage probabilistes, qui, en plus du hasard, intègrent une composante permettant un meilleur équilibre spatial entre les unités d'échantillonnage sélectionnées. Un des avantages de ces protocoles est l'assurance d'un recouvrement spatial entier du site d'étude (Brown, Robertson, et McDonald 2015). Ils évitent de ce fait d'éventuelles zones non échantillonnées. Ces stratégies produisent des estimateurs sur les populations ayant un remarquable taux de convergence de la variance, qui sont en plus, distribués normalement pour des grands échantillons (Barabesi et Franceschi 2011). Il est bien connu que les performances d'un protocole d'échantillonnage sont liées à sa capacité de dépendance à l'espace (Rajabi et Ataie-Ashtiani 2014). Ces protocoles d'échantillonnage spatialement équilibrés produisent de très bonnes estimations pour des variables qui ont des tendances spatiales (Stevens et Olsen 2004 ; Theobald et al. 2007 ; Anton Grafström 2012, 2012 ; Anton Grafström et Lundström 2013 ; Robertson et al. 2013).

Une technique classique pour permettre d'équilibrer selon la procédure suivante est connue depuis longtemps : une grille régulière est appliquée sur l'aire d'étude, une première unité d'échantillonnage est tirée au hasard dans cette grille et ensuite, les autres unités sont placées à équidistance les unes des autres (Cochran 1977 ; Zinger 1963). Cette technique, la plus basique des protocoles d'échantillonnage spatialement équilibrés, est appelée échantillonnage systématique (SS). Elle est réputée pour fournir des résultats d'estimation plus précis que la méthode du SRS avec un même nombre de points (Le Cacheux 1955). Malgré cela, l'échantillonnage systématique ne permet pas d'avoir une probabilité d'inclusion identique entre les unités statistiques et connue sur tous les sites. L'écart entre les sites est choisi arbitrairement par l'utilisateur et la validité de l'estimation n'est donc pas assurée (Hasel 1938), caractéristique qui confère à ce protocole l'adjectif de quasi-aléatoire. Aussi, si le phénomène étudié admet une cyclicité spatiale, l'échantillonnage systématique est inefficace.

Gibson et Lenzmeier (1981) présentent pour la première fois en 1981, l'idée de partitionnement et d'adressage de l'espace par le biais de leur méthode GBT (general balanced ternary). Une fois l'espace partitionné et adressé sur ordinateur avec l'avènement des Systèmes d'Informations Géographiques à la fin du XX^e siècle, les algorithmes d'échantillonnage spatialement équilibré se sont développés.

Aujourd'hui, le GRTS (generalized random tessellation stratified) (Stevens et Olsen 2003) est l'un des protocoles spatialement équilibrés des plus connus et des plus utilisés. Il a été initialement développé par et pour l'agence de protection environnementale américaine au sein du programme "Environmental Monitoring and Assessment program". Dans un premier temps, le GRTS crée une grille numérotée et la place aléatoirement sur la zone d'intérêt. Ensuite il linéarise les localisations (numéros) et utilise une transformation (hiérarchique inverse) sur ces localisations. Finalement il tire un point au hasard et sélectionne les autres points à intervalle régulier. Passer par une transformation de réversion hiérarchique permet de maintenir les propriétés spatiales de départ et donc d'équilibrer l'échantillonnage. À l'intérieur de la population statistique, aucune aire n'est sur-représentée par beaucoup de stations, aucune aire n'est sous-représentée avec peu de stations. Le GRTS n'est utilisable que sur une ou deux dimensions (équilibre sur deux variables, le plus souvent les coordonnées latitudinales et longitudinales), et donc applicable à des problématiques environnementales soit spatiales

(équilibre des unités statistiques sélectionnées sur une aire donnée), soit temporelles (équilibre des unités statistiques dans une période de temps donné).

Or, il semble parfois opportun de prendre en compte une troisième dimension (pour, par exemple, équilibrer les unités statistiques sélectionnées en fonction de l'espace et du temps) voire plus encore (caractéristiques des milieux, saisonnalité, statuts de conservation, profondeur…). C'est dans l'optique d'étendre le GRTS et son équilibre spatial à plus de dimensions, que le BAS (balanced acceptance sampling) a été développé par Blair Roberston, Jennifer Brown, Trent Mcdonald et Peter Jaksons en 2013 (Robertson et al. 2013). Pour étendre l'équilibre à d'autres variables que l'espace, il est possible de créer une couche GIS de probabilités d'inclusion des unités statistiques en fonction de données environnementales adéquates. L'efficacité des protocoles d'échantillonnage spatialement équilibrés peut être majorée en augmentant la probabilité de sélectionner des unités statistiques là où la variable d'intérêt a des chances d'admettre une grande variance (Foster et al. 2017). Les zones ayant une probabilité d'inclusion faible seront échantillonnées avec un effort d'échantillonnage plus faible que celles en ayant une forte.

Une grande partie de la thèse de Peter Jaksons (Jaksons 2014) se consacre au BAS et à l'évaluation de ses performances. Il le compare aussi à d'autres protocoles spatialement équilibrés comme le GRTS, sur la base de simulations et sur deux dimensions. Le BAS présente un équilibre spatial un peu meilleur que celui du GRTS (Robertson et al. 2013). Ce protocole utilise les séquences de Halton (Halton 1960) pour déterminer les coordonnées des points d'échantillonnage. Ces séquences sont utilisées en statistiques pour générer des points dans l'espace pour des méthodes numériques comme celles de Monte Carlo. Elles génèrent des points mieux espacés dans l'espace que les séquences pseudo-aléatoires mais sont par définition des séquences déterministes, connues pour être équidistribuées (Wang et Hickernell 2000).

Pour générer une séquence de 2, on commence par diviser l'intervalle (0,1) de moitié, puis par quatre, huit.

$$\{\frac{1}{2};\frac{1}{4};\frac{3}{4};\frac{1}{8};\frac{5}{8};\frac{3}{8};\frac{7}{8};\frac{1}{16};\frac{9}{16};...\}$$

Pour générer une séquence de 3, on commence par diviser l'intervalle (0,1) par trois, puis par 9, puis en 27^{ieme}

$$\{\frac{1}{3};\frac{2}{3};\frac{1}{9};\frac{4}{9};\frac{7}{9};\frac{2}{9};\frac{5}{9};\frac{8}{9};\frac{1}{27};...\}$$

On peut créer des coordonnées de points (x,y) en combinant deux séquences de Halton, par exemple :

$$\{\left(\frac{1}{2},\frac{1}{3}\right);\left(\frac{1}{4},\frac{2}{3}\right);\left(\frac{3}{4},\frac{1}{9}\right);\left(\frac{1}{8},\frac{4}{9}\right);\left(\frac{5}{8},\frac{7}{9}\right);\left(\frac{3}{8},\frac{2}{9}\right);...\}$$

C'est cette séquence de Halton que le BAS utilise. Pour créer un échantillon à deux dimensions, le BAS utilisera deux séquences de Halton. Pour créer un échantillon à trois dimensions, le BAS utilisera trois séquences de Halton, etc… La séquence de Halton est complètement déterministe. Mais on peut y intégrer de l'aléatoire sans affecter les propriétés de distribution de la séquence. L'aléatoire est ajouté en sélectionnant aléatoirement le point de départ de la séquence de Halton (Wang et Hickernell 2000). Le BAS produit des localisations d'unités statistiques spatialement équilibrées dans un certain ordre. Les deux premiers points sont spatialement équilibrés entre eux, les trois premiers aussi etc. Un des avantages de ce protocole est que si l'échantillonnage ne peut être effectué sur un ou plusieurs des points, l'équilibre spatial ne sera pas perturbé (Brown, Robertson, et McDonald 2015).

Le BAS est un protocole relativement nouveau, de ce fait, il a encore très peu été utilisé ou amélioré. Mais Foster et al. (2017) proposent déjà une méthode intéressante pour intégrer des sites historiques à un échantillonnage spatialement équilibré (le BAS). Un problème récurrent dans l'optimisation de protocoles d'échantillonnage déjà en place est la prise en compte des sites historiques (mêmes sites suivis depuis longtemps). Ces sites sont suivis sur le long terme avec comme objectif de déterminer une évolution, et optimiser le protocole en excluant ces sites serait perdre une information temporelle importante. L'équilibre spatial est gardé autour des sites historiques en y altérant la probabilité d'inclusion. Le BAS aura une probabilité plus faible de tirer des sites proches de ces sites historiques.

Le protocole d'échantillonnage nommé "Halton iterative partitionning (HIP)" (Robertson et al. 2018) a été développé en 2018 par la même équipe que celle qui créa le BAS 5 ans plus tôt. En effet, il s'est avéré que le BAS requiert une puissance informatique intensive quand il dessine l'échantillonnage d'une population discrète avec des unités statistiques très petites. De plus, passer par une phase d'acceptation/rejet des unités statistiques impacte la probabilité d'inclusion choisie par l'opérateur au départ. Le HIP pallie ces problèmes, tout en gardant les propriétés du BAS qui étaient intéressantes pour les études environnementales. Conceptuellement, le HIP partitionne en boites les dimensions choisies par itérations pour équilibrer l'échantillon en utilisant les propriétés de la séquence de Halton (*cf.* partie sur le BAS). Les points sont ensuite dessinés dans un ordre spécifique à partir des boites pour sélectionner des échantillons spatialement équilibrés. Comme pour le BAS, la partition peut être faite dans deux dimensions ou plus, suivant les besoins de l'étude. De plus, l'ordre spécifique utilisé par le HIP permet de créer une liste complémentaire de points d'échantillonnage qui gardera l'équilibre spatial si les points de la liste principale ne peuvent être effectués.

Une seconde équipe s'est intéressée aux protocoles d'échantillonnage spatialement équilibrés. Anton Grafström (2012) a développé le protocole appelé "spatially correlated poisson sampling (SCPS)". Il est dérivé de la méthode "correlated poisson sampling (CPS)" (Bondesson et Thorburn 2008a). Ce protocole peut être utilisé pour dessiner un plan d'échantillonnage équilibré sur plusieurs dimensions. C'est une méthode très générale, utilisable sur des populations finies, qui pondère négativement la probabilité d'inclusion des unités statistiques proches les unes des autres (distances euclidiennes). Les unités statistiques sont sélectionnées séquentiellement, c'est-à-dire une à une, jusqu'à avoir sélectionné le nombre d'unités désiré par l'utilisateur. La probabilité d'inclusion de chaque unité statistique est mise à jour dépendamment du résultat de l'événement de tirage de l'unité précédente. Parallèlement, la même

équipe développe deux autres protocoles d'échantillonnage spatialement équilibrés : les "local pivotal methods" (LPM1 et LPM2) (Anton Grafström, Lundström, et Schelin 2012). Le mode de fonctionnement de ces protocoles est aussi séquentiel, mais cette fois les unités statistiques sont sélectionnées deux à deux. La différence entre le LPM1 et le LPM2 réside dans la façon de sélectionner ces deux unités. Le LPM1 a un meilleur équilibre spatial des unités statistiques, mais le LPM2 est plus simple et rapide à exécuter. Pour l'instant, Le LPM est le seul protocole d'échantillonnage équilibré sur plusieurs dimensions qui soit disponible sous le logiciel gratuit R via le package {BalancedSampling} (Grafström et Lisic 2016).

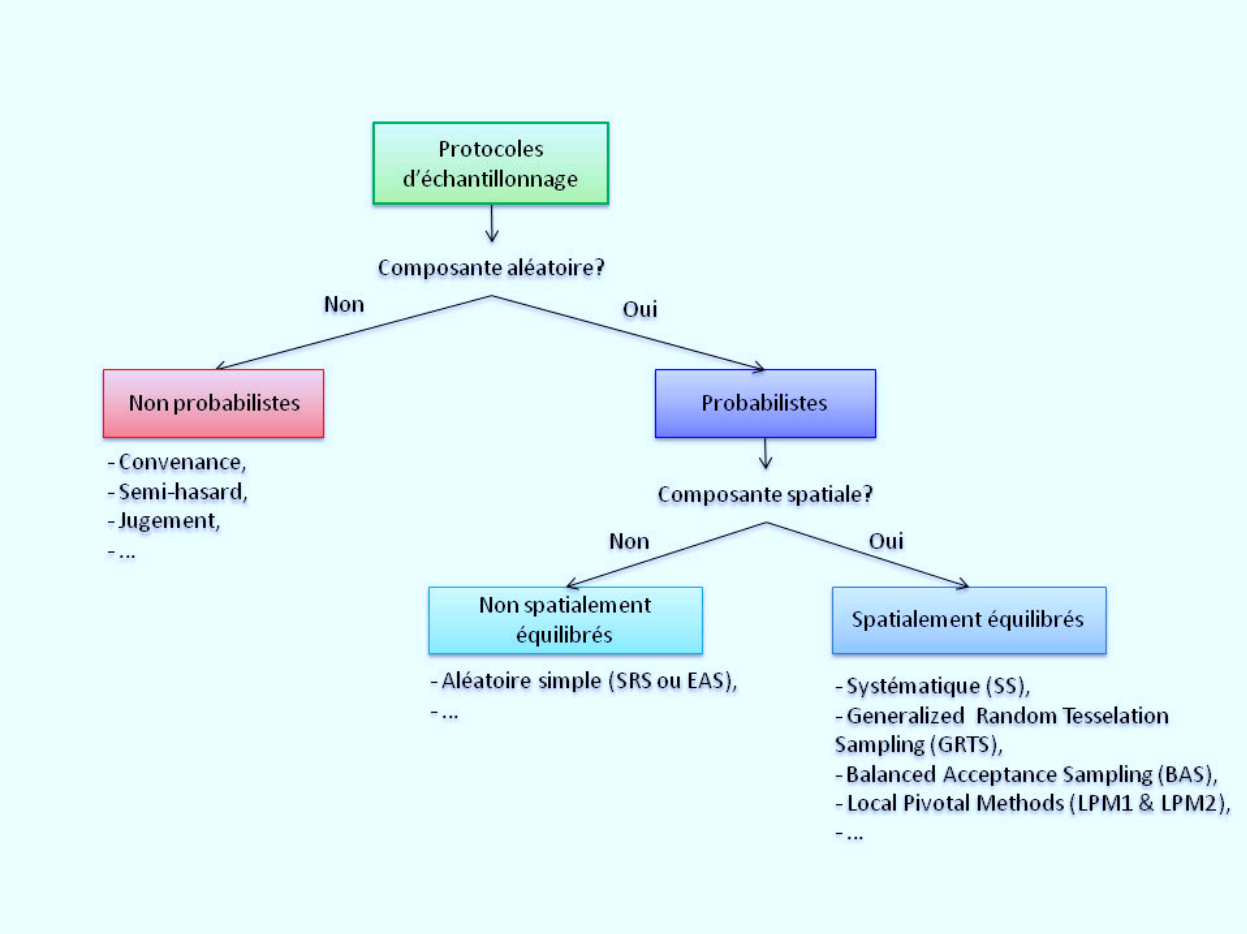


FIG. 2 : Les grands types de protocoles d'échantillonnage

Parfois, les unités statistiques sélectionnées lors du premier tirage ne sont pas accessibles sur le terrain. Cela a une incidence directe sur la précision des résultats du suivi puisque la précision est proportionnelle au nombre d'unités statistiques contenues dans l'échantillon. Si certains domaines où l'échantillonnage est utilisé ne sont pas touchés par ce problème, l'étude des populations naturelles dans leur milieu en fait très souvent les frais. On peut aisément citer 3 cas où cette situation peut se produire : 1/ le point sélectionné par le protocole d'échantillonnage tombe dans une propriété privée où l'opérateur n'a pas les droits d'entrée,

2/ le point se trouve dans un lieu où il serait dangereux pour l'opérateur de travailler (pieds de falaise ou zone contaminée chimiquement) et 3/ le point n'est pas accessible par l'opérateur (au milieu d'un lac par exemple).

Pour remédier à cette difficulté récurrente, certains protocoles d'échantillonnage ont la capacité de pouvoir fournir une liste d'unités statistiques de remplacement (SRS, GRTS, HIP). Cette liste garantie que le nombre d'unités statistiques prévues pour le suivi puisse effectivement être récolté. Nous devons cependant avertir l'utilisateur de certaines limites à leur utilisation. Même si certains sites ne sont pas échantillonnés, il faut garder une trace de ce qui s'y passe, cela garantit l'aléatoire. Il faut aussi faire attention à bien distinguer les sites qui ne font pas partie de la population d'intérêt de ceux qui le sont mais qui ne peuvent pas être échantillonnés ; d'où l'importance de bien définir la population statistique avant d'aller sur le terrain. Sinon, il faut être sûr qu'il n'y a pas de différences significatives entre les sites qui peuvent être échantillonnés et ceux qui ne le peuvent pas. Dans ce dernier cas l'inférence doit être faite seulement sur la partie de la population échantillonnable.

Il existe une technique, qui permet d'augmenter la précision de l'estimation avec un nombre d'unités statistiques fixé : la stratification. On sait que l'erreur d'estimation est liée à la variance du caractère étudié dans la population. Il faudra plus d'unités statistiques pour appréhender cette variance dans une population hétérogène que dans une population homogène. Autrement dit, à nombre d'unités statistiques égal dans l'échantillon, la précision de l'estimation de $\hat{\Theta}$ atteinte sur une population hétérogène sera moindre que celle atteinte sur une population homogène. L'idée de la stratification, est donc de découper la population en sous-ensembles homogènes, mais hétérogènes entre eux, puis d'échantillonner indépendamment dans ces sous-ensembles. La somme du nombre d'unités statistiques nécessaires dans chaque sous-ensemble sera alors inférieure au nombre d'unités statistiques qui aurait été nécessaire si la population n'avait pas été stratifiée. Lorsqu'elle est bien pensée, cette technique peut permettre une baisse du nombre d'unités statistiques contenues dans l'échantillon, et donc du coût total du suivi. Cette méthode rend donc un meilleur taux d'estimation de la population si le site d'intérêt est divisé en zones relativement homogènes (Yoccoz, Nichols, et Boulinier 2001) mais aussi lorsque ces zones sont construites à partir de caractéristiques environnementales corrélées au phénomène étudié (Zhao et al. 2016). Mais, si les strates ne sont pas bien construites, la stratification peut s'avérer inutile. Il faut avoir une connaissance de la distribution de la variable étudiée avant de dessiner les strates, ou au moins connaitre la distribution de variables exogènes liées à la variable d'intérêt et affectant sa variance. Lorsque seulement quelques variables exogènes sont disponibles, il est aisé de les utiliser pour construire des strates, mais lorsqu'il y en a trop, la stratification peut devenir complexe et mener à de petites strates très nombreuses (Grafström et Schelin 2014). La stratification est une méthode qui peut être utilisée avec n'importe quel protocole d'échantillonnage à l'intérieur des strates.

Certains protocoles d'échantillonnage avancés (LPM, BAS et HIP) ont la capacité de prendre en compte des probabilités d'inclusions différentes en fonction des unités statistiques. Ils ont la capacité de faire une "stratification automatique" dans des hautes dimensions (plusieurs variables exogènes) (Grafström et Schelin 2014), en changeant la probabilité d'inclusion des unités statistiques de la population.

Traiter les données

Une fois la phase de terrain terminée, et éventuellement après un traitement des échantillons en laboratoire, une valeur mesurée ou observée de la variable Y est disponible pour toutes les unités statistiques ω_j prélevées et donc une estimation du paramètre Θ est possible. Les méthodes permettant l'estimation $\hat{\Theta}$ du paramètre Θ de la variable Y sont différentes en fonction du mode de sélection des unités statistiques (*design-based* ou *model-based*).

- Cas où les données ont été récoltées avec un protocole d'échantillonnage probabiliste (*design-based*)

Lorsque les unités statistiques ont été sélectionnées suivant la méthode *design-based*, c'est-à-dire à l'aide d'un protocole d'échantillonnage probabiliste, l'inférence du paramètre étudié se fait simplement par un estimateur. L'estimateur le plus connu, puisque linéaire et sans biais pour tout échantillon, est celui de Horvitz-Thompson (Horvitz et Thompson 1952). Il est disponible pour une estimation du paramètre "moyenne" $\hat{T}_{\overline{y}}$ et du paramètre "total" $\hat{T}_t$ et leurs variances respectives $V(\hat{T}_{\overline{Y}})$ et $V(\hat{T}_t)$ peuvent, entre autres, être calculées. Pour tout plan d'échantillonnage simple, c'est-à-dire pour lequel toutes les unités statistiques ont la même probabilité d'inclusion ; l'estimateur de la moyenne de la population est simplement la moyenne de l'échantillon :

$$\hat{T}_{\overline{y}} = \frac{1}{n} \sum_{\omega_j \in S} Y_{\omega_j}$$

Sa variance est donnée par la formule :

$$V(\hat{T}_{\overline{y}}) = (1 - \frac{n}{N}) s^2_{\omega_j}$$

L'estimateur du total de la population vaut :

$$\hat{T}_t = \sum_{\omega_j \in S} \frac{Y_{\omega_j}}{n/N}$$

et sa variance :

$$V(\hat{T}_t) = N(N-n) \frac{s^2_{\omega_j}}{n}$$

Les plans d'échantillonnage simples (à probabilités d'inclusions égales) sont souvent utilisés lorsqu'il n'y a pas d'informations au préalable sur la population. Mais il est reconnu que si l'on connait déjà la population, l'utilisation d'une ou de plusieurs variables auxiliaires peut mener à un plan plus judicieux. Comme détaillé précédemment, une première façon d'intégrer ces variables est de les utiliser pour stratifier Ω en strates homogènes. Il suffira de calculer un estimateur par strate puis de les additionner (en les pondérant en fonction des probabilités d'inclusions utilisées) pour calculer le paramètre $\hat{\Theta}$ d'interêt. Une seconde façon d'introduire ces variables dans le plan est de sélectionner les unités d'échantillonnage avec

des probabilités d'inclusion π inégales. Dans ce cas, les probabilités d'inclusion ne sont plus égales à $\frac{n}{N}$. L'estimateur du total peut être calculé par la formule :

$$\hat{T}_Y\pi = \sum_{\omega_j \in S} \frac{Y_{\omega_j}}{\pi_{\omega_j}}$$

et sa variance :

$$Var(\hat{T}_Y\pi) = \sum_{\omega_j=1} \frac{Y^2_{\omega_j}}{\pi_{\omega_j}} + \sum_{\omega_j \neq \omega_k} \sum \frac{Y_{\omega_j}}{\pi_{\omega_j}} \frac{Y_{\omega_k}}{\pi_{\omega_k}} (\pi_{\omega_j\omega_k} - \pi_{\omega_j}\pi_{\omega_k})$$

avec ω_j et ω_k deux unités d'échantillonnage différentes et π_{ω_j} et π_{ω_k} leurs probabilités d'inclusion d'ordre 1 respectives, c'est-à-dire la probabilité que chacune d'elles appartienne à l'échantillon S indépendamment de l'autre. La probabilité d'inclusion d'ordre 2, $\pi_{\omega_j\omega_k}$, correspond à la probabilité que les deux unités ω_j et ω_k soient sélectionnées conjointement dans le même échantillon S.

- Cas où les données n'ont pas forcément été récoltées avec un protocole d'échantillonnage probabiliste (*model-based*)

Comme elles ne sont pas forcément générées par un protocole d'échantillonnage probabiliste, les données obtenues peuvent être très variées. En fonction du type de données dont on dispose, plusieurs méthodes de modélisation de $\hat{Y}$ sur Ω sont envisageables pour ensuite calculer $\hat{\Theta}$. Si les données sont suffisamment nombreuses, géo-référencées et qu'il existe un grand nombre de distances différentes entre elles, on peut faire appel aux méthodes géostatistiques. Ce sont alors des méthodes d'interpolation spatiale qui permettent, à partir de données ponctuelles de terrain, de reconstruire la variable $\hat{Y}$ sur la population Ω. La méthode d'interpolation spatiale la plus connue et la plus utilisée est le krigeage ordinaire. Le krigeage porte le nom de son inventeur D.G Krige, qui était un ingénieur minier sud-africain. Il développa dans les années 50 (Krige 1951) un ensemble de méthodes statistiques pour essayer de déterminer, à partir de forages miniers, la distribution spatiale des minerais. Mais c'est le français Matheron (Matheron 1963) qui mit en forme la méthode et lui donna son nom, c'est d'ailleurs lui aussi qui utilisa le mot "géostatistique" pour la première fois. Le krigeage analyse les points de données et y ajuste un modèle de distribution paramétrique pour créer un semi-variogramme. Les conditions d'utilisation d'un krigeage sont liées à celles du modèle paramétrique utilisé pour la construction du semi-variogramme. Webster et Oliver (1993) ont montré que pour obtenir un semi-variogramme isotropique (qui présente les mêmes propriétés dans toutes les directions), 150 unités statistiques devraient suffire, mais pour que celui-ci soit fiable (c.à.d. avec un intervalle de confiance raisonnable), 225 unités statistiques sont nécessaires.

Les difficultés commencent quand les données ne satisfont pas ces conditions d'utilisation. Parfois il est supposé que le modèle de variogramme est assez robuste pour passer au-dessus des conditions d'applications. Sinon des méthodes plus complexes doivent être utilisées, comme des méthodes non paramétriques ou des méthodes à distribution statistique libre (voir Henley (2012) pour ces méthodes).

Une autre méthode de modélisation de $\hat{Y}$ sur Ω est la modélisation statistique. L'objectif est de déterminer le modèle statistique qui régit la distribution de la variable Y sur Ω. Il faut donc sélectionner une loi de distribution, une fonction de lien et des variables explicatives X. Ensuite il faut valider l'ajustement du modèle aux données, tester son pouvoir prédictif et enfin le confronter à d'autres modèles pour sélectionner le meilleur. Les différents modèles testés peuvent donc différer les uns des autres de par : 1/ la loi de distribution choisie, 2/ la fonction de lien choisie et 3/ les variables explicatives - en fonction de leur nature et de leur agrégation les unes en fonction des autres (variables additives, multiplicatives, emboitées, élevées au carré…). Pour pouvoir utiliser ces variables explicatives, il faut qu'elles soient disponibles pour toutes les unités statistiques ω de Ω. Cette technique de modélisation est souvent utilisée en sciences environnementales. Par exemple on modélise assez intuitivement des niches écologiques d'espèces en fonction de variables environnementales. Cela permet de prédire, entre autres, des probabilités de présence en fonction de conditions biotiques et/ou abiotiques (Soberon et Peterson 2005). Plusieurs outils pour la création d'algorithmes ont été développés pour inférer des points de données d'occurrence jusqu'à des cartes de probabilité de présence, en prenant en compte des données environnementales (voir, par exemple, MAXENT (Elith et al. 2011), et le package {unmarked} du logiciel R (Fiske et Chandler 2011)). Utiliser un modèle de distribution pour reconstruire $\hat{Y}$ qui représente le mieux possible Y sur Ω requiert qu'une étude similaire ait déjà été effectuée autre part, et que les mêmes variables explicatives X soient disponibles pour tout ω_j sur Ω. Si ces conditions ne sont pas respectées, le modèle peut ne pas être adapté et prédire une population $\hat{Y}$ très différente de la population originale Y, et donc calculer un $\hat{\Omega}$ biaisé.

Répondre à la problématique

Les deux méthodes *design-based* et *model-based* permettent une inférence des données pour estimer le paramètre d'intérêt. Bien suivre les recommandations du *design-based* et construire un modèle valide en *model-based* sont les pré-requis pour que les résultats d'estimations soient précis et non biaisés, ou du moins que cette précision et ce biais soient calculables. Cela permet d'atteindre des résultats scientifiquement non contestables, qui peuvent être utilisés pour répondre à la problématique.

Résumé de la procédure d'échantillonnage

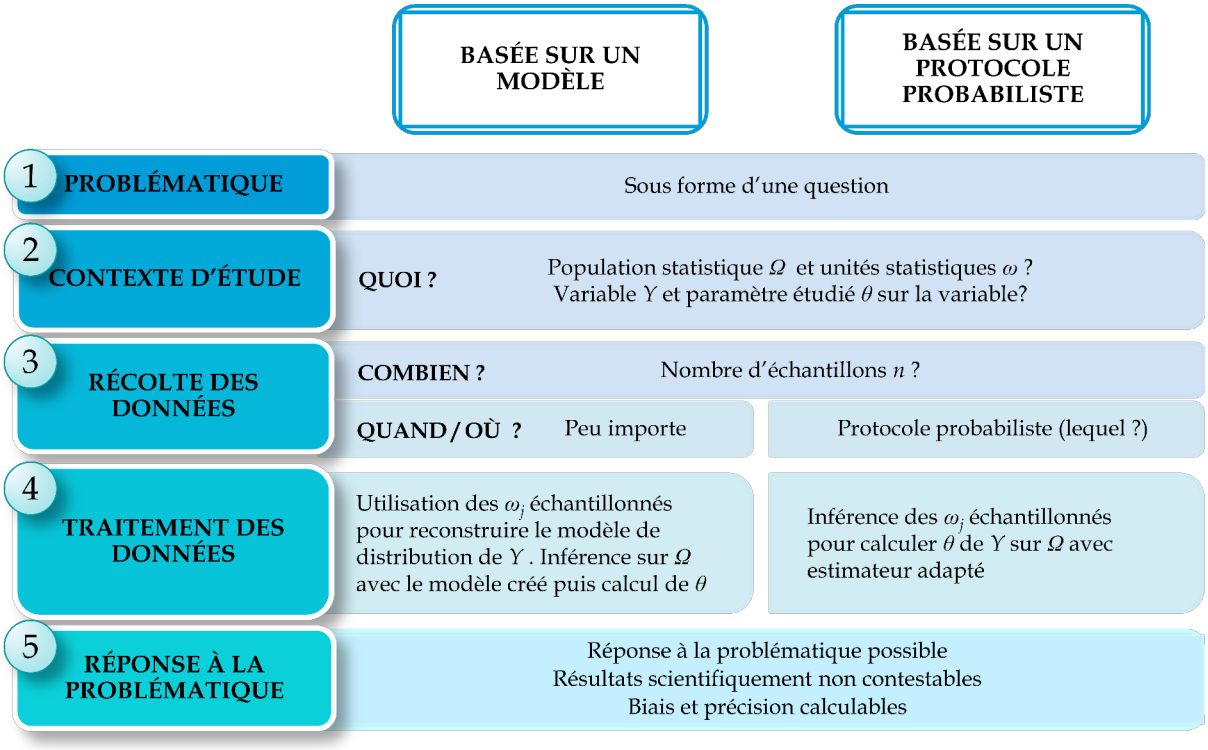


FIG. 3 : Déroulé d'une procédure d'échantillonnage

Objectifs, méthodologie retenue et organisation du manuscrit

Cette thèse s'intègre dans un contexte où les méthodes utilisées pour la mise en place de suivis environnementaux sont souvent problématiques et mènent à des résultats controversables. L'objectif est de proposer une méthodologie adaptable à la plupart des suivis environnementaux qui permettra aux utilisateurs de produire des suivis scientifiques efficaces. Nous avons développé une méthodologie qui permet à l'utilisateur de fixer la précision qu'il veut sur ses résultats d'estimation et qui renvoie un protocole d'échantillonnage optimal associé à un nombre d'unités statistiques à échantillonner. Une fois le nombre d'unités statistiques connu, il est simple d'estimer le coût de mise en place de la procédure d'échantillonnage sélectionnée sur le terrain.

Nous sommes partis de la définition même de la performance d'un protocole d'échantillonnage pour élaborer une méthodologie qui permet de tester, puis de choisir, le protocole le plus performant pour chaque étude. Plus un protocole d'échantillonnage est performant, moins il nécessite d'unités statistiques pour atteindre une précision voulue. La méthodologie présentée permet donc de déterminer puis de comparer, pour une (ou des) précision(s) désirée(s) sur les résultats, le nombre optimal d'unités statistiques à échantillonner pour différents protocoles. Le protocole qui renvoie le nombre d'unités statistiques le plus faible est considéré comme le plus performant entre tous ceux testés. Or, plus un protocole est performant, moins il est coûteux à mettre en place ; puisqu'il demande une phase de terrain moins intense. Cette méthodologie permet d'obtenir le meilleur rapport coût-efficacité pour une étude nécessitant un échantillonnage dans un objectif d'inférence, autrement dit, de baisser son prix tout en garantissant une précision adéquate.

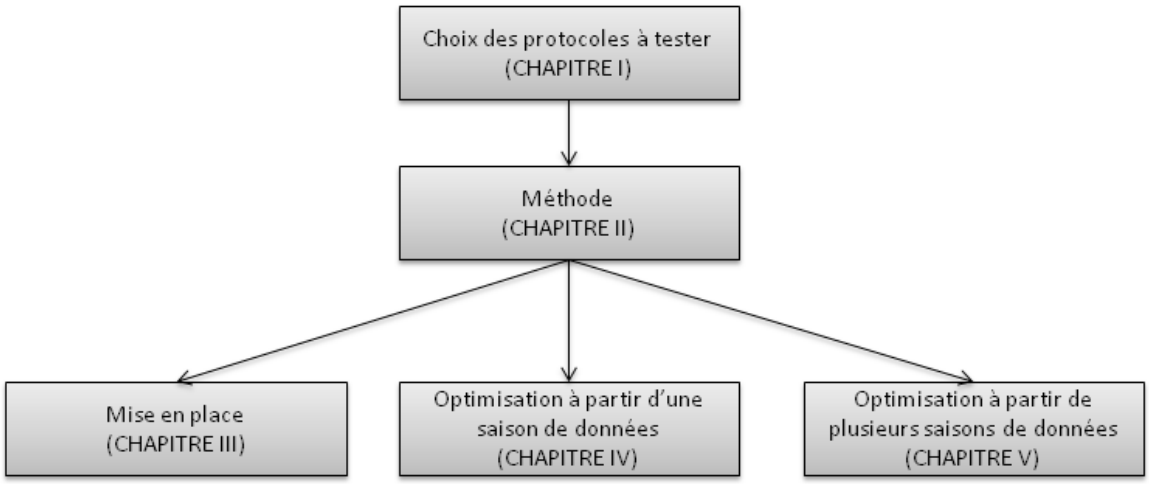


Fig. 4 : Schéma introduisant les liens entre les différentes parties du manuscrit

Le chapitre I de la thèse est une revue des protocoles d'échantillonnage existants, et plus particulièrement des protocoles d'échantillonnage spatialement équilibrés (spatially balanced sampling designs - SBS). La méthode présentée dans cette thèse nécessite de choisir des

protocoles à comparer entre eux et son utilisateur doit donc être au fait des dernières innovations en la matière. Le chapitre suivant (chapitre II) présente ladite méthode, développée pour mettre en place un suivi efficace ou optimiser un suivi déjà en place. Viendra ensuite un exemple de mise en place de suivi de population dans le chapitre III. L'exemple choisi est celui de la population de moustiques tigres *Aedes albopictus* dans les zones méditerranéennes urbanisées françaises. Le chapitre IV traitera de l'optimisation d'un suivi dans le cas de données issues d'une étude existante. L'exemple retenu est la campagne 2012 du suivi de la palourde *Ruditapes philippinarum* dans le bassin d'Arcachon. Ce même suivi est optimisé en prenant en compte six années de relevés dans le chapitre V.

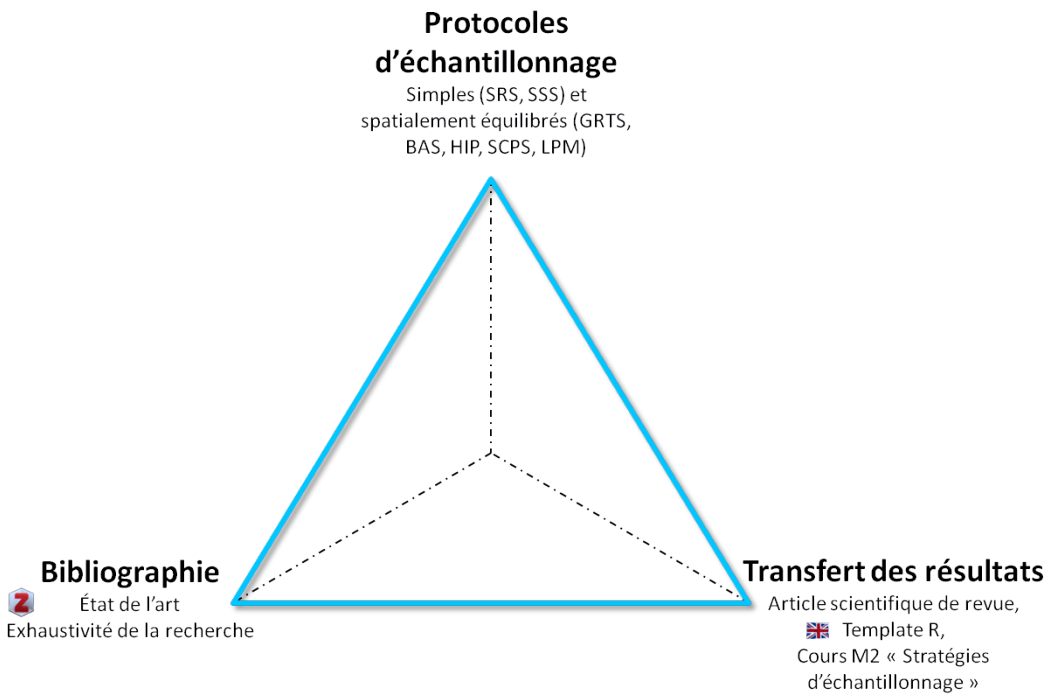
CHAPITRE I : Des protocoles d'échantillonnage spatialement équilibrés pour les suivis environnementaux

Synopsis :

L'objectif de cette thèse est de développer puis de tester une méthode permettant la sélection d'un protocole d'échantillonnage performant couplé à un nombre d'unités statistiques optimal dans l'échantillon. Ces derniers doivent permettre d'atteindre, avec un nombre minimal d'échantillons, une précision dans les résultats d'estimation assez fine pour pouvoir répondre à la problématique initiale, sans échantillons superflus. La méthode pourra être utilisée pour la mise en place d'un suivi efficace, ou pour optimiser un suivi déjà opérationnel. Une étape importante à effectuer en amont de son application sera de bien cibler les protocoles d'échantillonnage à tester. Lors de l'étude d'une variable spatialisée (abondance totale dans une aire de répartition par exemple), il a été prouvé que les protocoles d'échantillonnage spatialement équilibrés pouvaient être plus intéressants à utiliser que des protocoles plus simples. Ce premier chapitre est une revue de ces protocoles spatialement équilibrés.

Certaines études environnementales utilisent des protocoles d'échantillonnage non probabilistes pour sélectionner des échantillons dans des populations spatiales. Malheureusement, ces échantillons peuvent s'avérer complexes à analyser statistiquement et les résultats finaux d'estimation de la population d'intérêt peuvent être biaisés. Les protocoles d'échantillonnage spatialement équilibrés sont des protocoles probabilistes, qui permettent de bien distribuer les échantillons sélectionnés dans l'aire d'étude. Ces protocoles sont particulièrement utiles pour les études environnementales puisqu'ils produisent une bonne couverture spatiale de la ressource, ont des estimateurs précis de la population d'intérêt et peuvent potentiellement réduire le coût total du suivi. Le protocole d'échantillonnage spatialement équilibré les plus populaire est le "generalized random tessellation stratified" (GRTS), qui a plusieurs capacités intéressantes comme celle de générer des échantillons spatialement équilibrés, d'avoir des estimateurs connus et d'avoir la capacité de générer une liste de points de secours. Ce papier considère la popularité de l'échantillonnage spatialement équilibré en examinant plusieurs protocoles et en montrant que ceux-ci peuvent être implémentés sous le logiciel de programmation R. Nous espérons augmenter la visibilité de l'échantillonnage spatialement équilibré et encourager les environnementalistes à utiliser ces protocoles.

Compétences développées/utilisées dans le cadre de ce chapitre :



Valorisations de ce chapitre :

- **Publication :**

KERMORVANT Claire, D'AMICO Frank, BRU Noëlle, CAILL-MILLY Nathalie, ROBERTSON Blair. Spatially balanced sampling designs for environmental surveys. Environmental Monitoring and Assessment, 2019, vol. 191, p. 524. https://doi.org/10.1007/s10661-019-7666-y

- **Poster en conférence internationale :**

KERMORVANT Claire, BRU Noëlle, CAILL-MILLY Nathalie, D'AMICO Frank. De nouveaux packages pour sélectionner des points d'échantillonnage spatialement équilibrés sous R. Les sixièmes rencontres R. Juin 2017 - **Prix du meilleur poster**. Disponible en annexe.

Abstract

Some environmental studies use non-probabilistic sampling designs to draw samples from spatially distributed populations. Unfortunately, these samples can be difficult to analyse statistically and can give biased estimates of population characteristics. Spatially balanced sampling designs are probabilistic designs that spread the sampling effort evenly over the resource. These designs are particularly useful for environmental sampling because they produce good sample coverage over the resource, have precise design-based estimators and they can potentially reduce the sampling cost. The most popular spatially balanced design is generalized random tessellation stratified (GRTS), which has many desirable features including a spatially balanced sample, design-based estimators and the ability to select spatially balanced oversamples. This article considers the popularity of spatially balanced sampling, reviews several spatially balanced sampling designs and shows how these designs can be implemented in the statistical programming language R. We hope to increase the visibility of spatially balanced sampling and encourage environmental scientists to use these designs.

When sampling an environmental resource, it is important to randomly choose the sampling locations over the study area to provide formal statistical inference from the sample to the population. Smith, Anderson, et Pawley (2017) established that in ecology, 12% of field studies selected samples using simple random sampling (SRS) and 9% used systematic sampling. These methods are probabilistic sampling designs (meaning there is an element of randomness in selecting their samples) and have well established statistical properties. However, most of the ecological studies in Smith et al.'s (2017) review were not probabilistic sampling designs. Some studies used haphazard or subjective judgement sampling methods and some studies did not specify how their samples were drawn (Smith et al. 2017). This is troubling because data gathered in a haphazard or subjective way can produce unrepresentative samples and biased estimates of population characteristics (Albert et al. 2010 ; Levy et Lemeshow 2013). Choosing an appropriate sampling design for a particular study can be difficult and there is no best design for all research questions (Kenkel, Juhász-Nagy, et Podani 1990 ; Stehman et Overton 1994). This choice depends on many things including the study objectives, available sampling frames and known auxiliary variables. This paper focuses on making an inference from a sample to the entire population using a specific class of probabilistic sampling designs called spatially balanced sampling designs. These designs were chosen because they are particularly useful for sampling natural resources (Stevens et Olsen 2004). For a full treatment on the subject, the reader is referred to Benedetti, Piersimoni, et Postiglione (2015).

What is spatially balanced sampling ?

To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the spatial pattern of the population. However, the spatial pattern of the response variable may not be known before the sample is drawn. Fortunately, a common spatial feature in environmental sampling is that nearby locations tend to be more similar because they interact with one another and are influenced by the same set of factors (Stevens et Olsen 2004). Therefore, an effective strategy is to spatially spread the sample evenly over the resource. A sample that is evenly spread over the resource is called a spatially balanced sample. Stevens et Olsen (2004) introduced the phrase spatially balanced sampling and proposed a statistic that measures the spatial balance or regularity of a sample using Voronoi polygons.

Why should environmental scientists use spatially balanced designs ?

The potential advantages of using spatially balanced sampling have been demonstrated in the field of environmental science (Stevens et Olsen 2004 ; Christianson et Kaufman 2016 ; McGarvey, Burch, et Matthews 2016). The first advantage is that spatially balanced samples are evenly spread over the resource. Covering the resource avoids under-coverage and over-coverage, which can happen with probabilistic sampling designs with poor spatial balance (Stevens et Olsen 2004 ; Christianson et Kaufman 2016). If a researcher's analysis requires clusters of nearby observations, spatially balanced cluster sampling could be useful (Robertson et al. 2017). Spatially balanced samples can be very efficient when the response variable

has a strong spatial trend (Stevens et Olsen 2004 ; Barabesi et Franceschi 2011 ; Anton Grafström et Lundström 2013 ; Robertson et al. 2013 ; Benedetti, Piersimoni, et Postiglione 2017), because their design-based estimators take into account spatial heterogeneity (J.-F. Wang, Stein, et al. 2012) and spatial auto-correlation (Haining 2003). When estimating a population total or mean using the Horvitz-Thompson estimator, the local mean variance estimator (Stevens et Olsen 2003) is a popular variance estimator. There have been many studies showing the effectiveness of spatially balanced sampling with this estimator on a variety of populations with different spatial structures (c.f. Stevens et Olsen (2004) ; Anton Grafström, Lundström, et Schelin (2012) ; Anton Grafström et Lundström (2013) ; Robertson et al. (2013) ; Robertson et al. (2018)). If the spatial trend is weak or if there is no trend at all, there is no statistical advantage in choosing spatially balanced designs over other probabilistic designs (Robertson et al. 2013). Another potential advantage of spatially balanced sampling is reduced sample cost. As mentioned above, spatially balanced designs can produce precise design-based estimators when there is a spatial trend in the response variable. Hence, to achieve a desired level of precision, fewer observations may be required when spatially balanced sampling is used. This was illustrated by Kermorvant et al. (2017) ; Kermorvant et al. (2019) for a clam monitoring program in Arcachon Bay, France. They showed that when spatially balanced designs were used, the total survey cost was reduced by 30% when compared with simple random sampling. Another useful feature of some spatially balanced designs is that they can draw spatially balanced oversamples (replacement units). This is particularly useful for environmental sampling because when sample units cannot be observed (private property, inaccessible, too dangerous, etc.), replacement units are often required to achieve the desired sample size (Stevens et Olsen 2004 ; Robertson et al. 2018 ; Theobald et al. 2007). For example, in the Oklahoma state-wide stream and river monitoring program, only 130 of the 177 randomly chosen sites could be observed (Board 2013). Of the unobservable sites, eight were on private land and 39 had dry channels or were not accessible. To achieve the desired sample size, the researchers selected replacement sites from an oversample drawn using simple random sampling. The potential advantage of using spatially balanced oversamples is that the observed sample maintains some degree of spatial balance over the observable resource (Stevens et Olsen 2004 ; Robertson et al. 2018). Although oversampling is of practical importance, it does not eliminate the non-response from unobservable units or the bias of an inference (Robertson et al. 2018).

Generalized Random Tessellation Stratified

Generalized Random Tessellation Stratified (GRTS) (Stevens et Olsen 2004) is the most popular spatially balanced sampling design for environmental studies (Anton Grafström et Tillé 2013 ; Foster 2016). It was developed by the U. S. Environmental Protection Agency (EPA) for the National Environmental Monitoring and Assessment Program (Messer, Linthurst, et Overton 1991 ; Stevens et Olsen 2004). GRTS uses a complex algorithm to draw its sample and we briefly discuss its main steps here. The reader is referred to Olsen, Kincaid, et Payton (2012) for a full, non-technical description of GRTS. Initially, a grid is superimposed over the study area and each grid cell is hierarchically numbered using a base four numbering system. The numbered grid cells are then randomly permuted using reverse hierarchical ordering and mapped (in order) to the real line. A systematic sample from the ordered grid

cells is then drawn and one sampling unit is randomly selected from each of these grid cells. The selected units are then mapped back to their respective locations in the study area, to yield the sample locations. The systematic sampling and hierarchical ordering that GRTS uses ensures that the sample is spatially balanced (Stevens et Olsen 2004). To investigate GRTS's popularity, the Google Scholar search engine was used because it provided access to a wide range of publication types. It indexes journals papers (published online or in paper format), conference proceedings, posters, and technical reports from research organizations in both the public and private sector. The keyword "GRTS" was used for the search and we did not include citations from 2018. All the documents found were categorised as either "publications" or "reports". Publications included peer-reviewed journal articles and refereed conference proceedings, and reports included all other publication types. Results are displayed in Figure 5.

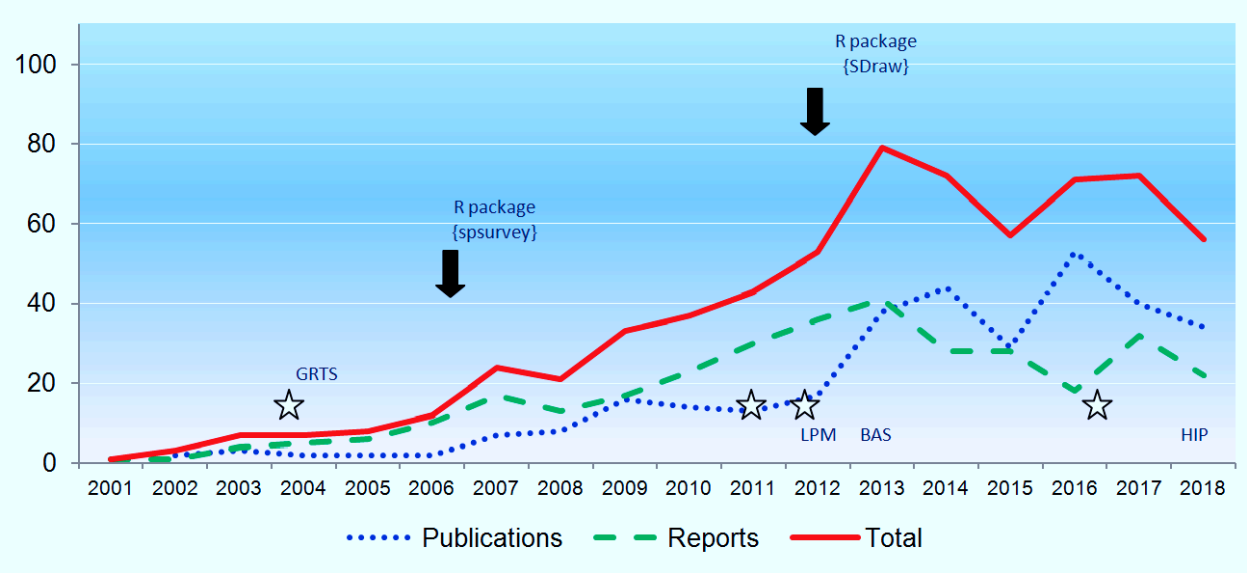


FIG. 5 : Flow representation of use and/or citations of GRTS in the literature, publication date (stars) of several spatially balanced designs and R packages (arrows) are shown.

Our analysis found 600 documents citing GRTS throughout the world. The citation and/or use of GRTS showed a steady increase until 2013, after which it flattened out. At the beginning, there were more reports than publications, but this trend appears to have reversed since 2014. Most of the documents found were in the fields of environmental science (mostly ecology but also environmental chemistry) and statistics (new designs, tests and comparisons). There were only two publications from other fields. The first was in economics, where GRTS was compared to existing sampling designs for business surveys (Dickson et al. 2014) and the second was a thesis on sampling standards for maintenance management quality assurance (Chen 2018).

Other spatially balanced sampling designs

Many spatially balanced designs have been proposed in the literature. In this section we mention several approaches. The Local Pivotal Method (LPM) (Anton Grafström, Lundström, et Schelin 2012) is a flexible spatially balanced design that can draw equal and unequal probability samples in multiple dimensions. Unequal probability sampling can be more efficient than equal probability sampling if there is a positive correlation between the inclusion probabilities and the response values (Robertson et al. 2013). Additional dimensions could include auxiliary information such as ecological threats, time intervals, species population structure or environmental data (Brown, Robertson, et McDonald 2015). The Swedish national forest inventory, for example, has implemented LPM with five auxiliary variables (Grafström et al. 2017). LPM is a popular method with 100 citations (using Google Scholar), where most of its applications were related to forestry. Anton Grafström (2012) also presented spatially correlated Poisson sampling (SCPS). This design is a modification of correlated Poisson sampling (Bondesson et Thorburn 2008b) that draws spatially balanced samples. LPM is algorithmically easier than SCPS, but SCPS may produce better results for some populations (Grafström et Schelin 2014). Balanced acceptance sampling (BAS) (Robertson et al. 2013, 2017) is another spatially balanced design. BAS uses the Halton sequence (Halton 1960) to spread its sample across multiple dimensions. BAS is conceptually simple, computationally efficient and is particularly useful for drawing spatially balanced oversamples (Robertson et al. 2018). We found 34 publications citing BAS, where most of the papers were methodological rather than applied. BAS has been used to survey bats in Bighorn Canyon National Recreation Area (Keinath et NRA 2016) and is being used for New Zealand's national monitoring program (van Dam-Bates, Gansell, et Robertson 2018). BAS is well suited for areal resources (geographic areas), but it can be inefficient on some point resources (Robertson et al. (2018)). To improve the performance of BAS on point resources, Robertson et al. (2018) presented Halton iterative partitioning (HIP). This spatially balanced design uses properties of the Halton sequence to partition a point resource into nested boxes to draw its sample, rather than using the sequence itself. Benedetti et Piersimoni (2017) presented a flexible class of spatially balanced designs that draw their samples based on a within sample distance (Benedetti et Piersimoni (2017)). The algorithm is simple to implement in multiple dimensions and any distance or similarity measure can be used to define the within sample distance. Spatially balanced sampling packages in R Several R software (Team 2014) packages are freely available to draw spatially balanced samples. To draw GRTS samples, spsurvey(Kincaid et Olsen 2015) or SDraw (McDonald 2016) can be used. These packages can draw samples from point resources (geographic locations), linear resources (rivers) and areal resources (geographic areas), and can also draw spatially balanced oversamples. The spsurvey package can also draw stratified spatially balanced GRTS samples with user defined strata.
The other spatially balanced designs mentioned in this article can be selected using the following packages. BalancedSampling (Grafström et Lisic 2016) draws equal and unequal probability LPM and SCPS samples from point resources. BAS and HIP samples/oversamples can be selected from point, linear and areal resources using SDraw (McDonald 2016). Historical or legacy sites can also be incorporated into a BAS design (Foster et al. 2014) using the MBHdesign package (Foster 2016). Finally, the R package Spbsampling (Pantalone, Benedetti, et Piersimoni 2019) can be used to draw the within sample distance-based methods of

Benedetti et Piersimoni (2017). Figure 6 shows examples of equal probability spatially balanced samples of 150 points drawn from a point resource using SDraw and BalancedSampling. An oversample of 20 points is also illustrated for BAS and GRTS. Note how the oversample points are spatially balanced with respect to the primary sample. To illustrate the R syntax for these packages, an annotated R script that creates Figure 6 is given in the supplementary material section.

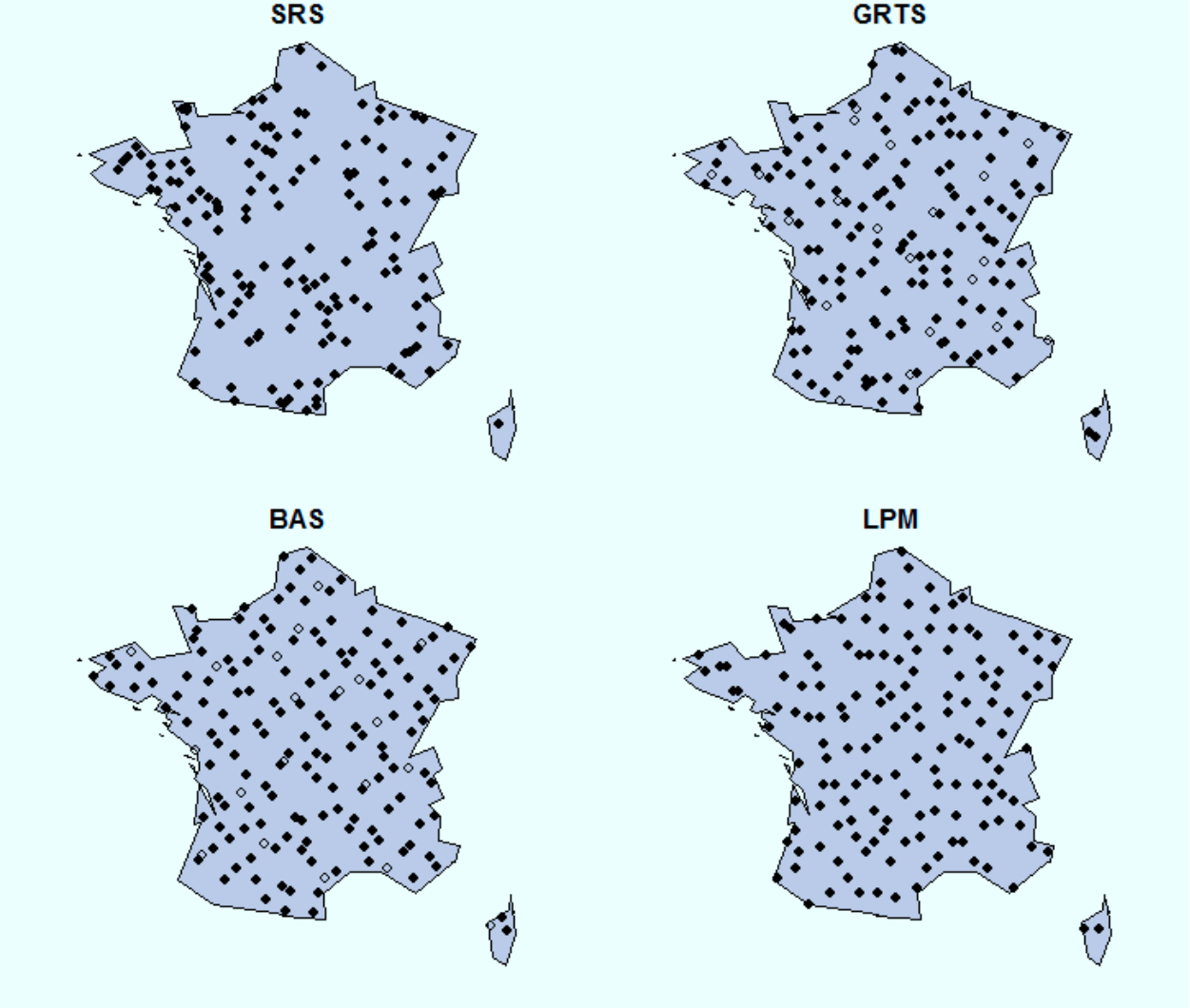


FIG. 6 : Several spatially balanced samples drawn using different designs, where open symbols denote oversamples sites.

Conclusion

Environmental scientists are beginning to use more advanced sampling designs to achieve robust statistical results. Spatially balanced designs are particularly useful for environmental science because they can produce good sample coverage over a resource, precise design-based estimators and potentially reduce sampling cost. GRTS is the most popular spatially balan-

ced sampling design and it is easy to implement using freely available R packages like spsurvey and SDraw. Another useful feature of GRTS is that spatially balanced oversamples can be drawn. Although oversampling is of practical importance, it does not eliminate the nonresponse from unobservable units or the bias of an inference. Several other spatially balanced designs are also available, each with an accompanying R package. LPM and SCPS samples can be drawn using BalancedSampling and SDraw selects BAS and HIP samples/oversamples. Although spatially balanced sampling has mostly been used in ecology, we encourage all environmental scientists to these designs when the research objective is to make an inference from a sample to the entire population.

Acknowledgments

We thank Jennifer Brown, Trent McDonald, Anton Grafström and Roberto Benedetti, and anonymous referees for valuable comments that improved this paper. This work was supported by "Communauté d'Agglomération Pays Basque - Euskal Hirigune Elkargoa" through a thesis grant.

Supplementary informations

Using R to draw GRTS samples

Several GRTS functions are available, for example GRTS() from the "spsurvey" package (Kincaid et Olsen 2016) and grts.line(), grts.point() and grts.polygon() (one for each frame type) from the "SDraw" package (McDonald 2016). They are easy to use and only require a shapefile (file format for GIS maps) and the output file. Once this information has been entered, the functions output the GPS coordinates of the selected locations. An example of an R script using the SDraw and the spsurvey packages is given below.

SDraw package

Draw a GRTS sample from a polygon shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#(where the polygon shapefile is located).
setwd(Chemin)

Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size

samples<-grts.polygon(Shape, nb, over.n)
plot(Shape, col=rainbow(length(Shape)))
points(samples, pch=16 )
```

```
samples
```

Draw a GRTS sample from a line shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#(where the line shapefile is located).
setwd(Chemin)

Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size

samples<-grts.line(Shape, nb, over.n)
plot(Shape, col=rainbow(length(Shape)))
points(samples, pch=16 )
samples
```

Draw a GRTS sample scheme from a point shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#(where the point shapefile is located).
setwd(Chemin)

Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size

samples<-grts.point(Shape, nb, over.n)
plot(Shape, col=rainbow(length(Shape)))
points(samples, pch=16 )
samples
```

spsurvey package

```
library(spsurvey)
library(fields)
library(sp)

Chemin <- "C :/Users/xxx/Desktop"
#set your working directory (where the shapefile is located);
```

```
setwd(Chemin)

att<-read.dbf("name of your shapefile without the extension")

nrow<-10    #set the sample size
  #Unstratified, equal probability sampling
  Equaldsgn <- list(None=list(panel=c(PanelOne=nrow),
                              seltype="Equal")
  )

  Equalsites <- grts(design=Equaldsgn,
                     src.frame="shapefile",
                     in.shape="name of your shapefield without the extension",
                     att.frame=att,
                     type.frame="area",
                     DesignID="EQUAL",
                     shapefile=TRUE,
                     out.shape= "my shape"
  )

# coordinates of sample sites
samples<-cbind(Equalsites$xcoord,Equalsites$ycoord)
samples
```

Save your survey plan in a text file

```
write.table(samples,file="myGRTSsurveyplan.txt")
```

Using R to draw BAS samples

The SDraw R package can be used to draw BAS samples. It works in a similar way to GRTS : the practitionner inputs a polygon, line or point shapefile and uses bas.polygon(), bas.line() or bas.point() respectively, to draw the BAS sample. To draw BAS oversamples, you simply increase the number of points drawn and use the ordered output. For example, a sample of size n will be the first n points in the ordered output.

SDraw package

Draw a BAS sample from a polygon shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#where the polygon shapefile is located).
setwd(Chemin)
```

```
Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size

samples<-bas.polygon(Shape, nb, over.n)
plot(Shape, col=ifelse(samplesbas$sampleID <= nb ,"black","red"))
points(samples, pch=16 )
samples
```

Draw a BAS sample from a line shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#where the line shapefile is located).
setwd(Chemin)

Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size

samples<-bas.line(Shape, nb, over.n)
plot(Shape, col=ifelse(samplesbas$sampleID <= nb ,"black","red"))
points(samples, pch=16 )
samples
```

Draw a BAS sample from a point shapefile

```
library(SDraw)
library(maptools)

Chemin <- "C :/Users/xxx/Desktop"  #set your working directory
#where the point shapefile is located).
setwd(Chemin)


Shape<-readShapeSpatial("name of your shapefile.shp")
nb<- 10                          # set the sample size
over.n<-10                       # set the oversample size
samples<-bas.point(Shape, nb + over.n)
plot(Shape, col=ifelse(samplesbas$sampleID <= nb ,"black","red"))
points(samples, pch=16 )
samples
```

Figure within the manuscript

The following R script was used to create Fig.2.

```
library(maptools)
data(wrld_simpl)
sp_area <- subset(wrld_simpl, NAME=="France")


nb<- 150                          # set the sample size
over.n<-20                       # set the oversample size

par(mfrow=c(2,2))
par(mar=c(1, 1, 1, 1))

samplessrs<-SDraw ::srs.polygon(sp_area, nb)
plot(sp_area, col="gray80", main= "SRS")
points(samplessrs, pch=16 )
samplessrs

samplesgrts<-SDraw ::grts.polygon(sp_area, nb, over.n)
plot(sp_area, col="gray80", main= "GRTS")
points(samplesgrts, pch= ifelse(samplesgrts$pointType == "Sample",16,1) )
samplesgrts

samplesbas<-SDraw ::bas.polygon(sp_area, nb+over.n )
plot(sp_area, col="gray80", main = "BAS")
points(samplesbas, pch= ifelse(samplesbas$sampleID <= nb ,16,1))
samplesbas

#LPM need a finite population; let's grided it.
area <- makegrid(sp_area)
spgrd <- SpatialPoints(area, proj4string = CRS(proj4string(sp_area)))
spgrdWithin <- SpatialPixels(spgrd[sp_area,])
sample_units<-length(spgrdWithin@coords[,1])

#plot(spgrdWithin, add = T)
sampleslpm<-BalancedSampling ::lpm2(rep(nb/sample_units,sample_units),
                                    spgrdWithin@coords)
plot(sp_area, col="gray80", main = "LPM")
points(spgrdWithin@coords[sampleslpm,], pch=16)
sampleslpm
dev.off()
```

Résultats issus de ce chapitre :

- Aujourd'hui très peu de suivis en écologie utilisent un protocole probabiliste, les résultats sont biaisés et sujets à controverses.
- Nous avons étudié la popularité des protocoles spatialement équilibrés, toutes matières scientifiques confondues, il semble y avoir une prise de conscience des scientifiques qui les utilisent de plus en plus, notamment en sciences environnementales. Mais cette utilisation reste encore marginale.
- Avec le développement des géostatistiques, il est maintenant simple d'utiliser des protocoles complexes. Nous avons édité des exemples de scripts pour leur utilisation sous le logiciel R.
- Nous espérons augmenter la visibilité et l'utilisation des protocoles spatialement équilibrés, connus pour leurs hautes performances.

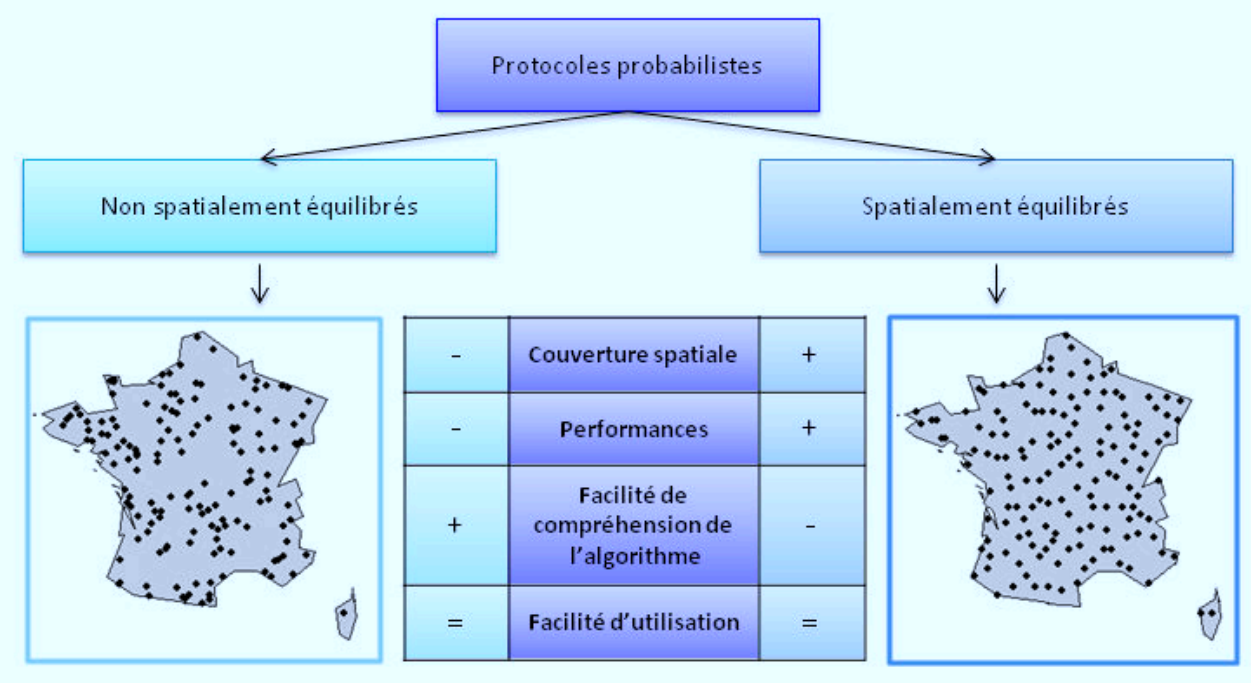


Fig. 7 : Conclusions du CHAPITRE I

CHAPITRE II : Une méthode générale pour choisir le protocole d'échantillonnage et le nombre de relevés garantissant un suivi efficace et rentable

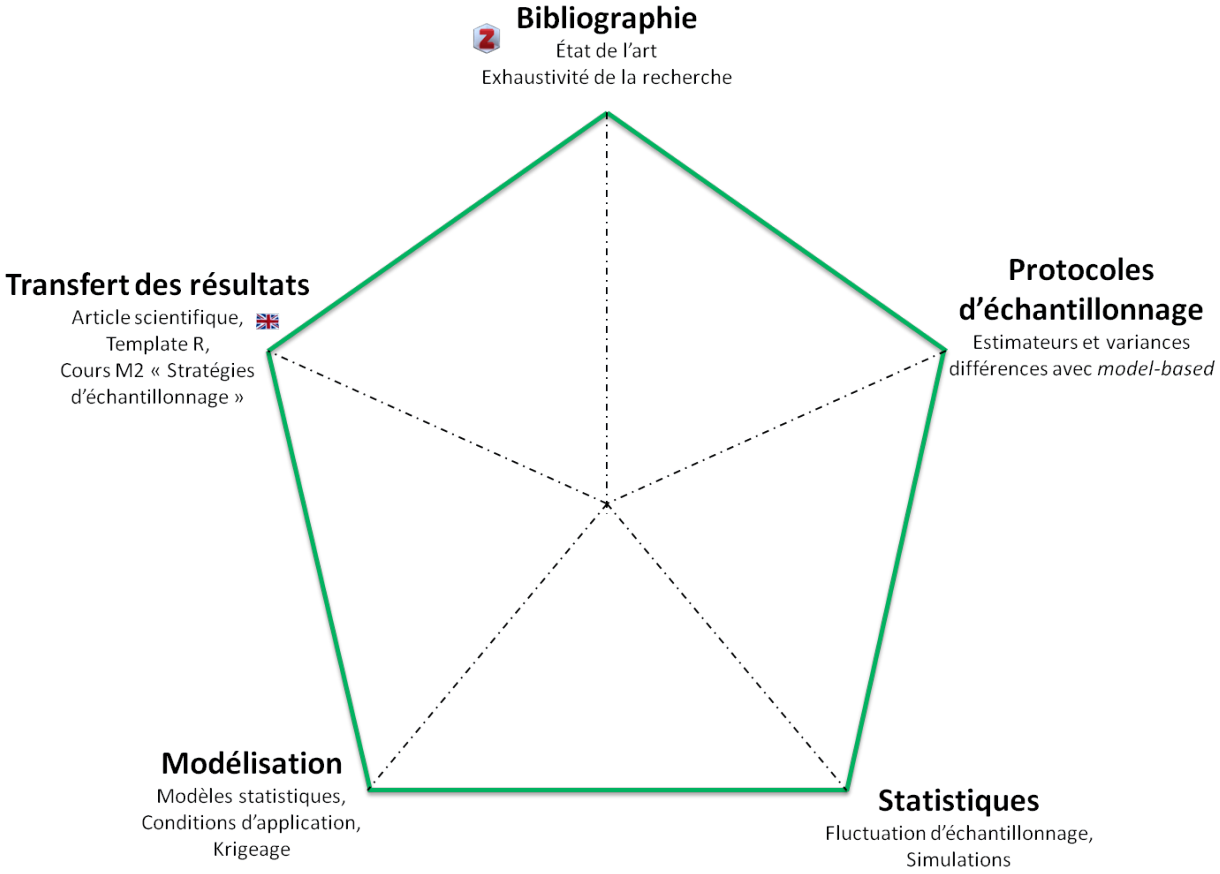
Synopsis :

Des problèmes dans la définition des procédures environnementales mènent de nombreux résultats d'études à être biaisés et sujets à controverses. Choisir un protocole d'échantillonnage et un nombre d'unités statistiques pertinents est une tâche difficile dans la mise en place ou l'optimisation d'un suivi. Le choix du protocole d'échantillonnage est important pour éviter le biais et augmenter le rapport coût /efficacité du suivi. Il peut avoir un fort effet sur le nombre d'échantillons nécessaires à atteindre une précision voulue dans les résultats finaux, et donc, dans le coût final de la procédure.

Dans ce chapitre, nous proposons une procédure séquentielle, mélangeant les théories de l'échantillonnage *design based* et *model based*. Cela pour aider à choisir un protocole d'échantillonnage et un nombre d'échantillons pour les suivis, quand l'inférence à la population est requise. L'idée principale est de reconstruire mathématiquement la distribution de la population que l'on cherche à suivre, puis de tester et comparer le rapport coût/efficacité de nombreuses procédures d'échantillonnage. La procédure permet de déterminer plusieurs niveaux de précision à atteindre dans les résultats finaux et de prendre en compte des données déjà disponibles sur la population. Les résultats issus de cette procédure sont un protocole d'échantillonnage couplé à un nombre d'unités d'échantillonnage optimaux à récolter sur le terrain pour atteindre une précision voulue. Cette précision est donc atteinte sans excès de points relevés.

Nous discutons de comment des données disponibles peuvent améliorer le suivi, depuis le cas où un jeu de données issues de plusieurs saisons de suivis sont disponibles jusqu'au cas où aucune données ne le sont. La force de cette procédure est qu'elle est basée sur des simulations. Cela permet d'essayer et de comparer un grand nombre de combinaisons protocole/nombre d'unités/précision à atteindre à moindre coût. Comme cela, l'utilisateur peut choisir la meilleure combinaison pour son suivi.

Compétences développées/utilisées dans le cadre de ce chapitre :



Valorisation de ce chapitre :

- **Publication :**

KERMORVANT Claire, D'AMICO Frank, BRU Noëlle, CAILL-MILLY Nathalie. Sequential process to choose efficient sampling design based on partial prior information data and simulations. *Submmitted in Ecography*

Abstract :

Issues on sampling procedure definition led numerous study results to be biased and object of controversy. Choosing relevant sampling design and number of samples is a difficult task when wanted to set up or optimize a survey. The survey design choice is very important to avoid bias and increase the survey cost-efficiency. It can have a strong effect on the sample size needed to achieve some targeted accuracy on results, and so, on the final cost of the procedure.

We propose a sequential process, melting design based and model based sampling theories, to help practitioners defining a sampling design and a number of samples for their survey where inference to the whole population is wanted. The main idea is to mathematically reconstruct the distribution of the surveyed population and then assess and compare cost-effectiveness of various sampling designs on this population. This process allows setting predetermined level(s) of accuracy to be reached in the targeted estimates and to take into account previous relevant data. Results are an optimal sampling design and an associated optimal sample size for a desired accuracy in the results. This accuracy is so achieved without excess sampling.

We discuss how to use available data to improve the survey, from the case were several historical data are provided to the case where no data are available. Strength of this process is that it is based on simulations. This allows trying a high number of combinations between sampling design, sample size and desired levels of accuracies. Sampling design performances can thus be compared. Thus, the user can decide which combination is the best for his survey and apply it for real.

Introduction

Monitoring programs are tools used in environmental science in three main tasks : to detect a change into a system, to measure success or failure of management actions and to identify effects of perturbations or disturbances (Legg et Nagy 2006). Likens et Lindenmayer (2018) reviewed the term "monitoring" in the ecological literature between 1985 and 2016 and more than 131 000 articles and numerous books were returned. Monitoring information is essential answering most ecological and environmental questions (Albert et al. 2010). For example they can be the basis for restoration programs or for endangered species conservation.

Sampling is very common because exhaustive information cannot be collected for almost all the cases. But its theory is complex and several environmental scientists are not trained to it. Thus, programs suffer from lack of details of problematic definition, hypothesis formulation, adapted sampling design (Smith, Anderson, et Pawley 2017) and so data quality (Legg et Nagy 2006). Poor method has numerous undesirable effects that can lead in the failure of a monitoring program (Legg et Nagy 2006). Issues with poor designs used in ecological studies often have led to significant controversy (Hayward et al. 2015) and ca have dramatics effects on estimates of populations trends (Fournier, White, et Heard 2019). It also means that it becomes difficult to evaluate management actions and results are not very useful for decision making (Vos, Meelis, et Ter Keurs 2000b). Roberts (1991) and Nichols et Williams (2006) deplore too many monitoring are "planned backward on the collect now (data), think-later (of a useful question) principle". A forum (Hayward et al. 2015) wrote after conflicting results were published in high-quality scientific journal. It emphases robust methods and appropriate experimental designs must be developed and used by practitioners, avoiding controversy in studies results.

Problematic understanding

Two theories are opposed for sampling and inference for finite populations : model-based and design-based sampling (Särndal et al. 1978). In model-based sampling theory, the population values are assumed to be generated by a stochastic model (J.-F. Wang, Stein, et al. 2012). The goal is so to find out this model to be able reconstructing the population. In design-based sampling theory the population values are unknown but fixed (J.-F. Wang, Stein, et al. 2012). Classical books of this theory are Cochran (1977) and Neyman (1934). Sampling design is the statistical tool used in design-based methods to choose samples within the studied statistical population. To ensure experimental designs to be robust and data integrity, at least two parameters of sampling procedure must be under particular attention : the sample size and the sampling design.

Challenges for environmental surveys

The choice of a sample size and a sampling design is a very important step in the establishment of a survey. Representativeness is brought by the random property of the sampling design (McDonald 2003 ; Sica 2006), precision by collecting enough data through a substantial sample size (Lohr 2009). A substantial sample size will increase precision on population estimation but may also increase the survey cost. This is particularly true in ecology where

sampling on field necessitates human and/or expert resources, sometimes expensive gears (boats, trucks…) differing than, for example, in sociology where sampling can just be a sample on the web. In the great barrier reef monitoring (Kang et al. 2016), divers must have the necessary skills and qualifications to do the monitoring and statisticians must curate the data. The practitioner has to found a trade-off (Stehman et Overton 1994) between a sufficient amount of sample, to achieve a precise estimation, and a price that will be reasonable for financers. Before constructing any survey procedure, clear objectives about total sample size and estimator quality have to be fixed. Priority can be given to maximize estimator quality (it's accuracy) or minimize total sample size (Guillera-Arroita, Ridout, et Morgan 2010) because no survey designs will be good for all purposes (Kenkel, Juhász-Nagy, et Podani 1990).

In ecology, the studied population is almost always a spatial population because species always display spatial distribution. With new developments in statistics and geostatistics (model-based), a large amount of probabilistic sampling designs (design-based) was developed last decades (McDonald 2014). Now, a significant number of theses tools are available and the issue is that it may be very tricky to determine which one is better to use for each surveys. Probabilistic sampling designs, displaying a random property, must be used for design-based sampling. Simple random sampling design, systematic sampling design, generalized random tessellation stratified (Stevens et Olsen 2004) sampling design, Balanced acceptance sampling (Robertson et al. 2013 , 2017), Halton iterative partitioning (Robertson et al. 2018), spatially correlated Poisson sampling (Anton Grafström 2012) and Local Pivotal Method (Anton Grafström, Lundström, et Schelin 2012) are good example of probabilistic sampling designs. Financial constraints are the main reason given for using qualitative (poor) methods (Legg et Nagy 2006), which does not guarantee survey success. Cost-efficiency of surveys is under scrutiny.

We are not the first team that is interested into cost-efficiency optimisation of monitoring programs (see, for examples : Perry et al. 2002 ; Field, Tyre, et Possingham 2005 ; Rudders 2011 ; Guillera-Arroita et Lahoz-Monfort 2012 ; Liberts 2013 ; Carvalho et al. 2016 ; Moore et McCarthy 2016 ; Vicente et al. 2016). In our mind, theses methods are very relevant and deserve to be used but they all are targeted to a specific problem. We need to develop a method to allow.

So, we develop a sequential process to help researchers and practitioners choosing the more performing sampling design and the optimal sample size for their studies. For this purpose, clear results about number of samples, final precision achieved and total survey cost must be available. The original part is that this process allows for taking into account prior knowledge, melting model and design-based sampling. The process starts from the study of available prior data. Final results are an optimal sampling design and sample size to achieve a fixed accuracy on results. From this, overall cost of the survey can be easily calculated. This methodology can be used to assess performances of any sampling design in any population already documented by field data.

Sampling issues to be taken into account

Values obtained from a sample differ from sample to sample because of random sampling fluctuations. This phenomenon is an important reason of observed differences between different samples configurations (Schwartz et Sagiv 1995), and so population estimates are almost always different from one sample to another. Sampling fluctuations is a cause of variance into sample and must be distinguished from meaningful variation of the targeted parameter (Fontaine et al. 2008). Sampling fluctuations are very difficult to determine because, in practice there is only one sample. They decrease when the sample size increases.

Reaching an acceptable precision on estimates results can also be problematic. The objective is to reach a high accuracy, i.e. a narrow variance around estimation result. MacKenzie (2006) defines accuracy as the sum of bias and precision. Bias is when the parameter value estimated with sampling is different from real population value. Bias must be removed and this can be done by using a probabilistic sampling design. The random characteristic of these designs ensures unbiased samples. Thus, with a probabilistic sampling design, accuracy of the sample only depends on precision. Precision is brought by collecting enough data through a substantial sample size (Lohr 2009). For better understanding, we will use only "accuracy" in this paper. Along with sampling fluctuations, accuracy on results increases (becomes narrower around real parameter value) while sample size increases. Population estimates calculated from a sample depend on sample size and on the values gathered at these sample units.

A substantial sample size will increase precision on population estimation and decrease sampling fluctuations but will also increase the survey cost. It is very important to well choose this sampling effort. The practitioner has to found a trade-off between a sufficient amount of sample, to achieve a precise estimation, and a price that will be reasonable for financers. Before constructing any survey procedure, clear objectives about total sample size and estimator quality have to be fixed. Priority can be given to maximize estimator quality (its accuracy) or minimize total sample size (Guillera-Arroita, Ridout, et Morgan 2010)).

Towards a cost efficiency procedure to have a gain on precision

The major challenge of this paper is to provide a sequential method permitting to choose optimal sampling design and optimal sample size for a survey. First of all, the process takes into account prior data. Prior knowledge of the studied area and the population can dramatically reduce the uncertainty in the sampling estimate (J.-F. Wang, Jiang, et al. 2012). Without this knowledge of the population distribution, optimization of the survey can be very tricky. This knowledge is very relevant to provide initial idea of quantification and spatial (and/or temporal) delimitations of the survey area (and/or duration). Secondly, these data will be the basis to compare several sampling designs to choose the most efficient on this population. Rajabi et Ataie-Ashtiani (2014) define performance (= efficiency) as the capacity of a design strategy to require fewer samples to reach a certain level of accuracy. In the comparison of various sampling designs, efficiency can be viewed as a measure of quality of these sampling designs (Brown 2003). Important issues with this challenge are to keep an acceptable accuracy on estimation results and emancipate from sampling fluctuations. Following this, we

choose to define the "optimal sampling design" as the more efficient design between those assessed. The "optimal sample size" is the corresponding sample size.

The developed process compares sampling designs efficiency and determines the most optimal of them. For each assessed sampling design, the method simulates a large number of samples (e.g $> 1\ 000$) of size n (to avoid sampling fluctuations) and calculates reached accuracy on targeted estimate for this n. If this accuracy is not smaller, or at least equal, than the accuracy fixed by the user, a greater n is assessed. Once the fixed accuracy is reached, the associated sample size is selected as optimal sample size for this sampling design. As sampling designs perform differently following the statistical population, optimal sample sizes are different following the used sampling design. The sampling design that has the smaller optimal sample size is chosen as the optimal sampling design and must be applied on field with its associated optimal sample size.

Theoretical framework

To understand the data

Let us denote Ω a finite statistical population composed of N elementary units ω. On a purely spatial research problem, statistical population would be the area of interest and an elementary unit would be a point, a line or a polygon. On a temporal research problem, statistical population would be a time lapse and an elementary unit would be a punctual date or a time interval. Finally, on a spatio-temporal problem, statistical population and elementary unit would be a combination of both spatial and temporal features.

Let us consider Y a numerical statistical variable of interest unknown on all statistical units ω_j of a spatial statistical population Ω. We will note $Y = Y_1; Y_2; ; Y_N$ all the possible values of Y on specific statistical units. We want to estimate a particular parameter of this variable. In this paper, we will focus on the total parameter. For example, if we work on total of abundance, Y would be the number of individuals. Let's note $T(Y)$ on Ω the total of Y variable on the statistical population. Performing an exhaustive survey of Y on Ω to calculate $T(Y)$ is almost impossible or may prove to be tricky (Chiarucci et al. 2003, MacKenzie 2006) because of time consuming (Cox, Cox, et Ensor 1997) and/or money lack (Theobald et al. 2007 ; Jackson et al. 2008 ; Lazarina et al. 2014). The common practice for dealing with this problem is trying to infer $T(Y)$ on the basis of samples from the statistical population Ω (MacKenzie 2006). In such way, we need to collect some information about Y on a sample S of n statistical units. We will note $y_1; y_2; ; y_n$ the values of Y sampled in one sample S of size n. To estimate the interest parameter $T(Y)$ on Ω from a sample S, we should then construct an estimator or choose between existing ones.

To understand the estimation procedure

The Horvitz-Thomson's estimator (Horvitz et Thompson 1952) is chosen for our total parameter estimation example because it is the best linear unbiased estimator (BLUE) (Tillé 2011). Horvitz-Thomson formula gives a linear estimator of total from samples, without bias and valid for all probabilistic sampling designs with non-null positive first order inclusion

probability (noted Π). It guarantees an unbiased estimator denoted $\hat{T_n}$ of $T(Y)$ total on Ω from a given sample of size n using the following formulae :

$$\hat{T_n} = \sum_{\omega_j S} \frac{Y_i}{\Pi_i}$$

WhereY_j is the value of Y for the statistical unit $_j$ and Π_j is the probability that elementary unit $_j$ is included in the sample S (called first order inclusion probability). For design with equal inclusion probabilities, Horvitz-Thompson's estimator of total only depends on number of statistical units in the population, number of samples and values taken by these samples. Inclusion probabilities are different depending on the used sampling design. They are known for classical designs and can be founded in sampling theory books (see Tillé (2011) for example).

Sequential process

We are interested on finding the optimal sampling design and associated sample size for a survey with prior data.

Original aspect

In this paper, the originality comes from the sequential method to choose efficient sampling design is designed from prior data to an optimized survey. The distribution of Y on the statistical population Ω should not be known at the beginning of the study. Prior data can come from a previous survey campaign or from a model and will help us reconstruct the targeted population. Then, from this reconstructed population we could optimize the survey. Thus, we use existing data to reconstruct $\hat{Y}$ on Ω. Taking into account available data on studied parameter when designing the study could improve survey efficiency. A simulation study is done to assess the optimal sampling design and sample size on this reconstructed $\hat{Y}$ on Ω. The aim is to obtain high quality estimates of population parameters with low samples (= low cost). Several frameworks are already developed for survey optimization (see introduction part) but most of them target a specific purpose and concern data where Y is known everywhere within Ω.

Data available and reconstruct Y on Ω

In the following we assume that Ω is a spatial domain. The entire distribution of Y variable on Ω is rarely known. For example, it is difficult to know the abundance of one species within its all living area. But, do we have some previous data on this space or not ? And, if yes, what can we do with such data ? These questions are the beginning of our process. We will now detail the two cases : data are available or they are not.

If previous data of Y were drawn with a probabilistic sampling design (design-based method), random ensures data independence. In this case estimates can be derivate directly from samples values, without assumption on Y distribution (Petitgas 2001). When data are not collected with a random process, a model of Y spatial structure need to be inferred. The

estimate is so model-based (Cochran 1977 ; Petitgas 2001). In this case, one type of model-based methods that can be used to reconstruct Y on Ω is geostatistical methods. These kinds of methods need available geo-referenced data. They use values of Y and calculate distances between theses values. Such class of methods computes a spatial interpolation using the information of the variable Y itself at one spatial point according to its neighbors. The more known and used geostatistical tool for spatial interpolation is the ordinary kriging method. Statistical assumptions to use kriging methods are related to the use of distribution parametric models for variogram construction. Webster et Oliver (1993) claim that, to built semi-variogram for an isotropic variable, $n = 150$ samples should suffice but $n = 225$ are needed to the variogram be reliable (i.e. with reasonable confidence interval). But they statute the question "how large sample is needed to obtain a satisfactory estimate of the semi-variogram ?" has, unfortunately, no easy answer. A practical rule is that variogram lags (distances between pairs at which the variogram is calculated) should both include at least $n = 30$ samples (Cressie 1985). The quality of the semi-variogram modeling comes from the size of the sample, the variety of distance between spatial units of the used sample, the marginal distribution of the data, anisotropy and trend (Oliver et Webster 2014). Another requirement for ordinary kriging is that the target variable must be normally distributed (Draper et Smith 1998). For skewed distributions, a linear estimate will not necessarily be of lower variance (Papritz et Dubois 1999 ; Petitgas 2001). Y can be transformed to become normal (Hengl, Heuvelink, et Stein 2004 through a generic framework). Otherwise, non-linear kriging methods must be used (Petitgas 1993 ; Gaus et al. 2003 ; Triantafilis et al. 2004).

When unable to use geostatistical approach, one way is to link data to exogenous variables known on Ω though a model (i.e. GLM, GAM...).One key assumption of this kind of models is that the residuals must be independent and identically distributed (Dormann et al. 2007). For these methods, data does not require to be geo-referenced but, for inference, there must be an establish link between Y values and exogenous variables X known on each statistical units of Ω. Main difficulties are to construct the statistical model. Statistical units and X variables must be independent. User must choose a distribution law and a link function. Then, goodness of fit to data has to be checked and predictions on space must have a slight confidence interval (Gregoire 1998). Problems may also rise when wanted to model from a qualitative variable and data are unbalanced on different modalities. As with any modeling approach, the interpretation and the quality of model output depend on the initial dataset and whether the model assumptions are met sufficiently (Guillera-Arroita et al. 2015).

When no data are available, a distribution model must be constructed elsewhere and adjusted onsite. Bayesian statistical models can also be used when no data are available (Choy, O'Leary, et Mengersen 2009). In this case, prior information may be obtained from expert knowledge. Initial model can be progressively updated once data are available. These two cases are not discussed in this paper. The proposed way to face the case when no dataset that could provide a representative image of Y on Ω is available, is to establish a pilot study (Legendre et al. 2002) . The pilot study results have to be as accurate as possible, so the larger possible number of samples has to be done. Concerning the sampling design for conducting a pilot study when the objective is to estimate population characteristics, we recommend the use of a spatially balanced sampling (SBS) design (Stevens et Olsen 2004 ; Anton Grafström, Lundström, et Schelin 2012 ; Robertson et al. 2013 , 2018). Because with

a same number of samples they reach a better accuracy in estimates than simple random sampling (Kermorvant et al. 2019). As SBS have a random component, a pilot study drawn with one of them and sufficient sample size ensures statistical requirements to geostatistical tools use, leading on a representative statistical population $\hat{Y}$ on Ω.

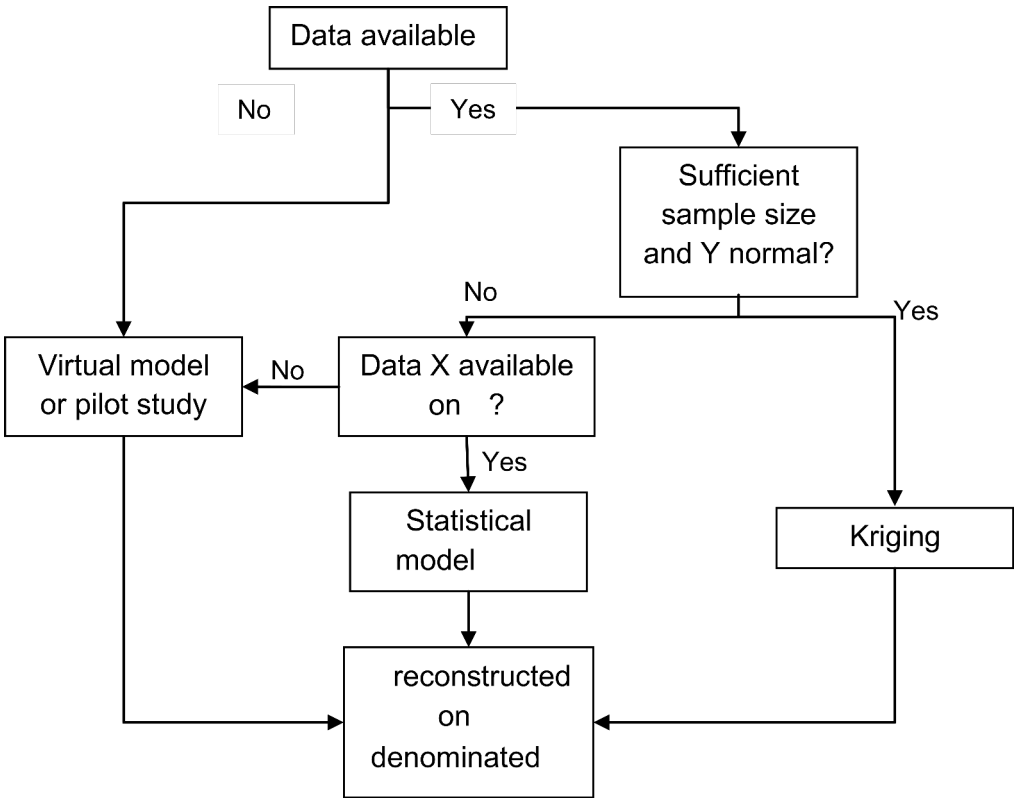


FIG. 8 : Process to reconstruct $\hat{Y}$ on Ω

Compare geostatistical and kriging methods predictive power

Cross-validation is a model validation technique mainly used when model goal is prediction. The dataset is divided into k groups, model is created on k-1 groups and the model's ability to predict new data is tested on the remaining group. Cross-validation statistical assumption is data are Independent and identically Distributed (i.i.d). Leave-One-Out (or LOO) is an exhaustive cross-validation where k equal to one sample. LOO cross-validation estimate of prediction error is calculated as the mean squared of the difference between the observed value and the predicted one :

$$CV_n = \frac{1}{n}{}_{j=1}^{n}(\frac{y_i - \hat{y_j}}{1 - h_j})^2$$

Where h_j is the diagonal element of the hat matrix. It tells how much influence an observation has on its own fit. It's a number between 0 and 1 that punishes the residual, because it divides by a number that's small, and it inflates the residual. The more this index is close to 0, the more the model's ability to predict value is good.

Example : spatial interpolation using the two methods from the same dataset

We choose "meuse" dataset from R package {sp}. It comprises four heavy metals measured in the top soil in a flood plain along the river Meuse. You can find further detail here : https://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf). We will focus on the spatial distribution of zinc concentration in soil (in ppm). As this dataset is geo-referenced and contains a set of covariate we can use the two previously presented methods.

Step 1 : Available data analysis

So applied to this problem, the statistical population Ω is the Meuse river watershed. We are interested on the zinc concentration parameter, previously called Y. 155 geo-referenced samples points $_j$ were done and are available (whatever the way of drawing them). Their spatial dispersion is represented on fig.9 a). Zinc concentration density is plotted on Fig.9 b), and log-transformation on Fig.9 c).

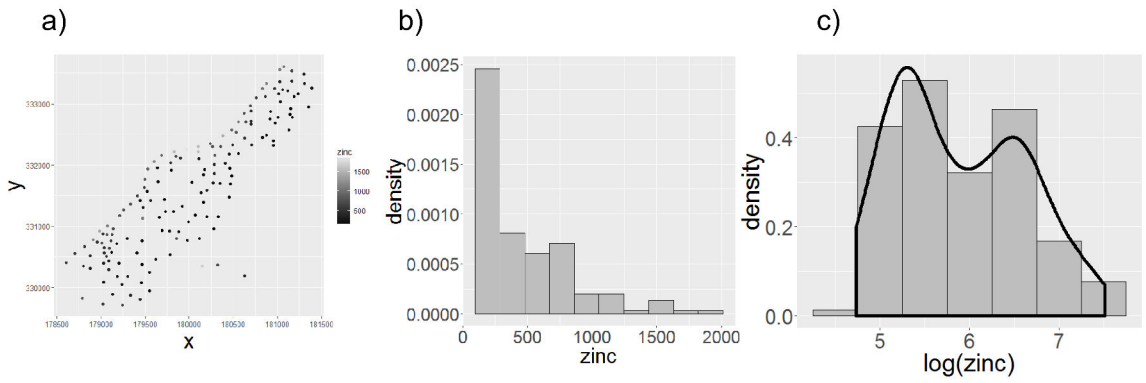


FIG. 9 : Visualisation of $\hat{Y}$

Step 2 : reconstruction of $\hat{Y}$ on Ω

First interpolation method presented in this paper is ordinary kriging. We have to check if statistical requirement for this method are followed by Meuse dataset before using it. A priori, 155 samples is enough and they display sufficient geographical distances. But Y variable, here zinc, is not normally distributed. We log transformed it to nearly approach a normal distribution and so one be more rigorous into using ordinary kriging method. It seems that we can use ordinary kriging method on zinc concentration to reconstruct it on the whole Meuse watershed. For result visualisation, we gridded the watershed into 3 103 cases. A semi-variogram is constructed with data (points on Fig.10 a) and a model is adjusted to it (smooth line on Fig.10 a). Here, model is exponential with a nuggets effect. This model is then used to interpolate zinc variable within the whole watershed.

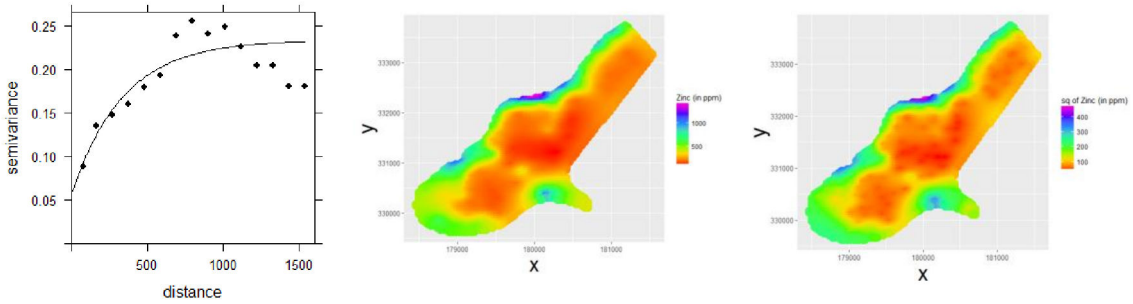
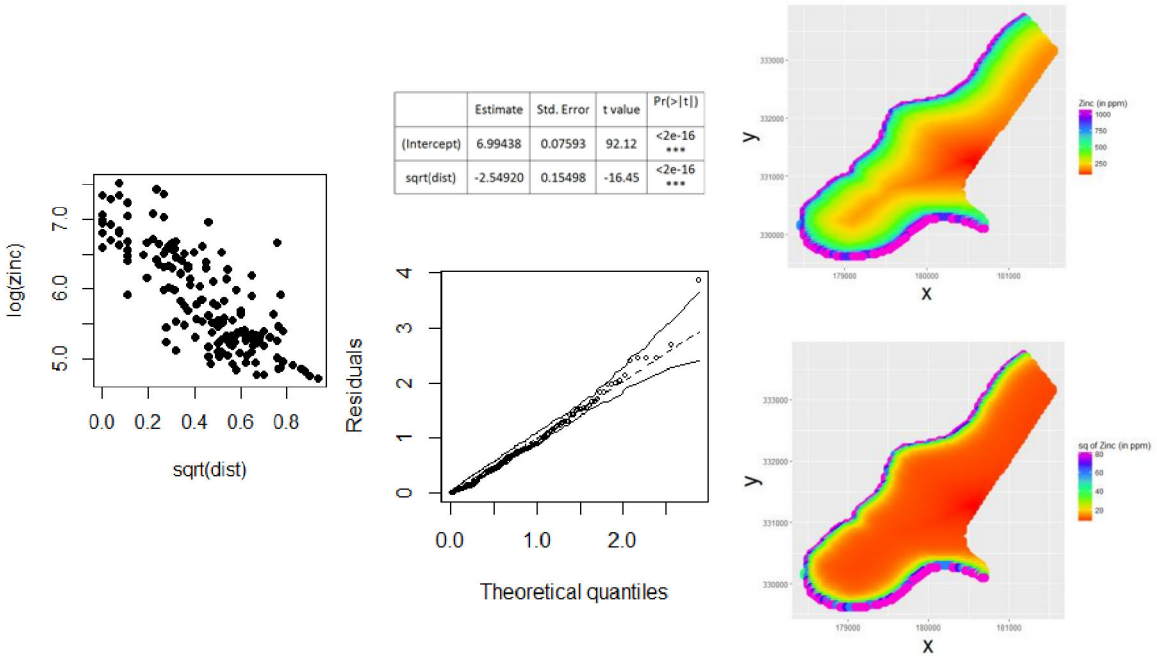


FIG. 10 : Spatial interpolation with ordinary kriging method (leave-one-out cross-validation = 0.1706)

Kriging is not the only method to reconstruct $\hat{Y}$ on Ω. Exogenous variable dist (previously called X) is also available for all statistical units $_i$ of . The statistical model-based $Y \sim X$ can be constructed. Requirement for statistical model method is to have an exogenous variable X that can be linked to explain Y and available on all statistical units of . Modelling is easier when there is a linear effect between X and Y (third plot). So here, data were log transformed and root-squared transformed. The statistical model used here (fig.11) is a linear model $loga(zinc = \alpha sqrt(dist)$. The tab in Fig.11 represents model summary. Residuals plot show the model fit well data. Last maps display prediction and related error of zinc concentration on Meuse watershed with the linear model.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.99438	0.07593	92.12	<2e-16 ***
sqrt(dist)	-2.54920	0.15498	-16.45	<2e-16 ***

FIG. 11 : Model fitting and spatial prediction (leave-one-out cross-validation estimate of prediction error = 0.1914)

We choose the Meuse example because the two model-based methods were applicable on it. Theoretically they both lead to a prediction map, but kriging method seems more accurate here since the only exogenous variable available on all Ω is the distance to river. We evaluate models predictive power by using leave-one-out cross-validation index. Kriging model display a cross-validation index little smaller than linear model and so should be preferentially applied for the Meuse example.
The distribution map of $\hat{Y}$ on Ω is the basis of the following step of our sequential process.

Compare sampling designs

Once the statistical population $\hat{Y}$ on Ω is reconstructed, we can start assessing performances of chosen sampling designs.

One by one sampling design simulation study of quality

As shown previously, the quality of a given sampling design can be assessed using efficiency of the corresponding estimator that depends on n. For a given sampling design, we will now show how to find the sample size that will permit to reach a wanted efficiency on the total

parameter estimation. The same process will be applied to each sampling design selected by the user to be assessed. Several values of n are so tested. For each n value, a large number (>1 000) of simulations of samples arrangements are computed following the idea of bootstrapping technique (Fontaine et al. 2008). This permits to remove random sampling fluctuations. Then for each combination $n \times j$ (one simulation) we calculate the estimate following the formula $\hat{T}_{n,j}$ $\hat{V}(\hat{T}_{n,j})$ and (for the case where we want to estimate a total) and use the mean of the 1 000 simulations $\overline{\hat{T}_n}$ and $\overline{\hat{V}(\hat{T}_n)}$. The process is described step by step in the following (Fig.12) :

1) Simulate 1 000 samples j for both increasing sampling efforts n ;
2) Values of samples j are values of corresponding statistical unit $_j$ on previously reconstructed population $\hat{Y}$;
3) For all simulations j calculate $\overline{\hat{T}_{n,j}}$ and $\overline{\hat{V}(\hat{T}_{n,j})}$;
4) Calculate mean of the 1 000 simulations $\overline{\hat{T}_n}$ and $\overline{\hat{V}(\hat{T}_n)}$.

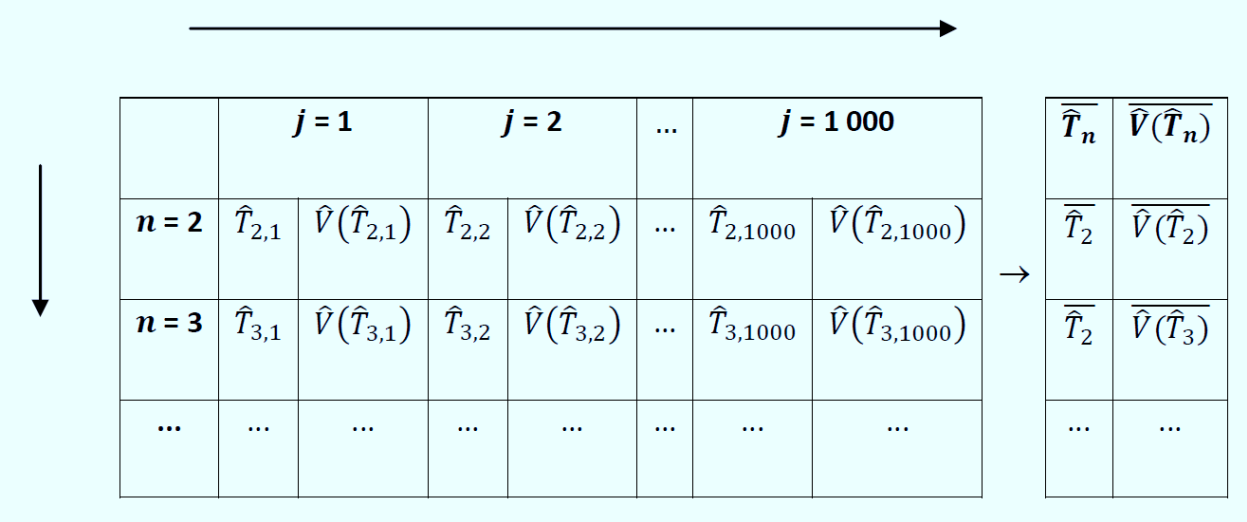


FIG. 12 : Simulation process

These steps have to be done for all assessed sampling designs.

Define optimal design and optimal sample size

For the previously assessed sampling designs, we have now $\overline{\hat{T}_n}$ and $\overline{\hat{V}(\hat{T}_n)}$ depending on sample size. The decision process to choose n_{opt} for each sampling design is based on an acceptable level of accuracy on results estimates. So the user needs to set this level called L and calculate margin of error with this level at both sampling size. To do so, we :

1) Set the level L of accuracy to be reached on estimates ;
2) Calculate margin of error ME_n of size L for all sampling size

$$ME_n = \frac{2t_\alpha \sqrt{\frac{\overline{\hat{V}(\hat{T}_n)}}{n}}}{\overline{\hat{T}_n}} \times 100$$

3) When margin of error ME_n is under the level fixed L in 1), optimal sample size n_{opt} is reached for this sampling design at this level of accuracy. Among the sampling assessed designs, the one that needs fewer samples than other to reach a same margin of error in total estimation is chosen as the optimal sampling design.

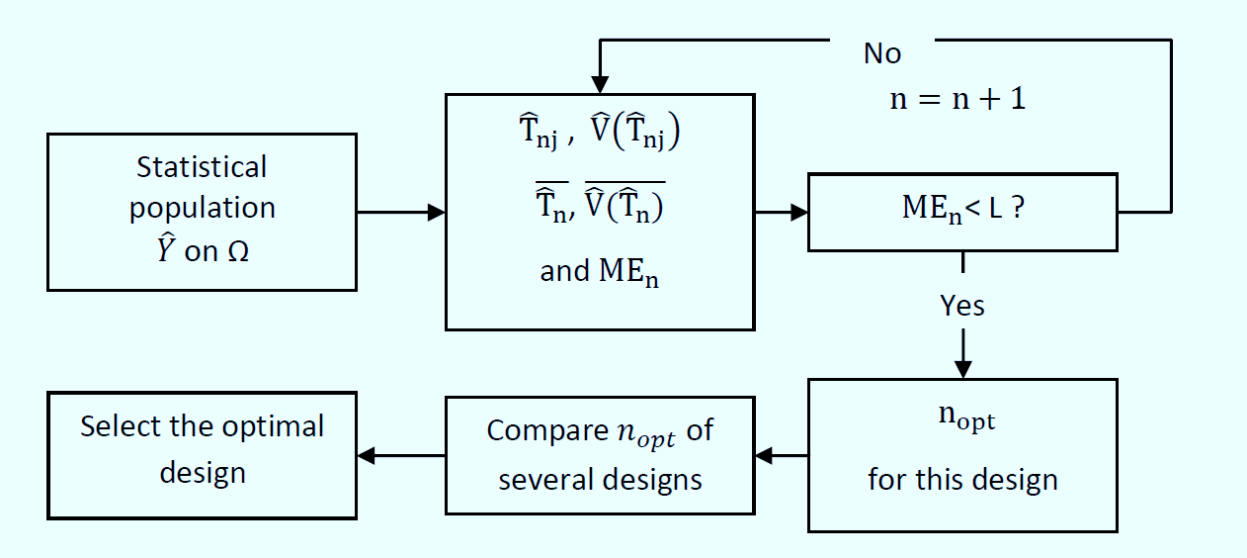


FIG. 13 : Process to assess optimal sampling design and sample size

Example : optimal sampling design and optimal sample size determination

Meuse dataset is still used here. We assessed simple random sampling (SRS) and systematic sampling (SSS) on the two Y distributions previously created. We set the levels of accuracy to reach on total estimator result on 10%. Sytematic sampling and simple random sampling designs needs respectively 23 and 18 samples to achieve 10% of accuracy on results on the population created by ordinary kriging, and 26 and 25 for the population built by linear model. Systematic sampling always need less samples than simple random sampling to achieve a same accuracy on total zinc estimation, SSS is so the optimal sampling design for this resource. The optimal sample size depends for others level of accuracy we want to reach on zinc total estimate can be easily calculated. Next and ultimate step is to estimate the price of the surveys with theses number of samples, but this step is very dependent of the study.

Conclusions

Where to sample and how many samples have to be done is often over-looked in surveys, resulting into non precise and non reproducible datasets. Results of inadequate surveys can

be misleading and hazardous not only because they fail to answer to the study problem but also because they can create the illusion that something useful were done (Peterman 1990). Following these lacks of robust survey designs and sampling strategies (also highlighted by Hayward et al. (2015)), we construct a robust and reproducible process permitting sampling strategies' results to be non contentious. We assess our sequential framework on the monitoring of manila clam resource on Arcachon bay (south west of France). By selecting a more performing sampling design for this survey, we could decrease the number of samples an save 30% of the price of the overall survey (Kermorvant et al. 2017 , 2019). We believe that our general process will be useful for scientists and managers. We developed it keeping in mind that it must be adaptable to any survey and its special features. It allows choosing the more efficient sampling design, leading in reducing sampling size and/or increase accuracy of results. This process can also be employed when attempting to develop a new monitoring, thus by selecting the best sampling design and the best sampling size from the beginning.

This method can have certain limitations. The reconstructed distribution of $\hat{Y}$ on Ω needs to be very representative of the targeted real population. If this requirement is not respected, the simulations will under or over estimate the needed sample size and can fail to select the best sampling design.

The process we presented here differs from already published ones because it allows taking into account prior knowledge of the population. One of its strength is that it is based on a simulation study (Zurell et al. 2010) and so all possible strategies can be assessed, without excessive expenditures. Framework result's is an optimal sample size by assessed sampling design for a desired accuracy in the results. But, the practitioner can define more than one a priori accuracies to be reached in the estimate and compare sampling designs and sample sizes needed to achieve them. As sample size and total survey cost are closely related, calculating the cost-effectiveness of several combinations is possible and the most appropriate one can be selected, before going on field. Having the possibilities to assess a large amount of sampling designs, choosing the better one and finding the optimal sample size are very relevant for studies where funds are often a limiting factor. In this sense, we choose to use semi-virtual populations. They reproduce as close as possible the distribution of the variable of interest and so they can be used to compare sampling designs (Albert et al. 2010), without being forced to assess all of them on the field. Virtual ecology is a powerful tool, that allows a quality assessment of sampling designs (Zurell et al. 2010).

Résultats issus de ce chapitre

- Nous avons développé une méthode permettant de définir un nombre d'unités statistiques à échantillonner et le protocole d'échantillonnage le plus adapté à utiliser pour un suivi.
- Cette méthode est basée sur de l'écologie virtuelle, un processus très puissant. De plus, elle combine les deux grandes méthodes d'échantillonnage que sont le *model-based* et le *design-based*.
- Le rapport coût (en nombre d'unités statistiques) / efficacité (en précision d'estimation) de chaque procédure testée par simulations est calculé pour ne garder que le plus intéressant.
- La méthode peut être utilisée pour optimiser un suivi déjà en place.
- Elle est adaptable à la pluspart les cas d'études car très flexible.

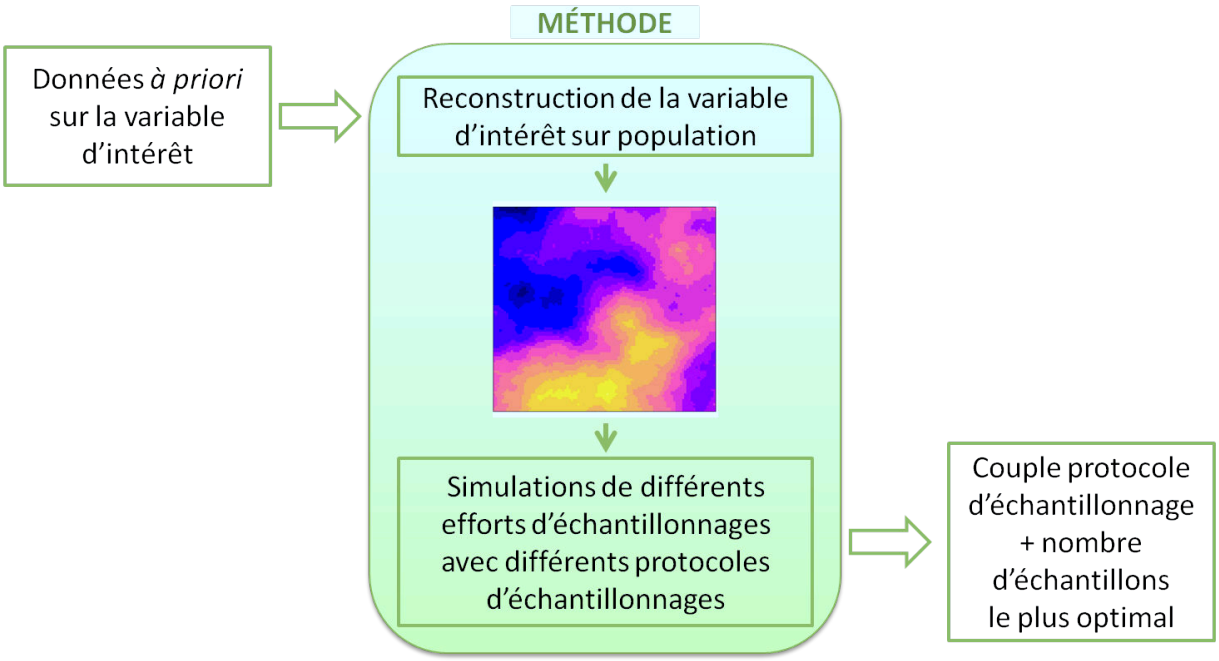


FIG. 14 : Conclusions du CHAPITRE II

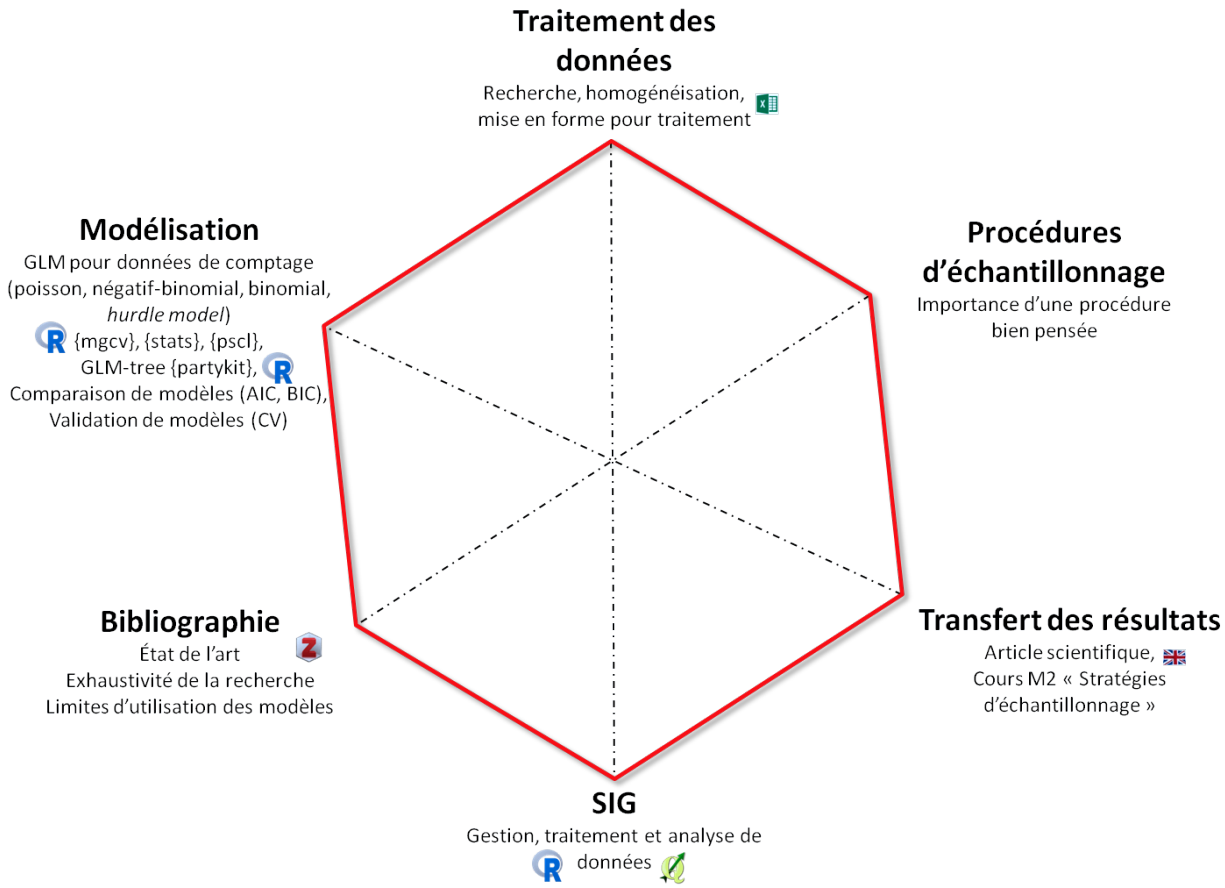
CHAPITRE III : Mise en place d'un suivi efficace sans données *a priori*

Synopsis :

La procédure séquentielle, coeur de cette thèse ayant été définie dans le chapitre précédent, les trois chapitres suivants en sont des illustrations. La première d'entre elles est l'utilisation de la procédure pour mettre en place un suivi efficace, lorsqu'aucune donnée n'est disponible sur site.

Mettre en place une procédure d'échantillonnage peut s'avérer complexe, surtout lorsqu'il n'existe pas de données *a priori* sur la population d'intérêt. Plus que non controversables, les résultats de suivis environnementaux doivent être assez précis pour permettre de répondre à la problématique initiale. La controverse peut être évitée en utilisant un protocole d'échantillonnage probabiliste ; la précision peut être augmentée en sélectionnant plus d'unités statistiques dans l'échantillon. L'objectif de cette étude était de mettre en place un suivi efficace pour le moustique tigre *Aedes albopictus* sur une aire nouvellement colonisée dans le sud-ouest de la France, où aucune données ne sont pour le moement disponibles. En suivant une méthode que nous avons publiée précédemment, nous avons utilisé la technique de l'écologie virtuelle pour sélectionner le protocole d'échantillonnage le plus efficace ainsi qu'un nombre d'unités à échantillonner pour maximiser le rapport coût/qualité de ce suivi. Cette méthode nécessite de recréer une population virtuelle sur le site d'étude à partir d'une modélisation de données disponibles autre part. Les différentes procédures d'échantillonnage sont ensuite testées sur cette population statistique. Cela permet de proposer différentes procédures d'échantillonnage pour le suivi du moustique tigre dans ces villes ; en fonction de la problématique de recherche initiale.

Compétences développées/utilisées pour ce chapitre :



Valorisations de ce chapitre :

- **Publication :**

KERMORVANT Claire, D'AMICO Frank, L'AMBERT Grégory. Setting up an efficient survey to estimate population parameter without prior knowledge.

Abstract

Setting up a survey may be may be very complex, especially when no a priori informative data are available on the population of interest. More than not controversial, surveys results must be enough precise to allow answering the initial research question. Controversy can be exempted by using a probabilistic sampling design ; precision can be enhanced by selecting enough samples. This study objective was to set up an efficient survey of *Aedes albopictus* on a newly colonized French agglomeration were no data are available yet. Following a method we previously published, we used virtual ecology method to select the most cost-efficient sampling design with appropriate sample size. This method request to recreate targeted population based on data from elsewhere. Then sampling procedure simulations are compared on this virtual population. This allows us recommend different sampling procedures on this agglomeration, allowing answering the initial research question.

Introduction

Sampling in environment

Monitoring information is essential for policy makers and scientists. For example they can be the basis for restoration programs or for endangered species conservation. Performing an exhaustive survey of any phenomenon is almost impossible or may prove to be tricky (Chiarucci et al. 2003 ; MacKenzie 2006) because of time (Cox, Cox, et Ensor 1997) and/or money lack (Theobald et al. 2007 ; Jackson et al. 2008 ; Lazarina et al. 2014). The common practice for dealing with this problem is trying to infer the targeted phenomenon on the basis of samples from the original population (MacKenzie 2006). Sampling procedure can be divided into four different components (Vos, Meelis, et Ter Keurs 2000b) : monitoring objectives (questions and hypothesis), sampling strategy (sampling design, number of samples, type of samples…), data gathering and data handling (analysis and response to the original question). The sampling procedure has to be well-thinking a priori (Conn et al. 2016) to minimize in fine this sampling error (Sica 2006). In this way, a "good" survey procedure ensures that the sample is non-significantly biased (MacKenzie 2006), gives precise estimates of the original population and results cannot be an object of controversy.

Issues when sampling procedure is not used

Schedule data gathering is an important step of the sampling procedure when a design-based inference is wanted. This step is capital for the results to answer the primary problematic (Martin, Kitchens, et Hines 2007). But environmental programs suffer from lack of details of problematic definition and hypothesis formulation, adapted sampling design and so data quality (Legg et Nagy 2006). Poor method has numerous undesirable effects that can lead in the failure of a monitoring program (Legg et Nagy 2006). Several studies deplore surveys are planed after data gathering and research problem is also thinking a posterior (Roberts 1991 ; Nichols et Williams 2006 ; Goldsmith 2012). Issues with poor designs used in ecological studies often have led to significant controversy (Hayward et al. 2015). It also mean that it become difficult to answer initial research problem and not very useful for decision making (Vos, Meelis, et Ter Keurs 2000b). For example it has not been possible to evaluate the effectiveness of US$15 billion projects of rivers restoration project all around US (37,099 projects) because of poor experimental design and lack of rigorous monitoring (less than 10 % of them indicate a form of assessment or monitoring of project efficiency) (Bernhardt et al. 2005). Roberts (1991) and Nichols et Williams (2006) deplore too many monitoring are "planned backward on the collect now (data), think-later (of a useful question) principle". More than be ineffective because of their inability to detect ecologically significant changes (Legg et Nagy 2006), inadequate monitoring can create the illusion that something useful has be done(Peterman 1990). A forum (Hayward et al. 2015), wrote after conflicting results, was published in high-quality scientific journal. It emphases robust methods and appropriate experimental designs must be developed and used by practitioners, avoiding controversy in studies results.

Prior knowledge for setting up a survey

When setting up a survey, it may be very relevant to have the opportunity to assess different sampling procedure to choose the most efficient. Prior knowledge of the studied area and the population can dramatically reduce the uncertainty in the sampling estimate (J.-F. Wang, Stein, et al. 2012). Expert knowledge is necessary and can help to better know the variable of interest before sampling for the first time. This knowledge is very relevant to provide initial idea of quantification and spatial (and/or temporal) delimitations of the survey area (and/or duration). But it cannot be used as a sole element to implement a new survey. Even if the expert involves are usually experienced and well-respected researchers, informal techniques that are implicit and unstructured are prone to cognitive bias (Webb et al. 2015). Virtual ecology can allow a quality assessment of sampling procedures. It's a rigorous research tool that is powerful and intuitive (Zurell et al. 2010). Kermorvant et al develop a method providing cost-efficient sampling design and sample size for survey setting. The main idea is to reconstruct a virtual variable of interest $\hat{Y}$ as close of possible of real Y on Ω statistical population and assess several sampling procedure on it. The sampling design needing fewer samples than the other to achieve precise estimate of variable parameter $\hat{\Theta}$ is selected with this associated sample size. But an issue arise when wanted to reconstruct $\hat{Y}$ and no data of Y are available on Ω.

A statistical model can be built and transferred if some data exist and are available in a similar environment. Statistical models such as GLM allow linking the targeted variable Y to exogenous variables X known on all statistical units ω of the statistical population Ω. Once this model created, its goodness of fit must be checked. When the model was developed for predictions of $\hat{Y}$ to all statistical units of the population Ω, the most critical metric to evaluate is how well the model predicts the target samples. This can be done, for example, with cross-validation. But these methods does not assess model generality or transferability to another statistical population, which can lead to overestimates of performances when predicting in other locations (Wenger et Olden 2012). Transfer of ecological models across geographic locations have shown limited results on literature (Dobrowski et al. 2011 ; Eskildsen et al. 2013 ; Peñalver-Alcázar et al. 2016 ; Wogan 2016). But scientific community is in front of a lack of methods to quantify transferability and this is, for now, a gap of highest priority (Yates et al. 2018). The transfer of model into a new population can provide a good base against which data can later be benchmarked once available (Yates et al. 2018). Once the model transferred parameter Θ' of the variable Y can be estimated on the new statistical population Ω'. Requirement to transfer the model built on Ω to Ω' is same exogenous variables X must be known on all statistical units of ω of Ω and all statistical units of ω' of Ω'.

In design-based sampling theory, probabilistic sampling designs must be used to guaranty data integrity. Some new sampling designs like generalized tessellation sampling design (GRTS) (Stevens Jr et Olsen 2004) are able to spread sampling units in one dimension (spatial) or in several dimensions (Robertson et al. 2013 , 2018 ; Brown, Robertson, et McDonald 2015). Theses dimensions could include auxiliary information such as ecological threats, time intervals, species population structure or environmental data (Brown, Robertson, et McDonald 2015). Algorithms will change the inclusion probability of samples by selecting them with an unequal probability. Unequal probability sampling can be more efficient than equal proba-

bility sampling if there is a positive correlation between the inclusion probabilities and the response values (Robertson et al. 2013). Local pivotal method (Anton Grafström, Lundström, et Schelin 2012), is for now the only spatially balanced sampling design (SBS) able to spread samples in several dimension available on R software (package {BalancedSampling}, (Grafström et Lisic 2016)). The Swedish national forest inventory, for example, has implemented LPM with five auxiliary variables (Grafström et al. 2017). When a survey is started to study the parameter Θ of the variable Y on population Ω ,distribution of Y variable on population is of course not known. In this typical case LPM requires are 1/ auxiliary variables X must be known on all statistical units ω of the statistical population Ω and 2/ there are significant correlations between auxiliary variables X and the inventoried variable Y. More the correlations between X and Y are significant, more the used X variables reproduce reliably Y distribution. But, by using previous data and statistical modelling, Y distribution on Ω' can be predicted dependently on X variables. Thus, in this second case Y distribution is known on Ω' statistical population. LPM can be used here to balance samples on the reconstructed targeted variable Y.

Study objectives

The main objective of this study is to set up a survey enough efficient to detect significant changes on *Aedes albopictus* presence probability in an agglomeration where it is not already well established (Bayonne-Anglet-Biarritz) by using data from another French region (Mediterranean cities) infected since several years. The three main steps will be 1/ to use data from cities where tiger mosquito's populations are now well-established and model presence probability in these cities, 2/ transfer this model to Bayonne-Anglet-Biarritz (BAB) agglomeration and 3/ find the minimal number of samples and the optimal sampling design to achieve 5% of precision on presence probability estimates.

Materials and methods

Aedes Albopictus ecology

Tiger mosquito, can be vector of several arboviruses including dengue (DENV) (Simmons et al. 2012), yellow fever (Jentes et al. 2011), chikungunya (CHIKV) (Jentes et al. 2011), but also zika (ZIKV) (Grard et al. 2014). Several studies already demonstrate that *Aedes albopictus* (= *Stegomyia albopicta*) (Reinert, Harbach, et Kitching 2009) distribution at large scale is mainly explained by bioclimatic variables. European centre for disease control (Disease Prevention and Control 2013) show this mosquito presence depends on precipitations and temperature. On a more reduced scale, tiger mosquito distribution would be more dependent on land cover (Rochlin et al. 2013). More precisely, urban, peri-urban or agricultural and forest landscapes have a positive effect on the probability of presence of tiger mosquito (Vanwambeke et al. 2007 ; Roche et al. 2015). Mosquito's larvae grow in human made containers as rain gutter, discarded tires, pots of flower… (Roche et al. 2015). When it's available human blood is preferred to other animal's species blood (Delatte et al. 2010). This makes urban and suburban zones an enabling environment for mosquito life cycle and preferential living space (Buckner et al. 2011 ; Beilhe et al. 2012 ; Rochlin et al. 2013 ; Roche et al. 2015 ; Samson et al. 2015).

Dataset from Ω population

In some cities of south east of France, tiger mosquito invasion is studied since 2008. Data are collected along a detection monitoring network of oviposition traps ; designed to detect as soon as possible invasions by this mosquito species. In this sense, sites are not randomly situated in space neither randomly picks up. Design-based inference is so not possible. We will use model-based inference. Oviposition traps are plastic buckets (Fay et Eliason 1966) with a square of polystyrene (layong support) imitating *Aedes Albopictus* behaviour of laying into human made containers (Boubidi 2016). Theses artificial containers are making more attractive by fulfilling them with a plant infusion (Reiter et Colon 1991). Ovitraps are the passive surveillance tool for detecting presence of gravid female mosquitoes used for longer time because of their easiness of conception and low cost of production. They are the most widely used routine approach for large-scale monitoring (Manica et al. 2017). More than 7000 results of checked trap at precise time and place are available. Abundance is recorded and so presence-absence can be easily calculated. For France land cover classification we used the OCS CESbio (Inglada et al. 2017). This classification is available for 10m squares and so adapted to a fine scale usage. Total precipitations, minimal, maximal and mean temperature by months were freely extracted from French meteorological company "Météo France"

Model Ψ and predict on Ω'

A GLM model will be created with these data to link the presence probability of gravid mosquitoes female Ψ to exogenous variables. Then, this model will be used on another population to predict presence probability of mosquitoes there.
We built the model from Mediterranean Ω statistical population to transfer it on BAB agglomeration. This second population is so called Ω'. Correspondingly, BAB statistical units are ω, the interesting variable is Y' (presence-absence) where we will measure the Ψ parameter of presence probability.

Land cover and meteorological data were extracted from the same database than Mediterranean one. 2018 data were used to predict month by month presence probabilities on BAB population.

Set up an efficient survey on Ω'

Once Ψ predicted on the new statistical population, several sampling procedures can be assessed by simulation. We use framework developed by Kermorvant et al. (submitted) to assess several sampling designs and choose the optimal one.

Chosen sampling designs to be compared are simple random sampling (SRS), simple systematic sampling (SSS), generalized random tessellation sampling (GRTS) and local pivotal method (LPM). SRS is the most easy to understand sampling design ; statistical units are chosen randomly. For SSS sampling design, only the first sampling unit is chosen randomly ; the other are placed equidistantly to cover the whole statistical population. SSS draws very well balanced sample but there is a huge drawback with it. If the variable is kind of cyclic SSS will not detect this cycle and results could be biased. GRTS is the most used spatially balanced sampling design (SBS) for now. It allows selecting well-balanced samples within

the study area but keeping a random algorithm. LPM is also able to draw spatially balanced samples, but it has the ability to balance the sample in more than just a spatial dimension. Theses dimensions could include auxiliary information such as ecological threats, time intervals, species population structure or environmental data (Brown, Robertson, et McDonald 2015).

Results

GLM model for presence-absence

We know from literature that tiger mosquito's dispersion in cities is almost entirely ruled by land cover and year period. We will reconstruct this model adding climatic data.

Data visualisation

Occurrence of presence and absence of *Aedes albopictus* eggs in traps is depicted in fig. 15. Number of picked up traps are very unbalanced between months and land cover types. Fall months display more occupied traps while winter ones display more unoccupied traps.

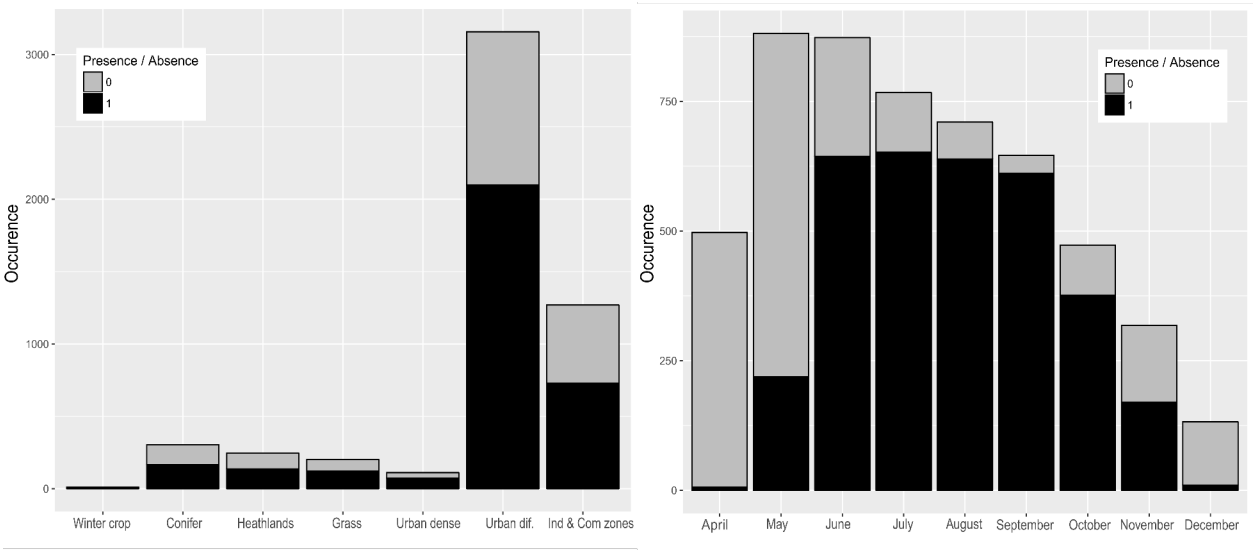


FIG. 15 : Data visualisation. Occurrence of traps with and without *Aedes albopictus* eggs depending on land cover and months

Modelling presence-absence

Presence-absence data allows predicting presence probability by using a binomial GLM. Once the model is created, a stepwise process was used to simplify it and keep only explicative variables. The remaining model is :

$$\Psi = \beta_1 \times mois + \beta_2 \times OCS + \beta_3 \times Prec + \beta_4 \times Tmin + \beta_5 \times Tmean + \beta_6 \times Tmax + \epsilon$$

Coefficients:	Estimates	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.0414094	1.0450614	-7.695	< 2e-16	***
May	2.9812379	0.4283355	6.96	< 2e-16	***
June	4.7350451	0.4655807	10.17	< 2e-16	***
July	5.2491292	0.4979465	10.542	< 2e-16	***
August	5.91609	0.4963592	11.919	< 2e-16	***
September	6.8447726	0.4790254	14.289	< 2e-16	***
October	5.7947301	0.4350674	13.319	< 2e-16	***
November	4.9289096	0.4325945	11.394	< 2e-16	***
December	2.5994522	0.541874	4.797	0.00000161	***
Conifer	1.7374063	0.8529901	2.037	0.041666	*
Heathlands	1.6671878	0.8549798	1.95	0.051179	.
Grass	1.7820175	0.8593385	2.074	0.038106	*
Urban dense	2.4153187	0.8826535	2.736	0.006211	**
Urban diffuse	2.3791546	0.8408169	2.83	0.004661	**
Ind. & Com. Zones	1.5412703	0.8418012	1.831	0.067113	.
Monthly precip	-0.0025963	0.0006757	-3.843	0.000122	***
Monthly minimal T°C	-0.0097319	0.0116766	-0.833	0.40459	
Monthly mean T°C	0.0602824	0.0178126	3.384	0.000714	***
Monthly maximal T°C	0.0347296	0.0183281	1.895	0.058108	.

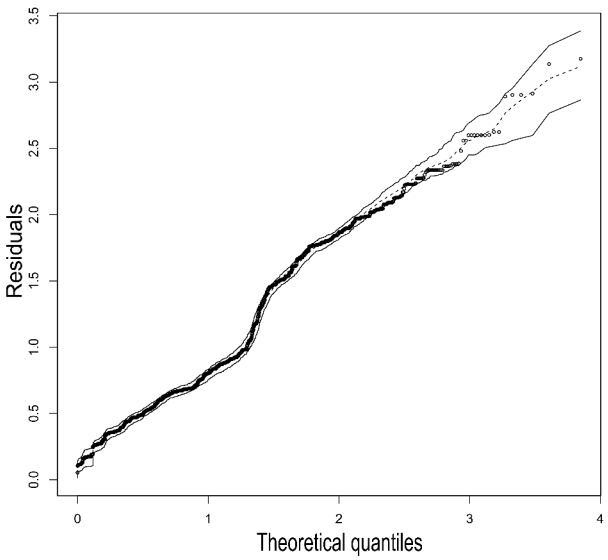


FIG. 16 : Summary and residual plot of built model

Residuals dispersion plot (Fig.16) shows the model well fit data. Furthermore, cross validation result is relatively low (leave-one-out-cross-validation's average mean squared error= 0.129) so this model is powerful to predict mosquitoes presence probabilities on Mediterranean cities.

Set up an efficient survey on Ω'

The population from we make the model is quite different from the population were we want to predict Ψ with the model in term of land cover (Fig. 17). Ω' is tree-quarter represented by diffuse urban land cover while in Ω urban land, conifer forest, grasslands and heartlands are both greatly represented. Speaking about mean minimal and mean maximal temperatures, in 2018 they were not very different between the two populations.

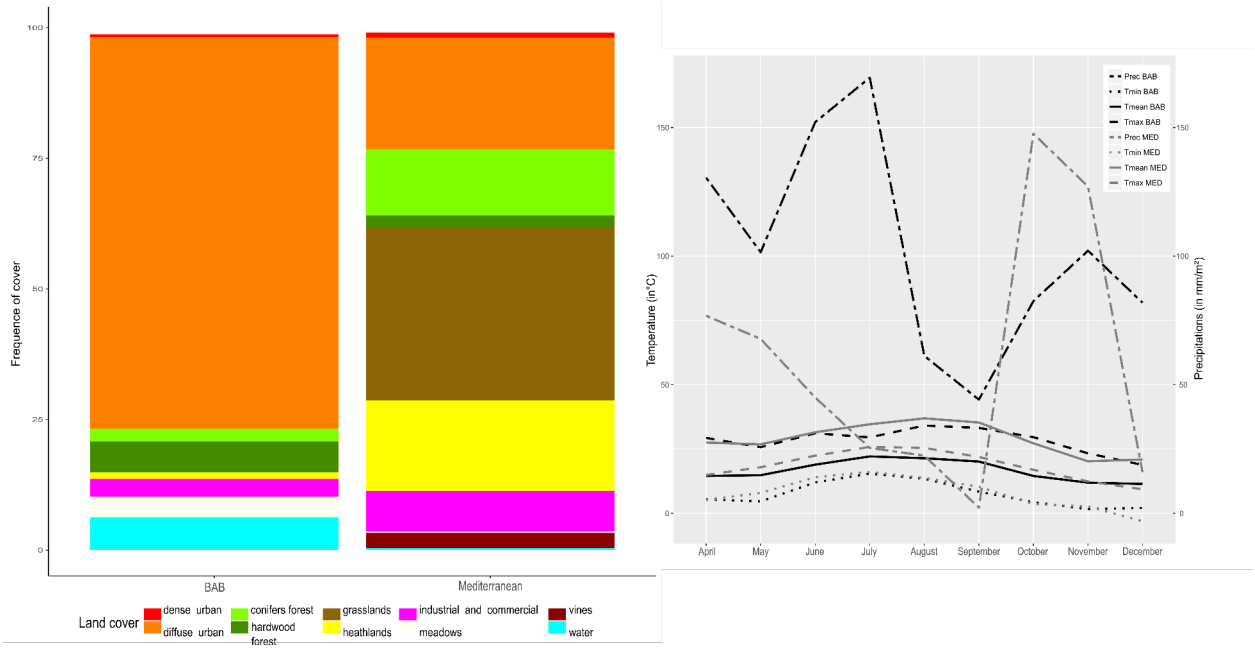


FIG. 17 : Differences observed between Mediterranean population (Ω) and BAB population (Ω') in term of land cover temperatures and precipitations

A map of tiger mosquitoes predicted presence probability Ψ' on BAB agglomeration was constructed for each month. Fig. 17 shows an example for June.

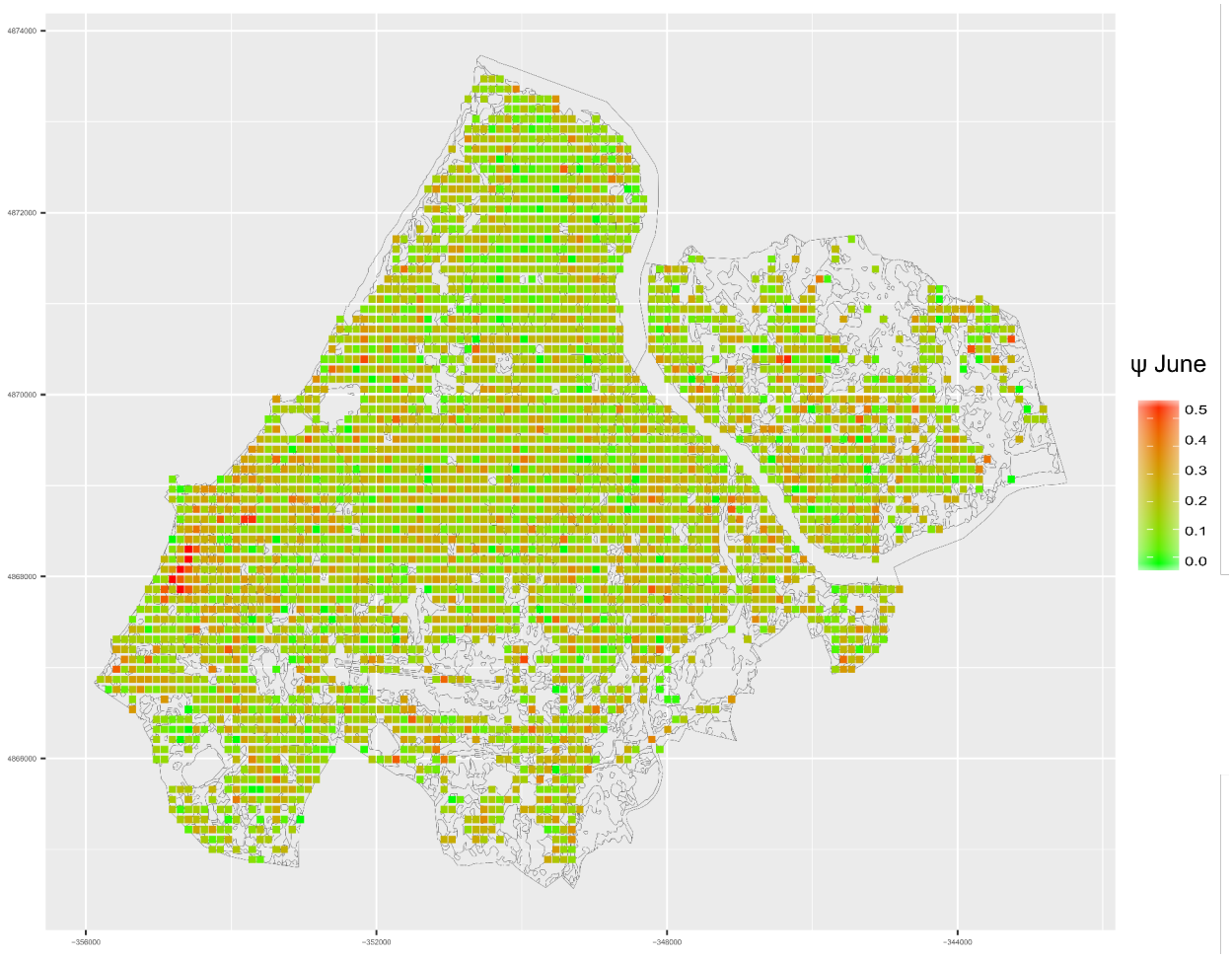


FIG. 18 : predicted presence probability map on BAB agglomeration (example for June)

Both of reconstructed populations serve as basis for sampling procedures simulations. Fig. 19 depict how many samples are necessary to be gathered by month and by assessed sampling design to achieve 5% of precision on presence probability estimates.

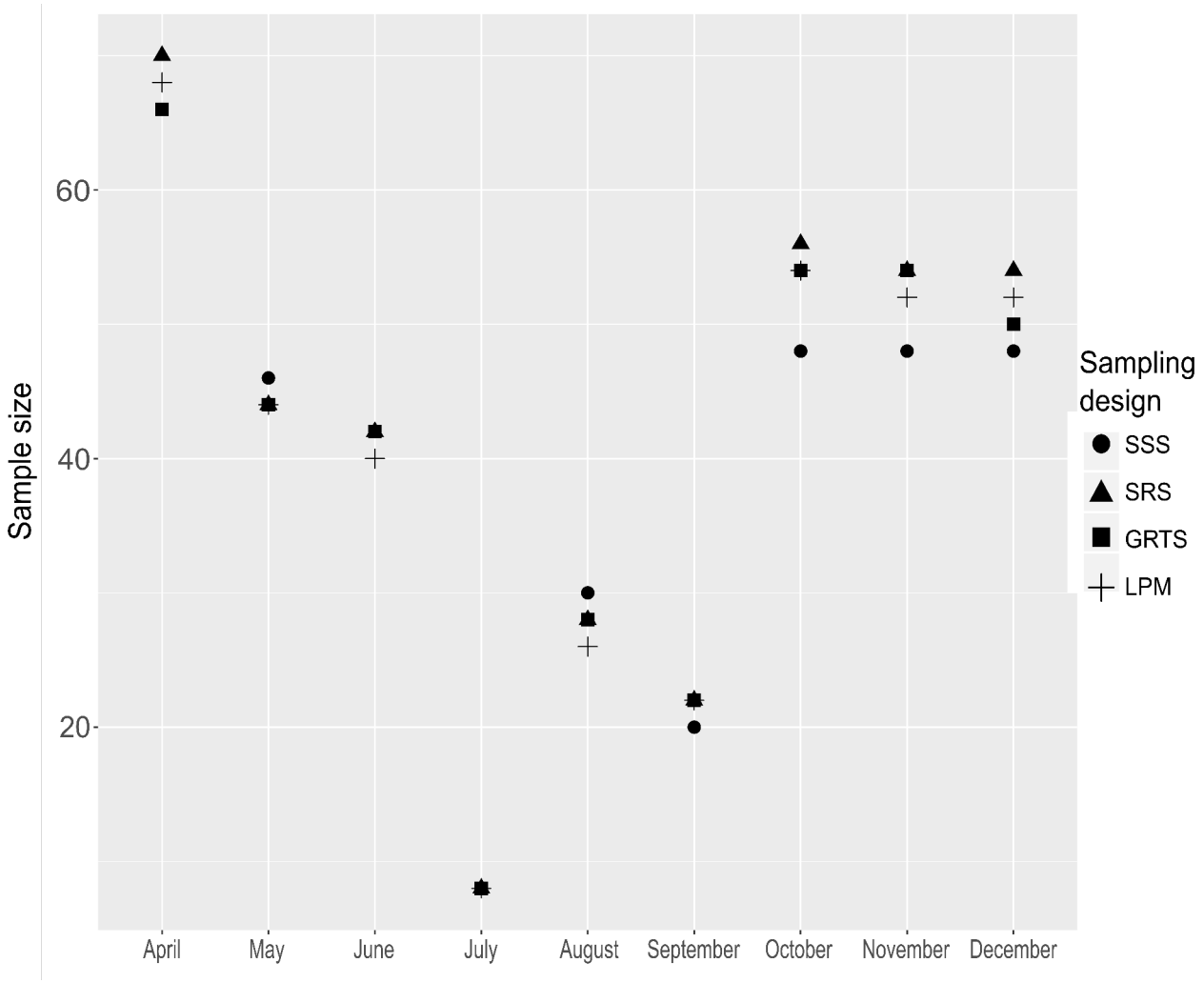


FIG. 19 : Number of samples needed with SSS, SRS, GRTS and LPM sampling designs to achieve 5% of precision on presence probability estimates

There are not big differences between sampling designs. Simple random sampling seems to perform less than the three others, needing little more samples to achieve 5% precision on presence probabilities. Months have an influence on number of samples needed to achieve the wanted precision. In summer time very few samples are needed (less than 10) but winter/fall require approximately 70 samples.

Discussion

In this paper, we demonstrated that we can provide an efficient sampling procedure when wanting to set a new survey on an unknown population. This by setting a number of samples to draw with a specific sampling design to achieve relevant precision on population estimates. Precision is chosen by the practitioner and depends on the initial research problem.

Results highlight that in BAB mosquito's population, the choice of sampling design is not very relevant for achieving required precision on presence probabilities estimates. Spatially balanced sampling (SBS) designs perform little better than simple random sampling, but we were expecting a higher difference. These poor results of SBS can be explained by an almost random distribution across the studied population. SBS use is very relevant when the targeted variable has a strong spatial trend (Stevens et Olsen 2004 ; Barabesi et Franceschi 2011 ; Anton Grafström et Lundström 2013 ; Robertson et al. 2013 ; Benedetti, Piersimoni, et Postiglione 2017) , because their design-based estimators take into account spatial heterogeneity (J.-F. Wang, Stein, et al. 2012) and spatial auto-correlation (Haining 2003).

The high monthly variation in number of statistical units to be sampled to achieve a 5% precision can be an issue. We cannot provide a survey which requires the field technician to change Ovitraps position and number each month. Our recommendations, if the initial research question is to detect a significant change between months on presence probability, are to keep the higher number of sample between months (April in this example), geographically place them with a SBS and gather them with a frequency of at least one month. It will certify that at least 5% of precision on presence probability will be achieved each month. If the initial research problem is to build a model of presence probability depending on land cover type, we recommend at least 30 traps on both land cover type gathered at least each month. Thus to guaranty statistical model assumptions will be met. Concerning sampling design, a SBS is also advised.

Ovitraps method for sampling mosquitoes is easy of conception and has a low cost of production. But the main issue with it is they cannot provide an accurate estimation of gravid female mosquitoes (Eiras et al. 2018). This is because a female can lay different numbers of eggs in an Ovitrap (Abreu et al. 2015). Another trouble with eggs sampling is the laboratory stage for count and species determination. It could take one or two weeks and this delay the population information (Resende et al. 2013). An improvement for Ovitrap can be to add a sticky surface to collect gravid female (Eiras et al. 2018). This will give a better estimation of adult population.

There are several differences between the two studied regions. The first comes from the fact that tiger mosquito presence is documented on BAB's region since 2015 only. Even if *Aedes albopictus* breads onsite the population is not yet well established, conversely to Mediterranean population. From Mediterranean dataset, we modelled a population occupying its realised ecological niche. But in BAB, mosquito is still spreading and so does not occupy yet its whole realised ecological niche. Second difference is exogenous variables are not exactly the same : land cover and climate are different. These differences can have an effect on mosquito's population ecology, dispersion and so colonisation of the urban area. They can have a strong effect on the model transferability, but we are not able to assess it yet because of methods lack (Yates et al. 2018). A previous study on *Aedes albopictus* distribution models worldwide transfer (Medley 2010) demonstrates a niche divergence between continents. This study also revealed that the transferred model between continents failed to predict all data, but predict better than expected by chance. This led us to warn about the necessity of updating the model on BAB once data will be available.

This must be checked on future by implementing the model dataset with data from BAB.

Data from Mediterranean cities where not collected in the aim to model mosquitoes presence probabilities with land cover and months. This leads into very unbalanced sample, winter months are not sampled at all several land cover type are sampled a lot of time (e.g. urban diffuse) other are almost not (e.g. hardwood forest). With few values, theses later land cover types were not included on the Mediterranean cities' model and so not predicted on Atlantic cities (which makes some gap in the predicted map).

Conclusion

We succeede on setting up an efficient survey when no data are available onsite. We follow guideline provided by Kermorvant et al. (submitted) to set an efficient *Aedes albopictus* survey procedure while no data were available. As models test of transferability is not yet availiable, we suggest to implement futures data in the créated model to update statistical population. Recommended sampling procedures depend on the future research question in this area.

Résultats issus de ce chapitre :

- Nous avons réussi à modéliser la probabilité de présence du moustique tigre sur les villes méditerranéennes françaises, puis avons transféré ce modèle à des villes de la façade atlantique (BAB).
- La méthode développée durant cette thèse a ainsi pu être appliquée pour mettre en place un suivi efficace sur la population de moustiques tigre du BAB, où très peu de données sont disponibles.
- Le nombre d'unités statistiques à échantillonner pour atteindre 5% de précision dans les résultats d'estimation de probabilité de présence est plus influencé par le mois de l'année que par le protocole d'échantillonnage utilisé ; néanmoins les SBS (protocoles spatialement équilibrés) restent plus performants que l'aléatoire simple.
- La transférabilité géographique d'un modèle statistique ne peut, à ce jour, pas être évaluée par manque de méthodes. Mais les résultats produisent une base pour l'utilisation de méthodes comme l'écologie virtuelle ; des données pourront être ajoutées une fois qu'elles seront disponibles pour implémenter ce modèle.

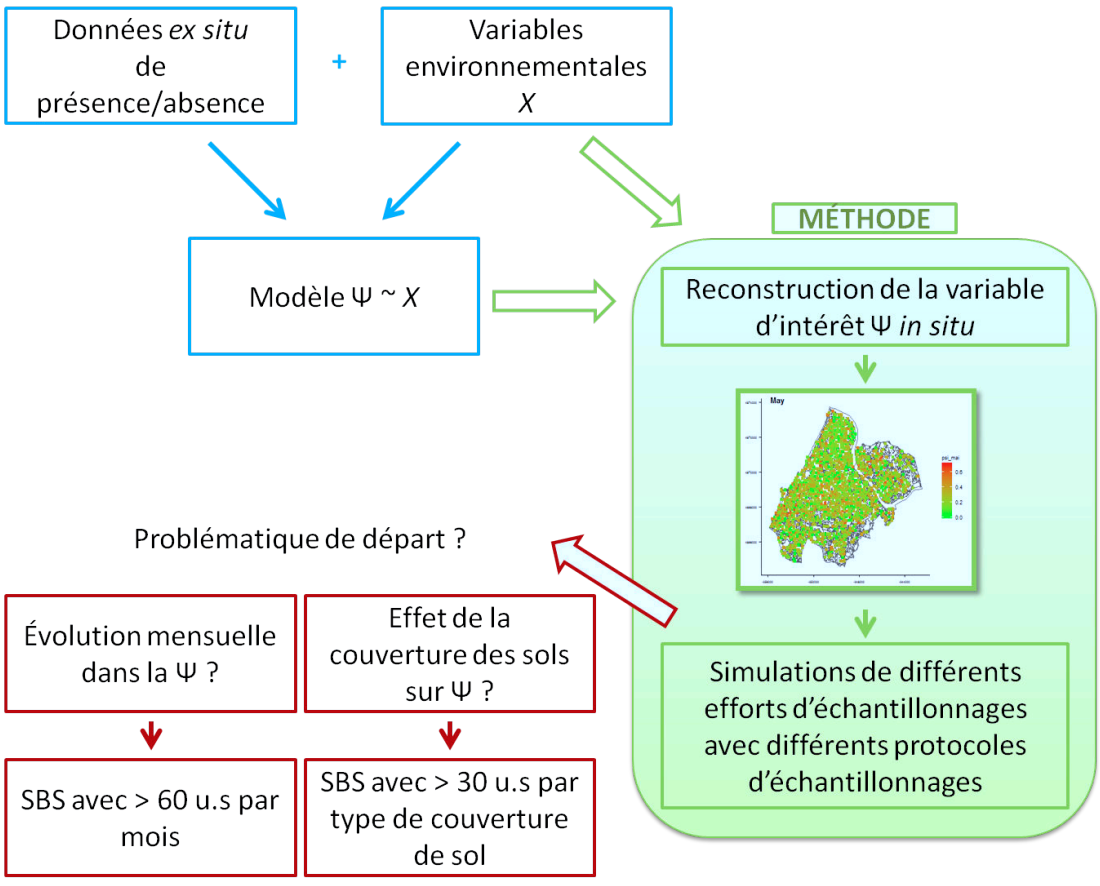


FIG. 20 : Conclusions du CHAPITRE III

CHAPITRE IV : Optimisation d'un suivi à partir d'une seule saison de données disponibles

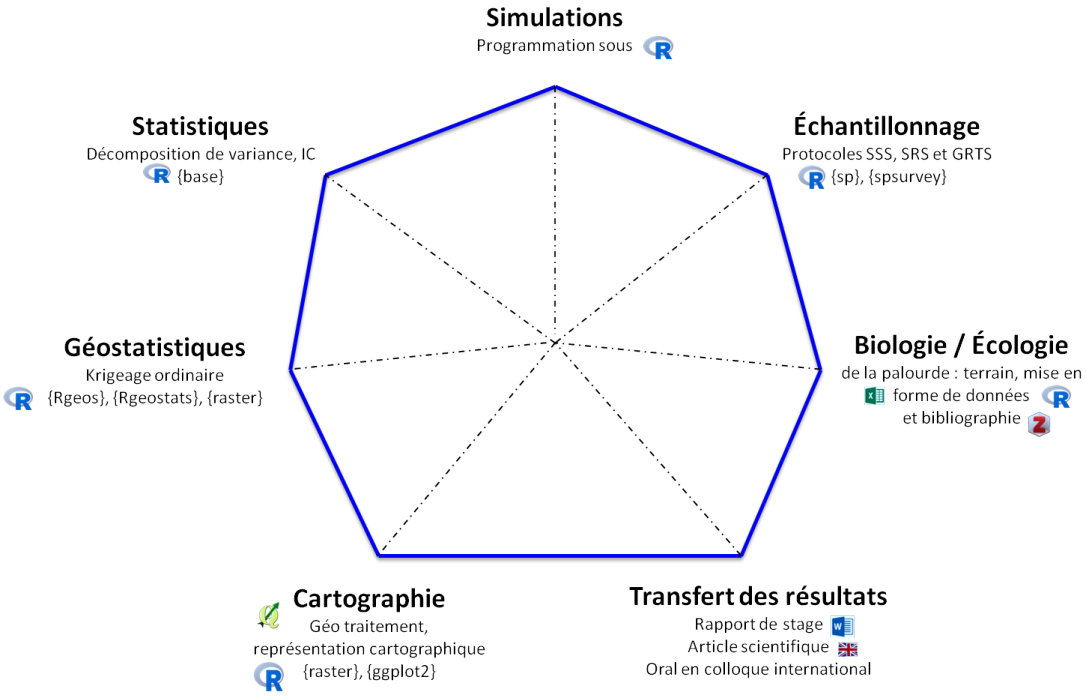
Synopsis :

Le chapitre qui suit est un exemple d'application de la procédure séquentielle présentée au chapitre 2 dans le cadre de l'optimisation d'un suivi avec une seule saison de données disponibles.

Les bivalves représentent une composante importante des écosystèmes marins benthiques et d'eaux douce à travers le monde. La palourde (*Venerupis philippinarum*) est l'un des bivalves les plus exploités pour la consommation humaine. Dans le bassin d'Arcachon (sud-ouest de la France), les pêcheurs professionnels, en collaboration avec les scientifiques, ont développé et mis en place un suivi sur le long terme pour estimer le stock de palourdes. Les résultats de ces suivis sont utilisés pour assister une stratégie de gestion durable de la ressource exploitée. Le suivi est actuellement basé sur un protocole en aléatoire stratifié (StRS) standard. Jusqu'ici, il a été effectué tous les deux ans depuis 2006. Le coût de chaque campagne de suivi, financées à hauteur de ~20% par les pêcheurs eux-mêmes, s'éléve à environ 50 000 €. Ce prix est trés élevé pour une ressource qui est, la plupart du temps, gérée à une échelle régionale. En 2016, une réduction des financements a engendré une annulation du suivi.

Des études récentes sur les protocoles d'échantillonnage se sont concentrées sur le développement de méthodes permettant une meilleure efficacité statistique (diminution de l'erreur d'échantillonnage) couplée à un effort d'échantillonnage réduit. Le protocole d'échantillonnage spatialement équilibré appelé "generalized random tessellation stratified" (GRTS) est l'une des ces méthodes. L'objectif de cet article est de comparer les performances du StRS communément utilisé avec celles du GRTS. Pour ce faire, nous avons recrée la population de palourdes dans l'ensemble du bassin d'Arcachon en extrapolant les données récoltées lors du suivi de 2012. Nous avons ensuite simulé plusieurs événements d'échantillonnage avec les deux protocoles à tester sur cette population semi-virtuelle. Finalement, nous avons défini les performances des deux protocoles d'échantillonnage en se basant sur trois niveaux de précision à atteindre dans les résultats (5%, 10% et 20% de précision) des deux estimateurs d'intérêts (biomasse et abondance). Nous recommandons l'utilisation du GRTS pour le suivi des palourdes dans le bassin d'Arcachon, puisque, pour atteindre un même niveau de précision, le GRTS a besoin de beaucoup moins d'échantillons que le StRS.

Compétences développées/utilisées pour ce chapitre :



Valorisations de ce chapitre :

- **Publication :**

KERMORVANT Claire, CAILL-MILLY Nathalie, D'AMICO Frank, BRU Noëlle, SANCHEZ Florence, LISSARDY Muriel, BROWN Jennifer. Optimization of a survey using spatially balanced sampling : a single-year application of clam monitoring in the Arcachon Bay (SW France). Aquatic Living Resources, 2017, vol. 30, p. 37.

- **Communication orale en conférence internationale :**

KERMORVANT, Claire, CAILL-MILLY Nathalie, BRU Noëlle, D'AMICO Frank, SANCHEZ Florence, LISSARDY Muriel, BROWN Jennifer. Clam monitoring : optimization of a recurring survey in the Arcachon Bay using spatially balanced sampling. Oral communication in International Symposium of Oceanography of the Bay of Biscay ISOBAY XV. June 2016. Bilbao, SPAIN

Abstract

Bivalves are important components of benthic marine and freshwater ecosystems throughout the world. One of the most exploited bivalves used for human consumption is manila clam (*Venerupis philippinarum*). In Arcachon Bay (SW France), commercial fishers and scientists have developed a monitoring survey to estimate clam stocks to assist in implementing a sustainable management strategy. The survey design that is currently used is based on standard stratified random sampling (StRS). The survey has been undertaken every 2 years since 2006. Each survey costs approximately €50 000, with funding provided by ~20% of the commercial fishers. The survey is quite expensive, given that this resource is managed mostly at a regional level. In 2016 for instance, the survey was not done because of a shortfall in funds to support it.

Recent studies on survey designs have focused on new developments that allow for higher statistical efficiency (lower sampling error) coupled with lower survey effort. Among these is the spatially balanced generalized random tessellation stratified (GRTS) design. The aim of this study is to compare the performance of the common StRS method with the GRTS design. To do this, we created a semi-virtual clam population by extrapolating the 2012 field survey results in the whole bay and simulated survey events with the two designs. We then assessed the two survey designs using three threshold precision levels (5%, 10% and 20% precision) for the two estimators of interest (biomass and abundance). We recommend the use of the GRTS design for clam surveys in Arcachon Bay. To achieve the same level of precision, GRTS requires less survey effort than StRS.

Introduction

Manila clam (Venerupis philippinarum) is one of the most exploited bivalves in the world and stocks are of concern in many locations. Because of the interest in the species, many studies have been carried out. For example, work has been done with the aim of assessing the geographic spread of these species (which can be invasive), such as in Poole Harbour in the UK (Jensen et al. 2004), San Francisco Bay, USA (Carlton et al. 1990), Venice Lagoon, Italia (Pranovi et al. 2006), Southern California, USA (Talley, Talley, et Blanco 2015) and Santander Bay, Spain (Bidegain et al. 2015). Others studies have been undertaken to focus on factors influencing mortality in Manila clam stocks (Park et Choi 2001 ; Paillard, Allam, et Oubella 2004) , to study hyperparasites (Le et al. 2015) and to report ingestion of microplastics (Davidson et Dudas 2016).

Despite the number and diversity of studies, there is no standardized design for bivalve sampling for population estimates. The studies cited above use different survey designs, each with different features such as randomized or stratified designs, quadrats or transect sample units, and once-off data collection surveys, or ongoing repeated field surveys etc. Interestingly, none of these studies has reported problems caused by the survey design, except for one (Davidson et Dudas 2016), which mentioned a potential limit in precision from having a small sample size. Many of these studies of Manila clam use a probability-based design, usually stratified random sampling (StRS) [James et Fairweather (1996) ; Pitel et al. (2004) ; J. Bald et al. (2005) ; Gray et al. (2014) ; C. A. Gray (2016a), C. A. Gray (2016b)). Others have used expert knowledge rather than a probability sample to locate sample sites. Sample designs not based on probability lead to a number of problems including lack of repeatability and the difficulty in estimating a valid measure of precision. This could lead to biased estimates (Albert et al. 2010). A fundamental reason for design-based sampling is that the a priori determination of inclusion probabilities allows for unbiased statistical inference (Särndal et al. 1978 ; Thompson 2012).

In France, Arcachon Bay represents, along with the Morbihan Gulf, the main Manila clam production area. Manila clam in Arcachon Bay have been studied since the 1990's to provide information on the species predators, mortality for species management (Robert et Deltreil 1990), and to identify suitable areas for Manila clam harvesting in the Bay (Robert, Trut, et Laborde 1993). Later, after abandoning the culture and installation of clams in the bay (Auby 1993), studies have been more oriented towards pathology (De Montaudouin et al. 2000) and, starting in 2003, stock surveys (Caill-Milly et al. 2003). Today, monitoring of the fishery stock is a co-management approach between commercial fishers and scientists. Surveys have been undertaken every 2 years since 2006 The surveys are conducted to assess selected indicators (densities, total number of clams expressed in number and mass, size structure, …for details, see (Caill-Milly, Duclercq, et Morandeau 2006 ; Caill-Milly et al. 2008 ; Sanchez, Caill-Milly, et De Casamajor Marie-Noelle 2012 ; Sanchez et al. 2014)) of the current stock and to detect any changes in these indicators over time. This information is the basis for adaptive management measures (e.g. issuing licences, defining protected areas, and identifying periods of no fishing). As there is no quota defined for this fishery, catch control can be made through the number of licences issued. If the indicators show a decline in the population, this number of licences is subsequently revised downward. Other options are to impose periods when fishing is banned,

and extending any protected areas. In this last example, accurate spatial data are mandatory for the choice of the protected areas, for example to select areas with sufficient densities, and containing a large proportion of adults.

A major concern with this last monitoring survey is that it is time-consuming (approximately 500 sample stations are visited on each survey) and costly, although the aim is to keep the biennial survey costs under a threshold of €50 000. Without the support (financial and in kind) of the commercial fishers, ongoing surveys are not assured. This indeed happened in 2016 and no surveys were undertaken. This lack of a survey in 2016 present new problems with reporting on the fisheries stock because there is no longer an unbroken time series of the stock status indicators to use in assessing population management measures. The Arcachon Bay monitoring survey is of particular importance because the clam population shows lower fitness compared with other French sites (De Montaudouin et al. 2015). Therefore, it is very timely and even necessary in this context to consider alternative monitoring survey designs that are less costly but do not compromise survey precision.

To monitor the state of the stock, the StRS method has been used for the Arcachon Bay's clam monitoring (Caill-Milly, Duclercq, et Morandeau 2006 ; Caill-Milly et al. 2008 ; Sanchez, Caill-Milly, et De Casamajor Marie-Noelle 2012 ; Sanchez et al. 2014) because it is viewed as the more powerful among classical sampling designs using strata. This survey design has already been used in the Morbihan Gulf (Berthou et al. 1997) and has provided information that has been accepted by both the science and fisheries communities. The main disadvantage of this design is that sometimes areas of the bay are not surveyed because there is no explicit spatial structure imposed on the sample locations within strata (Stevens et Olsen 2004 ; Christianson et Kaufman 2016).

Given these well-known caveats and weaknesses of StRS, new spatially balanced sampling designs have been developed for monitoring ecological resources(Stevens et Olsen 2004 ; Robertson et al. 2013 ; Brown, Robertson, et McDonald 2015). This presented an opportunity to review the use of the generalized random tessellation stratified (GRTS) spatially balanced design as an alternative to the current StRS used in the Arcachon Bay clam monitoring survey. Survey design starts with setting targets for total effort and desired precision of the survey estimates ; priority can be given to either maximize precision or minimize total effort (Guillera-Arroita, Ridout, et Morgan 2010). To formulate a survey design, the sampling practitioner must be aware of the trade-offs among objectives (Stehman et Overton 1994), indeed, no survey design will be ideal for all purposes (Kenkel, Juhász-Nagy, et Podani 1990). For example, James et Fairweather (1996) explained that overly small sample sizes will not provide precise descriptions of beach macrofauna species because the survey may fail to consider all sources of uncertainty, confounding large- and small-scale variation.

In this paper, we assess clam monitoring survey performance in Arcachon Bay, contrasting two survey designs : StRS, the current design, and GRTS. Using a semi-virtual population created from the results of a real survey (2012), we compare the performances of these two survey designs in terms of precision and survey effort. Finally, we estimate the impact of differences in sample size on the overall survey cost.

Materials and methods

Studied species description : biological and ecological aspects

Three different clam species are found in Arcachon Bay : the cross-cut carpet shell (*Tapes decussata*), the golden carpet shell (*Paphia aurea*) and the Japanese carpet shell (*Venerupis philippinarum*) (Bertignac et al. 2001). *V. philippinarum*, also called "Manila clam", is the most abundant species among them. It can tolerate salinities from 15 to 50 g L-1 (Le Treut 1986), but their growth is highly determined by temperatures and trophic resources (Melià, De Leo, et Gatto 2004 ; Tamayo et al. 2015). Other environmental factors can also have an impact on growth and survival, such as turbidity, immersion time, sedimentary characteristics, dissolved oxygen concentration and parasites (Goulletquer et Bacher 1988 ; Soudant et al. 2004 ; Gosling 2008). With *V. philippinarum*, growth does not stop during winter, unlike the other two species. This capacity allows Manila clam to reach an exploitable size very quickly (3 or 4 years) and makes this species cost-effective to harvest (Le Treut 1986). Research on the Arcachon Bay clams highlights a growth deficiency above 32 mm (Caill-Milly et al. 2012). This is not the case in other French or foreign production sites (Dang 2009 ; Caill-Milly et al. 2012). Manila clams are sexually mature from 20 mm and can reproduce several times a year.

Manila clams' natural habitat comprises the medio-coastal fringe of sheltered bays, estuaries and river mouths. They favour areas with low swell and preserved areas which create frequent water renewal (Le Treut 1986). Manila clams are benthic bivalve and live buried in the soil at a variable depth of 7-12 cm, depending on size and season, with the adults living deeper than juveniles. The environmental conditions such as regularity and duration of water flow, temperature, and the thickness and porosity of sediments have a direct influence on this clam micro-distribution (Olu K et al. 1996). Earlier, (Walker et Tenore 1984) found that Manila clam density varies widely depending on sediment sub-substrates. Their results match the ecological preferences of clams found by Tamura (1970), who cited an ideal living environment composed of 20-60% sand and 20-30% mud. Manila clams have an aggregative type of spatial distribution (Kalyagina 1995). This has been confirmed by many studies which also emphasized high spatial variability, regardless of the scale or the method of sampling (Juanes et al. 2012). This species can have lateral movements reaching 6 m per month (Tamura 1970). Its vertical distribution within sediment varies between year's periods and depends on individual age : juveniles are found near the surface whilst adults dwell at depth ranging between 7 and 12 cm, the latter being the maximal burrowing depth known so far (Le Treut 1986).

1. Objectives

Manila clam production in Arcachon Bay is important, with more than 500 T produced per year (Sanchez et al. 2014). In 2016, 41 commercial fishers depended directly on this harvest. In addition to this commercial activity, recreational exploitation of the Manila clam resource in Arcachon Bay takes place. Management strategies such as limiting the number of licences and established no-fishing protected areas have been used since 1996 to reduce over-fishing. To measure the impact of such management strategies, commercial fishers and scientists rely

on information from stock assessments.

2. Study area : environmental characteristics

Arcachon Bay is a 156-km² semi-sheltered lagoon in the southwest French coast (Fig. 21). Mostly composed of intertidal flats (110 km² within the inner lagoon), this mesotidal system has a sediment composition ranging from mud to muddy sand and is colonized by extensive sea grass meadows of Zostera noltei (Auby et Labourg 1996 ; Kombiadou et al. 2014). Influenced by external neritic waters and continental inputs (Dang 2009), the bay has a semi-diurnal macro-tidal rhythm. Temperature and salinity gradients within the bay are controlled by these water mass characteristics as well as by slow tidal water renewal (Bouchet et al. 1997 ; Plus et al. 2006). The Manila clam population is primarily located in an arc from west to south within the bay on its east side, in an area covering ~50 km². The primary area of location of *V. philippinarum* was divided into strata based on expert knowledge (fisheries scientists and commercial fishers). Each stratum represents an area that is as homogenous as possible (Yoccoz, Nichols, et Boulinier 2001 ; Zhao et al. 2016) (see Fig. 21) in terms of hydrology, sediment particle size characteristics, current patterns. This stratification was chosen because these factors were considered to be the most relevant for partitioning the spatial distribution of Manila clam. The stratification was mapped in the initial monitoring survey. Over time new strata in adjacent areas have been added, while some of the original strata are no longer sampled. In 2014, there were 17 strata in the survey. Previous to this there were 14 strata (from 2003 to 2006), 16 strata (from 2008 to 2010) and 19 strata in 2012.

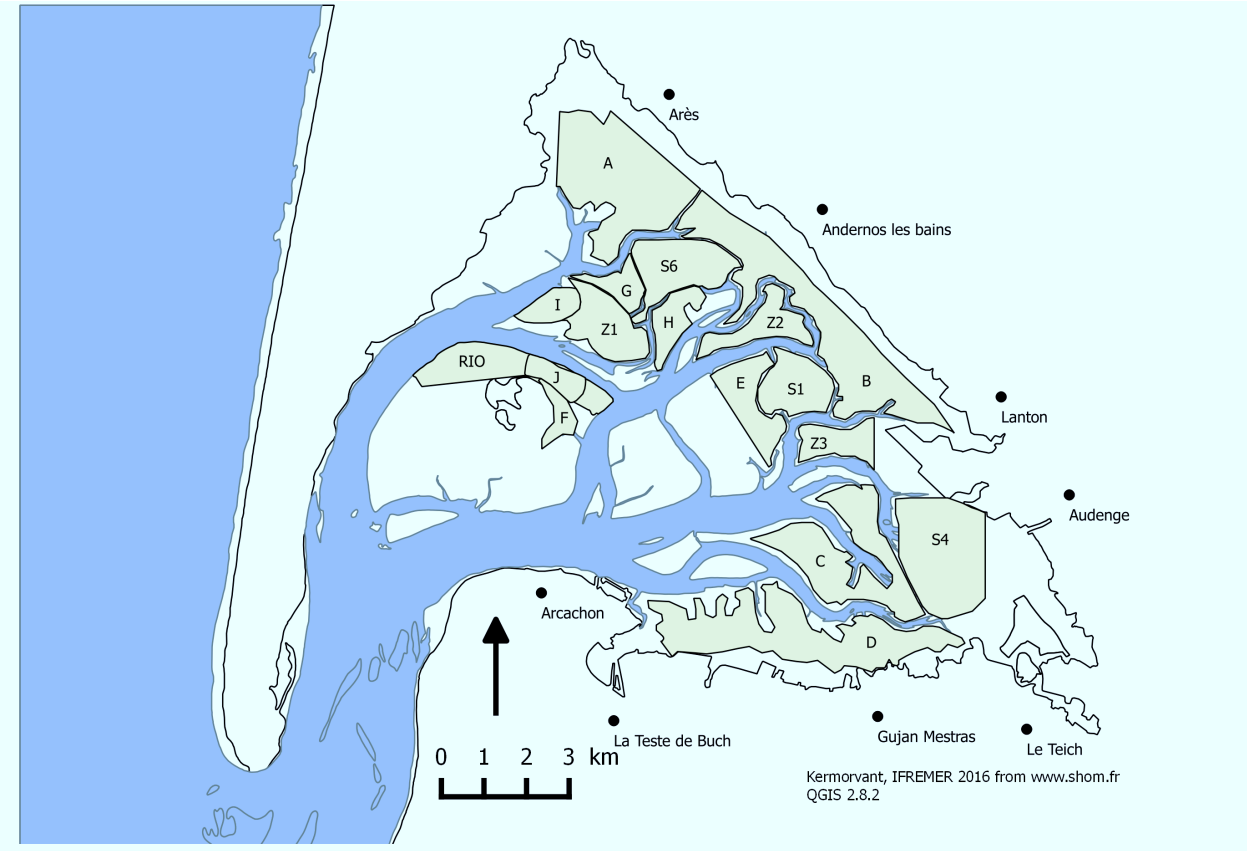


FIG. 21 : Survey site, Arcachon Bay, France, divided into 17 strata (A, B. RIO. Z3)

Monitoring surveys : classical methodology and new approach

The current survey design involves sampling at high tide using a Hamon grab aboard a professional boat. The Hamon grab is the recommended tool for sampling benthic macrofauna from coarse substrata (Le Treut 1986). It is regularly used on rough ground (Kingston 2009) and works well for sampling at the depth the Manila clam is buried at. Manila clams live buried at a mean of 12 cm of depth and the grab collects a sediment core of 0.25 m² (0.5 m x 0.5 m) on a 0.2 m depth. For this study, we will suppose that the sampling gear does not involve sampling bias. The core samples are filtered on board with running water over a set of three sieves with 2, 1 and 0.5 cm mesh size. All specimens of *Veneridae* are sorted and identified. Counting and measurement (to the nearest 1 mm using a slide calliper precision) are undertaken on board or in the laboratory, depending on the year. Surveys have been performed every 2 years since 2006 in late spring (and was undertaken every 3 years between 2003 and 2006) (Sanchez, Caill-Milly, et De Casamajor Marie-Noelle 2012). The whole field survey, including prior requirements, typically takes 18 days. The survey effort is 10 stations per km² (each sample station being identified by its geographic coordinates). Stations are randomly located within each of the strata. This gives a proportional stratified sampling, with a survey effort proportional to the strata surface size.

Stratified sampling is one of the most used designs in ecology. StRS is one of the most commonly used survey design in ecology, due to its ease of use and its flexibility. Additional samples can be easily added at the survey design a posteriori. The StRS technique involves dividing the study area in strata then randomly sampling within each stratum. For a better performance, individual stratum are created to be relatively homogeneous (Yoccoz, Nichols, et Boulinier 2001 ; Zhao et al. 2016). Recently, there have been a number of new spatially balanced survey designs that are becoming more popular (e.g. Stevens et Olsen (2004) ; Robertson et al. (2013) ; Brown, Robertson, et McDonald (2015)). One of the first spatially balanced designs was GRTS, designed for environmental monitoring over the long term and at a large scale (Stevens et Olsen 1999, 2003 ; Stevens Jr et Olsen 2004). The spatial balance provided by the GRTS design addresses a major disadvantage of StRS for our population. With GRTS, no sample station will be excessively far from another station (in this, it resembles but surpasses the systematic sampling strategy) and very few stations will be extremely close to another. Importantly, GRTS is known to have high efficiency compared with other designs. As with StRS, the main area can be divided in strata but the main difference is that GRTS use an algorithm for spatial balance instead of a simple random process.

GRTS has been used in several studies. For example, it has been used to determine bull trout (*Salvelinus confluentus*) population status through counts in basins of the Columbia River Plateau in the USA (Jacobs et al. 2009) and to develop ArcGIS tools via a forest biodiversity survey in a case study in Hunan Province, China (Li, Xu, et Zhou 2012). Here, we illustrate the use of GRTS, and provide a comparison with StRS for bivalve surveys.

Methodology for comparing two survey designs and choosing the best way to sample the clam population

The performances of both the GRTS and StRS were evaluated using estimators of total abundance based on the hypothesis of a known population (virtual one) of Arcachon Bay Manila clams.

The survey design comparison followed these classical steps : (i) the studied variable distribution was considered to be known in the population ; (ii) samples were selected from this population with both designs and results compared.

1. Building the Arcachon Bay clam population

In this study, a single year dataset was used (2012), but this methodology can be applied to the five other datasets (2003, 2006, 2008, 2010 or 2014). The spatial distribution of clam abundance (expressed in number per 0.25 m^2) and biomass (expressed in grams per 0.25 m^2) was estimated for the year 2012 from survey data. We used geostatistical analyses and a kriging technique from the RGeostats library (Renard et al. 2014) in the R environment (Team 2014) to create a semi-virtual population of clams. The analysis had the following steps : 1) exploratory data analysis ; 2) variography analysis with adjustment of a model to the experimental variogram (nugget effect and isotropic exponential model) ; 3) global assessment and associated variance. The implemented interpolation method used was block

kriging model with a 200m sliding neighborhood. This methodology created a model (usually a smoothing model) to build a complete semi-virtual population. It used the observed clam data from the sample sites to "fill in" the missing data from the sites that were not sampled. Fig. 22 present the virtual populations created for clam abundance and biomass. The kriging settings were chosen according to expert knowledge, this could bring some bias in results. But as the same semi-virtual population is used to assess the performance of both survey designs we decide to not take it into account.

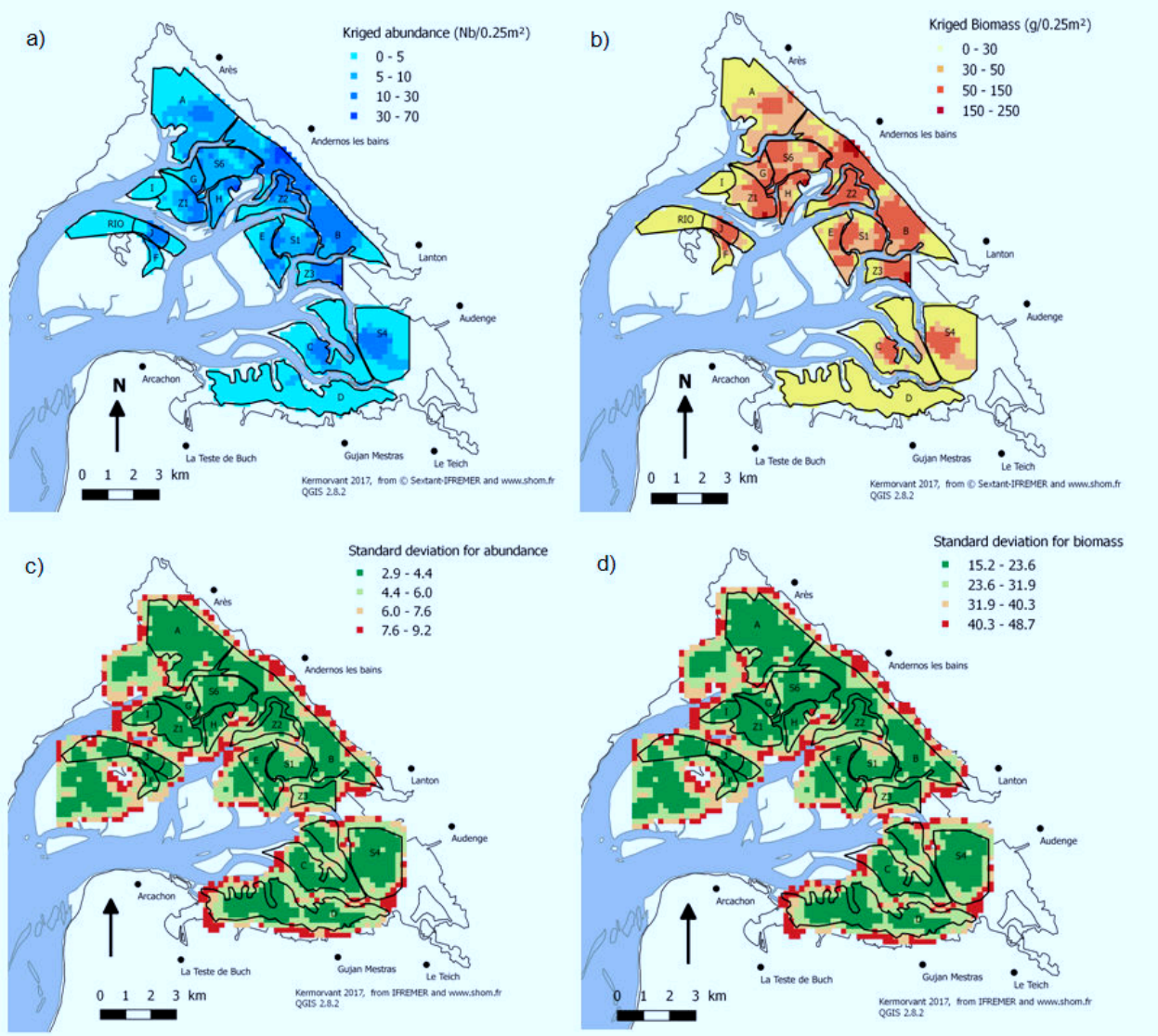


FIG. 22 : Semi-virtual clam populations for abundance (a) and biomass (b) parameters and associated standard deviation (c and d)

2. Performance assessment : the optimal number of sample stations per design and the corresponding precision

We decided to compare the two survey designs in terms of the optimal number of samples within each stratum which is directly linked to the cost. Here, we define the optimal sample size as the minimal number of samples that need to be collected in each stratum to achieve the desired precision of the clam population estimate. To determine this, selection of the GRTS and StRS points was performed by using a statistical sampling methodology with the "spsurvey" and "sp" packages in R software, respectively [Pebesma et Bivand (2005) ; Bivand, Pebesma, et Gomez-Rubio (2013) ; Kincaid et Olsen (2015).

MacKenzie (2006) defines Accuracy = MSE= variance + bias^2. As StRS and GRTS are probabilistic survey designs, every sample has a known non-zero probability of selection, leading to unbiased estimates of the mean and variance (and their confidence interval) for variable of interest (Albert et al. 2010). Also, bias can be considered as null when the sampling is repeated many time and thus accuracy become synonym of variance.
The steps to determine precision of estimations by survey design and then compare their performances were as follows : Step 1 : For a given size n (n = 1, 2, 3, .,N) of sample stations within each stratum, we took 1000 different samples for GRTS and for StRS. Each sample at the given survey effort is called a replicate. The same number of stations, n, was chosen within the stratum, first using the StRS design and then the GRTS design. Step 2 : For each replicate sample j (j = 1, 2, 3, .1000), at each level n of survey effort, the mean biomass and mean abundance were estimated, and the confidence interval (95%) of the estimated mean biomass and the estimated mean abundance were computed. The width of the confidence interval and the relative precision of the estimators of the two designs were calculated as :

$$e_{jn} = 2z \times \frac{s_{jn}}{\sqrt{n}},$$

where z is the quantile of the standard normal distribution (1.96 for a 95% confidence interval) n is the survey effort within the stratum and sjn is the standard deviation of the within-stratum sample with a sample size of n for replicate j. Then, for each replicate j, the precision was calculated as :

$$P_{jn} = \frac{e_{jn}}{\bar{x}_{jn}} \times 100,$$

where e_{jn} is the confidence interval width and $\bar{x}_{jn}$ is the estimated mean (biomass or abundance). The overall precision for the stratum for the sample size n was calculated as the average of the 1000 estimate of precision :

$$P_n = \frac{\sum_{j=i}^{1000} P_{jn}}{1000}.$$

Step 3 : The sample effort was varied within the stratum until three levels of nominal precision for biomass and for abundance (Pn = 20%, 10% and 5%) were achieved for the two survey designs. We defined this as the optimal sample size per stratum for a given level of precision. As there was a difference in the optimal size for estimating biomass compared with

abundance, we selected the larger of the two sample sizes as the value for the corresponding stratum.

Step 4 : Comparison of the two methods was carried out using statistical tests and by comparing the optimal sample sizes. We calculated paired Wilcoxon tests to evaluate the significance of the differences in the optimal sample sizes of the GRTS and StRS methods for the different precision levels. Tests were conducted with the R software, "stats" package (Team 2014).

This methodology is summarized in Fig. 23.

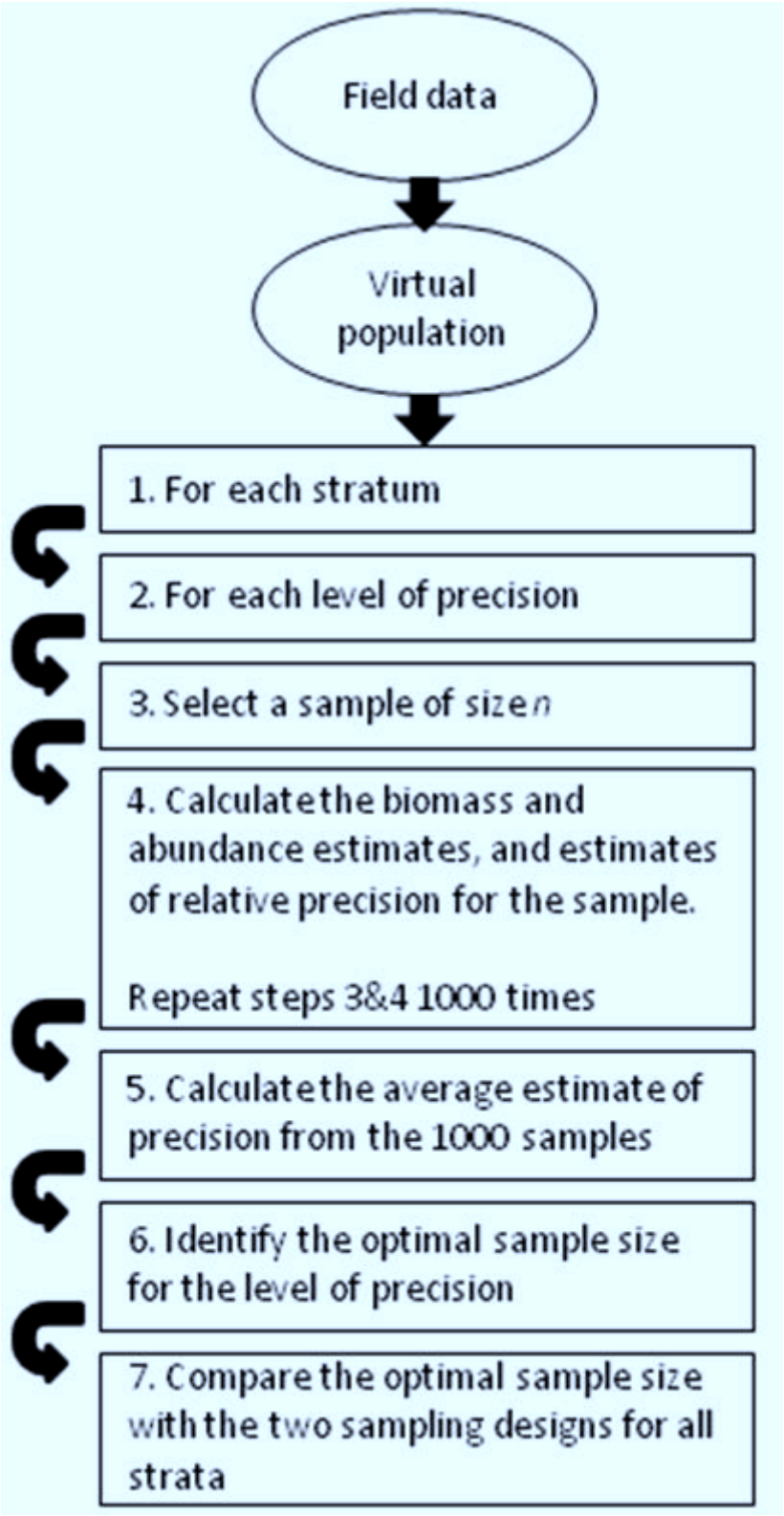


FIG. 23 : Methodology used to assess the performances of StRS and GRTS on Arcachon Bay manila clam population

Back to a practical point of view : assessment of monitoring cost

We based our calculations on the 2012 field survey budget to assess the survey costs. The survey cost was decomposed into two parts : a fixed cost (the costs of meetings to prepare the survey, material, data treatment, meetings to present the results to the commercial fishers and administration, etc.) and a variable cost which depended on the sample size (the costs of boat and grab rental, participation costs of scientists and commercial fishers, etc.). We estimated the fixed cost to be €12 000 and the cost of one sample was approximately €80. We assessed the overall cost of the simulated designs as follows :

Survey cost = fixed cost + number of samples x 80.

Results

Three main outcomes are presented : a map of GRTS and StRS applied in the same strata to visually observe the differences in their survey plan, then a comparison using the optimal sample size and finally a comparison using the field price.

Visually comparing the designs

To compare the spatial distribution of the two survey designs, we show (Fig. 24) an example of plans obtained from a single StRS and a single GRTS survey, with the same survey effort of 15 samples in three strata (G, Z1 and S6). In this part, only one random sample was used for both designs and mapped. The two survey maps illustrate one important feature of the benefit of GRTS, which is that it will always produce a sample that is well-balanced, and with more complete spatial coverage. In fig 24a there are noticeable clusters of sample points and areas devoid of samples. With GRTS (Fig. 24b), there is more consistent or even coverage of samples, without having a fixed regular pattern of sample points. In Fig. 24b, the sample points still appear to be randomly located without either extreme regularity or extreme clustering.

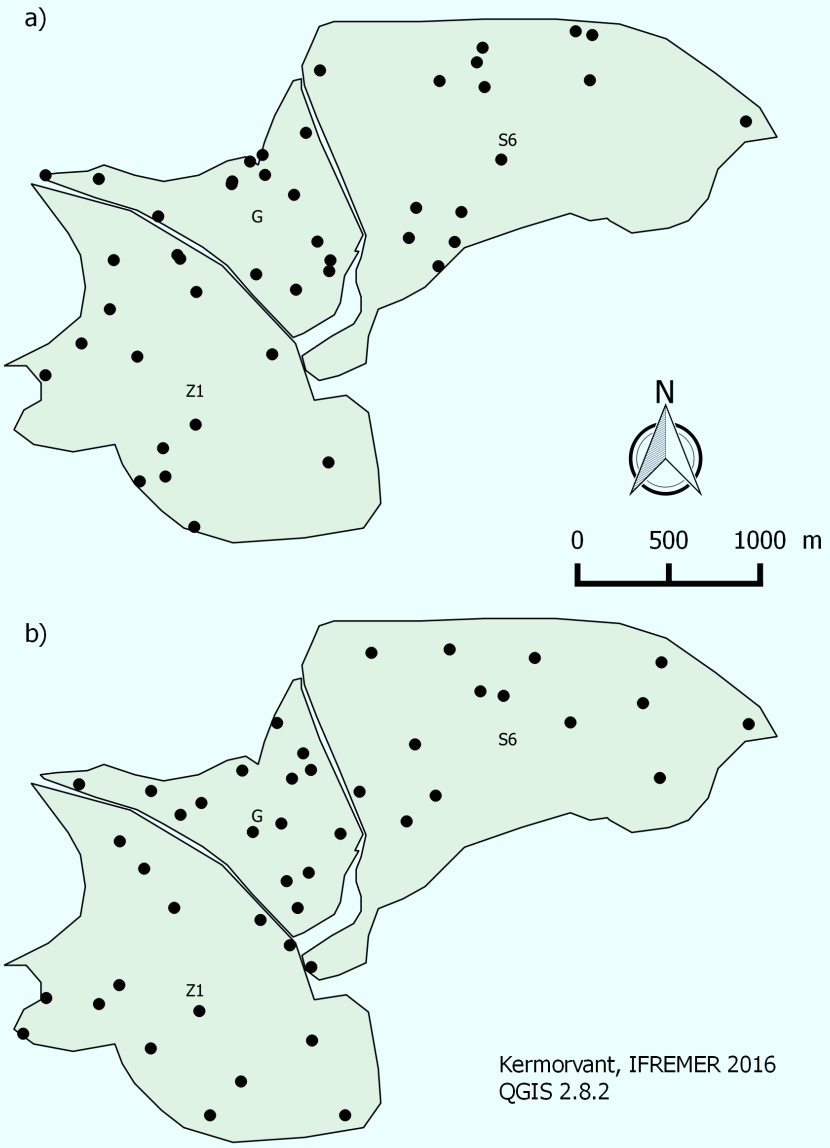


Fig. 24 : Examples of sampling plans with 15 sample units per stratum for the stratum S6, Z1 and G using (a) StRS design and (b) GRTS design.

Comparing the optimal sample size of each design

The optimal sample sizes for the three levels of precisions for estimates of biomass and abundance for each stratum are displayed in Fig. 25.

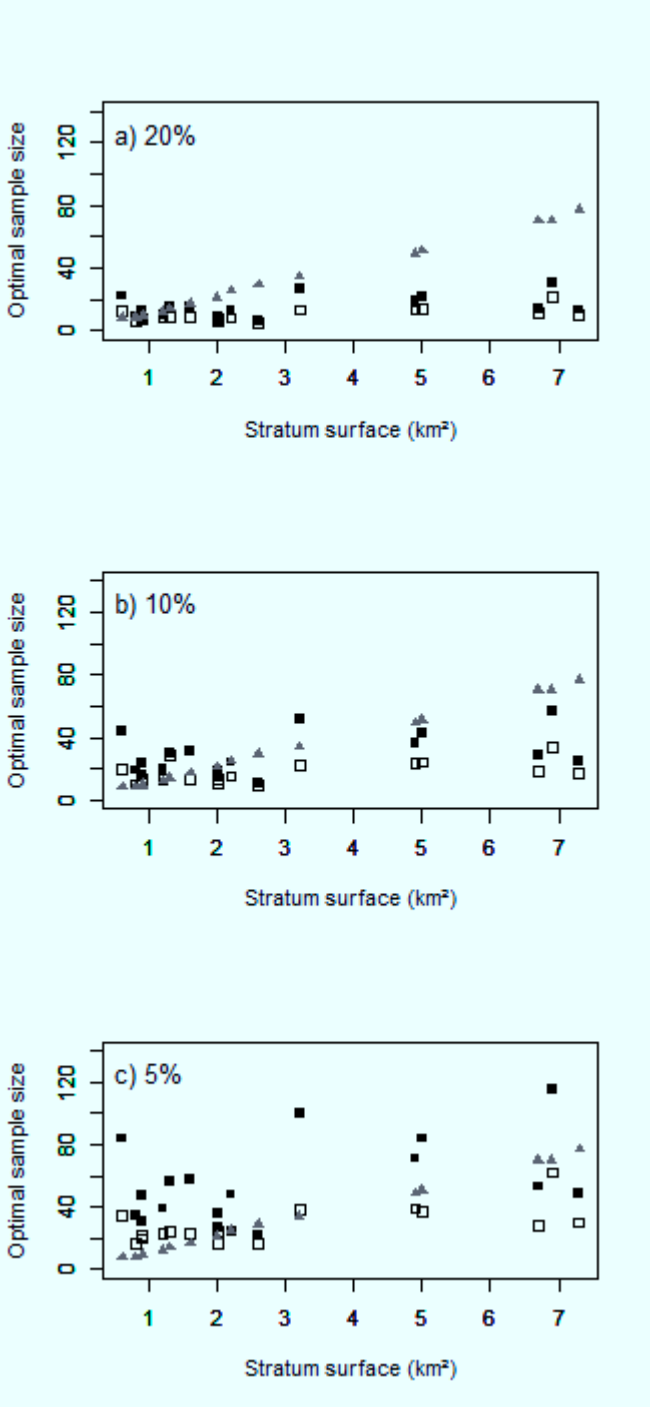


FIG. 25 : Optimal sample size estimated for the StRS and GRTS designs by stratum surface with three precision thresholds : (a) 20 percent, (b) 10 percent and (c) 5 percent. Full triangle = actual number of stations surveyed during the 2012 campaign ; full square = estimated optimal sample size using StRS, empty square = estimated optimal sample size using GTRS.

The optimal sample size for GRTS and StRS appears to be unrelated to the size of the stratum. The most prominent feature of the graphs in Fig. 25 is that fewer samples were

needed with GRTS than with StRS to achieve the same level of precision. The difference in the optimal sample size for GRTS compared with StRS was most pronounced for the more precise surveys (Fig. 25a).

The differences in the optimal sample size between GRTS and StRS for the different levels of precision were all significant (Table 1).

Table 1. Estimated optimal sample size and estimated survey costs for 5%, 10% and 20% level of precision with StRS and GRTS. Optimal sample size is computed for each stratum separately, values for the two survey designs are then compared with paired Wilcoxon test.

Precision	Wilcoxon tests		Overall sample sizes Actually = 525		Survey costs (= 1 000 €) Actually = 53	
	test statistic (V)	P-value	StRS	GRTS	StRS	GRTS
5%	0	< 0.001	955	481	76	38
10%	0	< 0.001	493	281	39	22
20%	0	< 0.001	248	167	20	13

Comparing the survey costs of the designs : highlighted issue

We used the optimal sample sizes for the three target precision levels described in the previous section to estimate the total survey costs (Table 1). The total sample size for GRTS and StRS was estimated from the sum of the individual strata observed using the optimal sample size (Table 1). The 2012 total sample size and cost are also shown. Given the reduced number of sample points for the same level of precision, in our study, GRTS always costs less than StRS. Further cost savings can be made by reducing the targeted precision level ; for example, at 5% precision with GRTS, the total cost of the survey would be €50 000 and half that cost if only 20% precision was acceptable. Interestingly, even with the highest level of precision we used (5%), the cost of the total GRTS survey would be less than that of the 2012 survey.

Discussion

Monitoring survey designs should be designed to accommodate the end-users' requirements for precision or performance of the survey results, and the cost budget (Yoccoz, Nichols, et Boulinier 2001 ; MacKenzie et Royle 2005 ; Guillera-Arroita, Ridout, et Morgan 2010 ; Moore et McCarthy 2016). In practice, this often translates to questions such as the allowed or permissible maximal imprecision, and the maximum allowable survey cost. For these fisheries, these two requirements are both priorities. In 2016 and 2017, for example, the clam survey did not occur in Arcachon Bay because the commercial fishers could not afford it. Therefore, there is some urgency to find a more cost-effective design where the survey results have enough precision to be able to contribute to the clam management strategy.

We have developed a methodology to assist commercial fishers to test design surveys. The desired level of precision for the survey can be set along with a variable cost component. The use of a simulation study allowed us to test a panel of different levels of precision and

compare two survey designs. This simulation approach has been used in many other studies. For example, simulations were used in an instrument-based survey to evaluate alternative survey designs for Arctic marine mammal populations (Conn et al. 2016) and to optimize animal detection given the breeding behaviour and logistical access for threatened species (Lanier, Bailey, et Muths 2016). We found only a few simulation studies using regional-scale populations. For example, Ene, Næsset, et Gobakken (2016) compared the performances of above-ground biomass estimation methods at a regional scale. Our study demonstrates the use of a simulation method to assist in designing a survey of a benthic lagoon population. In addition, Defeo (2011) pointed out the paradox that in reactions to the environment, the structure of regional populations as well as their dynamics are among the most poorly understood ; however, commercially exploited bivalve populations should be relatively easy to study (located close to the shore, no or limited animal movements). Our study provides a method to implement or optimize field studies and surveys. We cannot deny that some bias can occurs between the real population and the semi-virtual population we simulated. This bias can be due to the choice of kriging parameters and to gear efficiency during field sampling.

This study demonstrates that for clam population monitoring surveys in the Arcachon Bay, GRTS should prove more efficient than the current survey design (i.e. StRS). We observed in our study that the cost advantage of GRTS compared with StRS increased as the desired level of precision increased (where 5% precision is higher than 10% precision, etc.). In our simulation results, the relative similarity in optimum sample size across different sized areas may be a result of different patterns of spatial autocorrelation in different stratum, a topic to be considered further. It seems that the total area of the stratum had less effect on the precision than the heterogeneity of the clam population within the strata. Indeed, the optimal sample size defined by our methodology is a fixed minimal in each stratum. However, we caution that this finding is for this study only, where there is high heterogeneity in clam distribution, irrespective of the size and location of the strata. It is possible that each stratum includes parts with very high clam abundance and others without clams despite the best efforts to delineate homogeneous areas. This hypothesis is based on previous bivalve studies which found that clams often are patchily distributed (e.g. Kalyagina (1995) ; Armonies (1996)).

Surveys for populations with high heterogeneity are an ongoing area of research. For example, McGarvey, Burch, et Matthews (2016) suggested that systematic survey designs are superior to random designs for clustered populations, and Christianson et Kaufman (2016) highlighted that estimates of heterogeneity can be obtained with a well-chosen survey design. It seems important to have an idea of the heterogeneity and the aggregation of the studied variable or population in order to choose an appropriate survey design. However, complex designs such as GRTS have not been tested yet in benthic populations in a lagoon area. This hypothesis of non-homogeneous clam distribution in Arcachon Bay will be the subject of a future study.

The applied issue highlights that even if GRTS surpasses classical StRS in terms of requiring a smaller total number of samples, it is important to keep in mind that having more samples in both designs led to better precision in the clam population estimation. Countering that, having more samples leads to higher costs. Alternatively, relying on fewer samples involves

a loss of precision but produces a substantial reduction in total survey costs. Comparing the simulated studied designs results and the actual precision from this field survey is not an easy task considering that this field survey use a proportional to the strata surface size survey effort and so achieve different precisions for each stratum while, in the simulated study, the achieved precision is the same for all strata. With the 2012 clam survey in Arcachon Bay, some strata had very good estimated precision (2.5%) but others were far worse (~50%). We could average these values but this would hide this heterogeneity. Another issue with the current field design is that in the smallest strata, and because the number of sample stations is chosen proportional to the size of the stratum, there are too few stations to calculate a reliable estimate of within-stratum precision. Conversely, in the largest strata, proportional sampling results in very high survey effort with only a marginal benefit in precision. Our proposed methodology reveal that with more constant effort among strata we should improve the estimation of precision, and standardize precision across all strata in the Arcachon Bay.

The final decision of which survey design should be used in future surveys is for the commercial fishers to decide. The applied issues highlight that commercial fishers have to make a choice between overall survey cost and precision. The trade-off is the following : the survey cost can be lowered but it will reduce precision, or the survey cost can stay as it was with a gain in precision.

Conclusions, Perspectives

Our simulation study suggests that GRTS would be more efficient than StRS (based on the 2012 clam survey). We assume that the Manila clam population shows some spatial heterogeneity in their distribution within Arcachon Bay. It is also possible that some temporal heterogeneity may exist. These aspects will be tested in the near future by using our methodology to include other years of the monitoring survey (2003, 2006, 2008, 2010 and 2014) (Caill-Milly et al. 2003, 2008 ; Caill-Milly, Duclercq, et Morandeau 2006 ; Sanchez, Caill-Milly, et De Casamajor Marie-Noelle 2012 ; Sanchez et al. 2014). After conducting this work, the most appropriate design will be discussed with the stakeholders and the recommended sample size for each stratum will be set for future surveys, the next one being scheduled for 2018.

This first study confirms the possibility of improving clam monitoring. GRTS appears to be an efficient survey design that should be used instead of StRS. We note that sample design is a rapidly evolving field of science and other advanced spatially balanced designs have already been described, despite limited field testing (e.g. Robertson et al. (2013) ; Brown, Robertson, et McDonald (2015)). We will watch this emerging area of research for more ways to improve clam monitoring.

Acknowledgments

Data were collected in the context of the Manila clam monitoring programme financed by the French institute Ifremer, European grants (IFOP - instrument Financier d'Orientation de la Pêche ; and FEP - Fonds Européens pour la Pêche), the French government (MEEM - Ministère de l'Environnement, de l'Energie et de la Mer), Aquitaine Regional Council,

Gironde General Council and professional fishing organizations. This work was supported by Ifremer's scientific direction within the framework of site policy grants (DESCARTES 2 project) aiming to reinforce collaborations between Ifremer and regional academic partners. This work was also supported by "Communauté d'Agglomération Pays Basque - Euskal Hirigune Elkargoa" through a thesis grant. We also want to acknowledge the two reviewers of this paper for their good advises and recommendations.

Résultats issus de ce chapitre :

- Le suivi de la ressource en palourde dans le bassin d'Arcachon est financièrement lourd et n'a pu être effectué ni en 2016 ni en 2017 en raisons de difficultés financières d'un des partenaires.
- De récentes études ont montré qu'utiliser un protocole spatialement équilibré pouvait améliorer l'efficacité des suivis, et donc réduire leur coût.
- La méthode développée dans cette thèse est ici utilisée pour comparer l'efficacité d'un de ces protocoles, le GRTS, à celle du SRS - utilisé jusque-là en routine pour cette campagne - en se basant sur les résultats de la campagne de 2012.
- Basé sur un population reconstruite mathématiquement, le nombre d'unités statistiques à récolter dans l'échantillon pour attendre une même précision dans les estimateurs de population est quasiment divisé par deux (- 43 %) lorsque le GRTS est utilisé à la place du SRS sur les résultats de la campagne 2012.
- La méthode permet d'optimiser un suivi à partir de données issues d'une seule campagne.
- La stabilité (spatiale et temporelle) de ces résultats ainsi que la représentativité de la population reconstruite doivent être étudiées avant de les transférer d'un point de vue opérationnel aux campagnes à venir (voir Chapitre V).

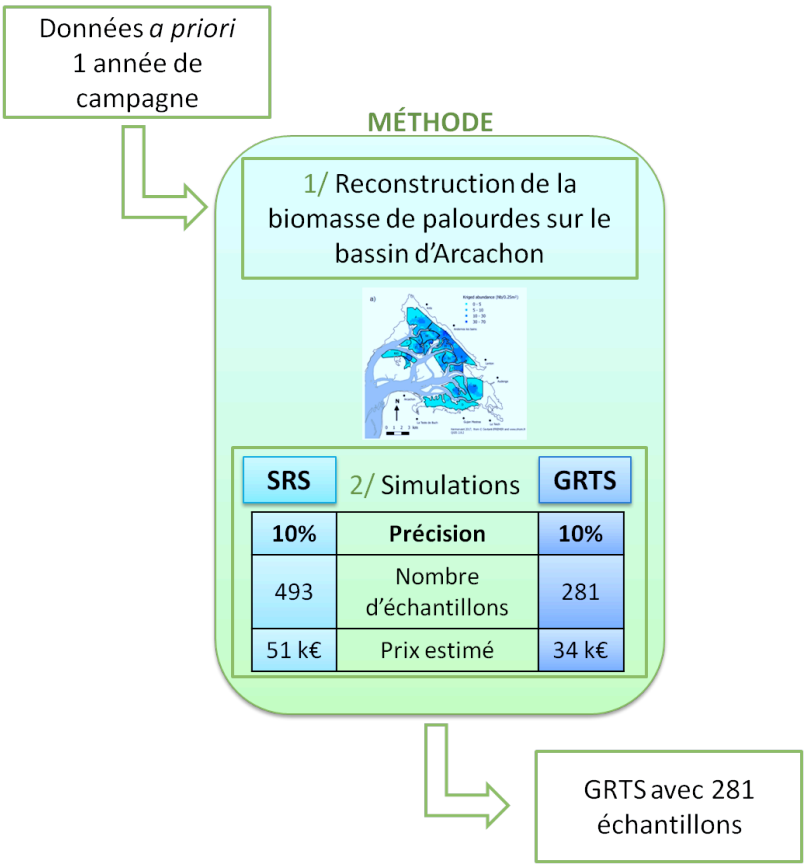


FIG. 26 : Conclusions du CHAPITRE IV

CHAPITRE V : Optimisation d'un suivi à partir de plusieurs saisons de données disponibles

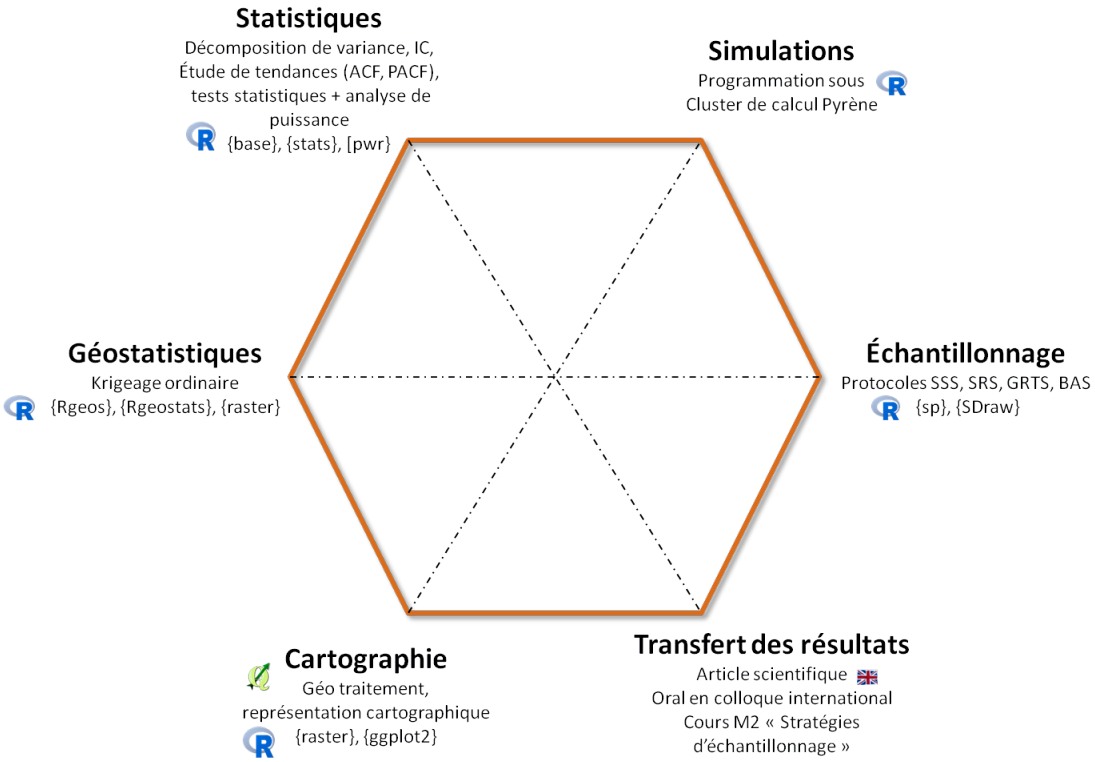
Synopsis :

Ce dernier chapitre s'applique à démontrer que la méthode séquentielle développée durant cette thèse peut être utilisée pour une optimisation dans pérenne de suivis environnementaux. Le suivi de la palourde dans le bassin d'Arcachon est en réalité effectué depuis 2006 tous les deux ans. Il existe donc un jeu de données conséquent, correspondant à 6 années de campagnes scientifiques.

Trois protocoles d'échantillonnage sont testés sur chacune des données récoltées lors des 6 années de campagnes passées. Le coût de leur application réelle sur le terrain est également calculé. Les protocoles sélectionnés sont : 1 - l'aléatoire simple (SRS - celui utilisé lors des campagnes passées de ce suivi), 2 - le Generalized Random Tesselation Sampling (GRTS - un protocole spatialement équilibré connu pour ses bonnes performances) et 3 - le Balanced Acceptance Sampling (BAS - protocole spatialement équilibré nouvellement développé, encore jamais testé sur une population réelle).

En premier lieu, il a été confirmé que les deux protocoles spatialement équilibrées avaient une performance meilleure que l'aléatoire simple. Ces deux protocoles avancés ont une performance comparable et permettent d'atteindre une même précision dans les résultats avec seulement un nombre moitié moindre d'échantillons que l'aléatoire simple. Les résultats de simulations montrent que le nombre d'échantillons nécessaires aux protocoles pour atteindre une précision fixée dans les résultats est constant au cours du temps. Cette caractéristique pourra nous permettre de proposer un nombre d'échantillon fixe à prélever dans toutes les prochaines campagnes, et cela, malgré l'existence de variations spatiales et temporelles dans la distribution de la palourde.

Compétences développées/utilisées pour ce chapitre :



Valorisations de ce chapitre :

- **Publication :**

KERMORVANT, Claire, CAILL-MILLY, Nathalie, BRU Noëlle, D'AMICO, Frank. Optimizing cost-efficiency of long term monitoring programs by using spatially balanced sampling designs : The case of manila clams in Arcachon bay. Ecological Informatics, 2019, vol. 49, p. 32-39.

- **Communication orale en conférence internationale :**

KERMORVANT Claire, BRU Noëlle, D'AMICO Frank, CAILL-MILLY Nathalie. Using spatially balanced sampling designs to optimise cost-efficiency of long term monitoring programs : application to Manila clam in Arcachon Bay. Oral communication in International Symposium of Oceanography of the Bay of Biscay ISOBAY XVI. June 2018. Anglet, FRANCE - **Prix de la troisième meilleure communication orale étudiante**.

Abstract

Lack of funds is one major issue in ecology, in particular at local scale. It is known that sustainable management of a natural population requires a good understanding of its functioning, itself dependent on a good long term monitoring program. Such programs are usually very difficult to implement, especially for resources char- acterized by high spatio-temporal variation in their distribution, resulting in a trade off between efficiency and costs. Today, thanks to rapidly evolving statistical theory, new survey designs are developed, some with the characteristic of well balancing samples in the study area. This paper aims at demonstrating that theses advanced sampling designs perform better than the usual ones for long term monitoring program of local resources, with the added benefices of saving money and also increasing results accuracy. To prove it, and for it high spatio- temporal variation in it distribution, we choose the example of Manila clam's stock monitoring in Arcachon bay. This stock is under high scrutiny and last campaigns could not be done because of lack of funding (at least 50,000€/survey). We use a simulation study based on real data to assess and compare performances of news and older sampling designs on this survey. Three sampling designs are tested in both of the 6 past monitoring campaigns data and we estimate the cost of their application in the field. Selected sampling designs are : 1 - simple random sampling (SRS - the one used in the past years of this monitoring program), 2 - generalized tessellation sampling (GRTS - a recent spatially balanced sampling design known for its high performance) and, 3 - balanced acceptance sampling design (BAS - a newly developed spatially balanced sampling design, never tested yet in a real population). We first confirm that the two spatially balanced sampling designs perform better than simple random sampling. Both of the advanced sampling designs perform equally and allow achieving same accuracy in the results with almost half sampling intensity than SRS. This makes them so cost-effective that 30% of each campaign price could be saved if they were used. Moreover, the three sampling designs need a constant sample size thought years to achieve a fixed accuracy in results. This will permit us to fix one sample size that could be done for all future campaigns ; and this, despite the existence of spatial and temporal variations in clam's distribution.

Introduction

Most of ecological studies involve spatial or temporal data, even both of them. But performing an exhaustive survey of any phenomenon is almost impossible or may prove to be tricky (Chiarucci et al. 2003 ; MacKenzie 2006) because of time (Cox, Cox, et Ensor 1997) and/or money lack (Jackson et al. 2008 ; Lazarina et al. 2014 ; Theobald et al. 2007). The common practice for dealing with this problem is trying to infer the targeted phenomenon on the basis of samples from the original popu- lation (MacKenzie 2006). Results of these sampling procedures are more and more accurate when sample size increases. But the sampling procedure costs increases also according to the sample size. There is a trade-off between the results accuracy and the cost of the survey. A good resource management involves a good knowledge of the popula- tion of interest, and also accurate results in sampling procedures. But, when it comes to reality, especially at local scale, funds devoted to ecological studies are usually unsubstantial and far from being large enough to get accurate estimates. In this study, we want to illustrate that local scale long term monitoring program could be optimised (meaning in the context of our study being less expensive but providing a better accuracy) only by changing the survey design. In this respect, we choose to apply our methodology on manila clam long term mon- itoring in Arcachon bay.

Manila clam (*Venerupis philippinarum*) is one of the five highest produced bivalves in the world (Astorga 2014) with 4,010,702 T produced in 2014 (source : Fao FishStat). It is worldwide present and economically important for fisheries industries. In many countries, bi- valves are harvested for food and baits (McLachlan et al. 1996). But stocks are of concern in many locations. Numerous studies around the world on natural populations were undertaken with the aim to assess the geographic spread of this species (which can be invasive), in Poole Harbour in UK (Jensen et al. 2004), San Francisco Bay (Carlton et al. 1990), Venice Lagoon (Pranovi et al. 2006), Southern California (Talley, Talley, et Blanco 2015) and in Santander Bay, Spain (Bidegain et al. 2015). Other studies were carried out to further understand the influence of factors associated with mortality in manila clam stocks (Dang 2009, @Dang-Correlationperkinsosisgrowth2013 ; Paillard, Allam, et Oubella 2004 ; Park et Choi 2001), to study hy- perparasites (Le et al. 2015), and to report micro plastics ingestions (Davidson et Dudas 2016).

Arcachon bay is a 156 km2 semi-sheltered lagoon located in the southwest coast of France. It represents, with the Morbihan Gulf, the main manila clam production area for this country. Manila clams are studied in Arcachon bay for the purposes of following descriptors of the stock, assessing its status and proposing an adapted management plan of this resource to the fishermen (co-management process between scientists and professional fishermen). Not only fishermen request and order the survey but they also help scientists during field procedure ; in turn, scientists suggest improved management solutions to fishermen.

The first survey was undertaken in 2003 (Caill-Milly et al. 2003) and subsequently carried out every two years since 2006 (Caill-Milly, Duclercq, et Morandeau 2006 ; Caill-Milly et al. 2008 ; Sanchez et al. 2010, 2014 ; Sanchez, Caill-Milly, et De Casamajor Marie-Noelle 2012). A stratified random sampling (StRS) was applied on 17 strata, with a sampling effort of 10 stations per km2 (this sampling effort is an empirical compromise between field

expertise, biological information and avail- able fund) ; each sampling station being identified by their geographic coordinates. A major concern with this monitoring survey, given that sampling effort, is that it is time consuming (approximately 500 sample stations are visited on each survey) and costly, although the aim is to keep the survey costs under a threshold of 50 K€. Without the support (financial and in kind) of fishermen, ongoing surveys are not assured. This indeed happened in 2016 and 2017 when no surveys were un- dertaken. This lack of data in 2016 and 2017 rose new problems, both for statistical and management reasons, because it broke the time series of stock's status indicators and impeded sustained effort in identifying population best management measures (e.g. restricted harvesting areas and periods, limited licenses number). Indeed, the Arcachon Bay monitoring survey is of particular socio-economical importance ; for example, on the socio-economical point of view, there were 70 manila clam's professionals fishermen in 2016 (source : IFREMER personal communication). Therefore it has been crucial and urgent to consider alternative monitoring survey designs that will be less expensive without jeopardizing accuracy of estimates.

Clams are spatially structured species, patchily distributed in their living area (McLachlan et al. 1996 ; Dugan et McLachlan 1999 ; Defeo 2003 ; Denadai, Cecília Z. Amaral, et Turra 2005) and this pattern makes them challenging to sample (C. A. Gray 2016a). Formers monitoring programs results in Arca- chon bay confirm the patchy spatial structure of manila clams in this area, but also highlight a temporal variation of abundance and biomass between both campaigns (Sanchez et al. 2014). For bivalve distribution studies, as for the whole ecological studies (Smith, Anderson, et Pawley 2017), there are only few papers which indicates the used sampling design. Some of them use systematic design (Bald et Borja 2001, 2005 ; Borja et Bald 2000) and other random survey designs transects selected randomly in space and time (Wekell et al. 1994). For studying distribution of intertidal macrofauna including bivalve on beach, nested survey design with transects was also applied (James et Fairweather 1996).

Stratified version of simple random sampling design was used for bivalve in the French Normano-Breton Gulf (Pitel et al. 2004), in Arcachon bay (Caill-Milly et al. 2003), in Morbilhan gulf (Berthou et al. 1997) and for exploited bivalve in the swash zone on exposed ocean beaches (Gray et al. 2014). It was proved recently that some sampling designs show better performance for clam's studies (Kermorvant et al. 2017). Here, we define sampling design's performance as the minimum sampling in- tensity needed by a sampling design to reach a fixed accuracy in the results. As generally accepted, we consider that the lowest is the sample size for the sampling design, the highest its performance. Under this acceptation, a previous study based on one year of Arcachon bay's manila clam data showed that generalized random tessellation sampling - GRTS (a variant of spatially sampling design) performs better than simple random sampling (SRS). As it allows obtaining a fixed accuracy with fewer samples, GRTS proves to be less expensive (and so more cost-effective) to use than SRS (Kermorvant et al. 2017). This conclusion was raised from a study based on a small dataset, limited to a one year-survey ; it is important to definitely prove that this conclu- sion is robust and highly generic, encompassing any kind of spatially balanced design. The first aim of this study is thus to confirm that GRTS does perform better than SRS for manila clam monitoring despite nat- ural inter-annual variability. To enlarge the picture, and conclude that any spatially balanced design can do better, we also aim at assessing performance of another spatially balanced sampling design, namely the balanced acceptance sampling (BAS)

recently brought available (Robertson et al. 2013).

This is the first time that this new sampling design will be used to optimise a survey, and that it will be compared with other sampling designs in term of cost-efficiency.

As a corollary, a second task we aim at bringing to life for practitioner in this paper is to address the existence of one sampling size by strata that give good estimates of abundance and biomass throughout a medium-term survey (here six years of manila clam monitoring data). And this, despite the existence of spatial and temporal variations in clams abundance and effectives. We argue that if such sampling size exists, it is then possible to propose the better sampling design that could be applied for the future field campaigns and provide managers and fishermen with the answer to the key question they all ask : how many samples have to be taken by strata to achieve a wanted accuracy in clam's biomass and abundance throughout the whole bay ?.

Beyond Arcachon bay manila clam population's specificities, other local populations throughout the world have to be surveyed with the same questions and limits (especially in terms of cost). The ultimate goal of our study is thus to explore transferability of the approach so that other monitoring programs could be optimised worldwide so that more cost-effective management of local resource can be made every- where.

Material and methods

Field survey

The primarily area of location of V. philippinarum in Arcachon Bay was divided into 17 homogeneous strata (Fig. 27) from expert knowledge (hydrological, sediment particle size characteristics, currents patterns, management point of view) (Caill-Milly et al. 2003) as expressed above. 14 strata are sampled since 2003, two (namely I and J strata) only since 2008, and one (namely RIO stratum) since 2012.

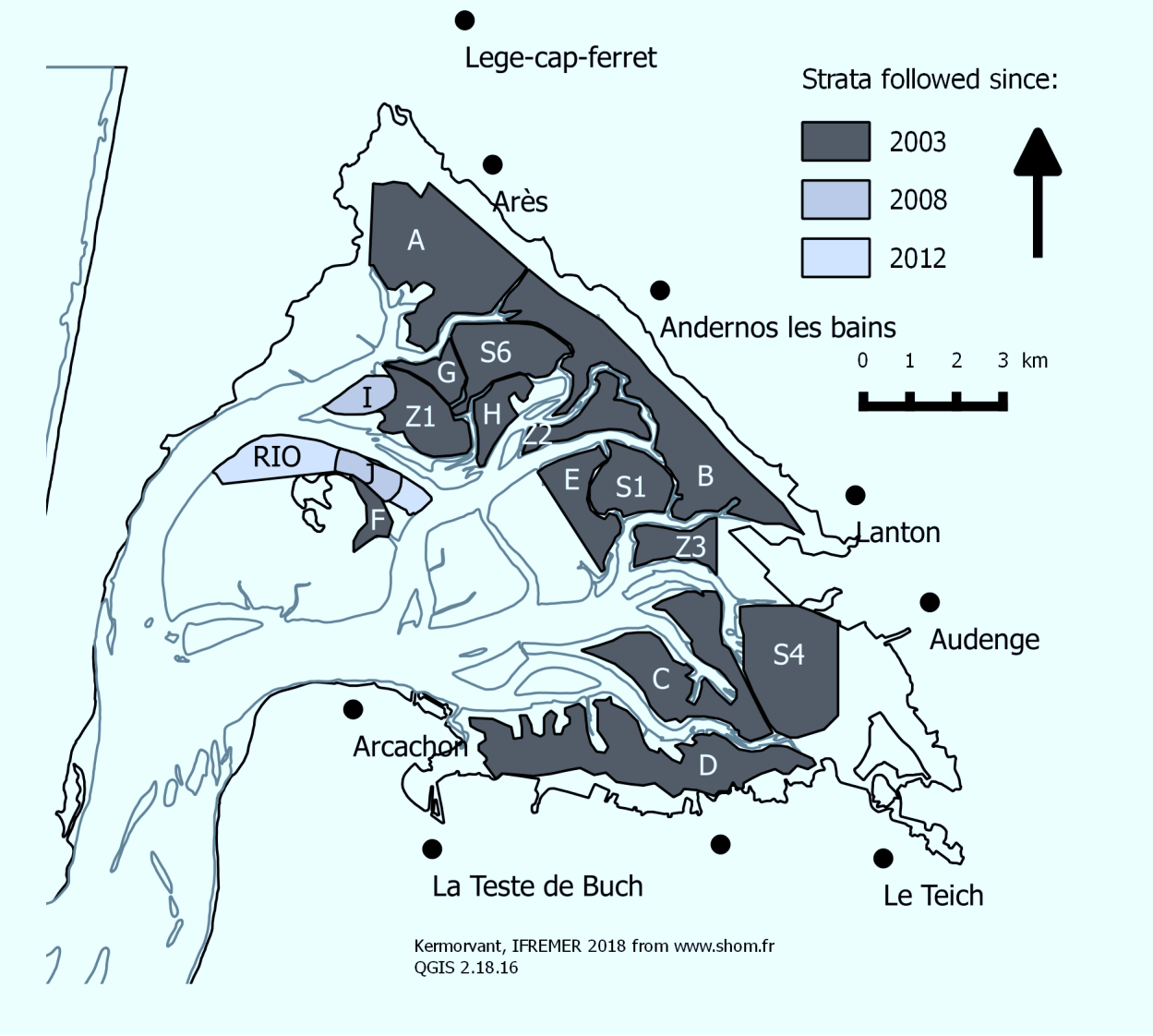


FIG. 27 : The survey site divided into 17 strata (A, B. RIO. Z3)

The survey gear is a Hamon grab which collects a sediment core of 0.25m2 (0.5 m 0.5 m) on a 0.2 m depth at the ebb tide. We assume that this measuring tool permits an optimal detection (the grab keep sedi- ment deeper than the maximal deepness of clam burring). Therefore, it allows detecting clams distribution in the bay and its space-time variation without sampling bias.

Data interpolation

The point data obtained during previous monitoring campaigns were used to recreate two maps for each survey year (one of abundance and other of biomass) for the whole bay. Geostatistics along with the statistical method of kriging, which permits guessing values of non-sampled points from the known value of their sampled neighbours were applied to

the existing database rich of information on spatial distribution of abundance (expressed in number per 0.25 m^2) and biomass (in gram per 0.25 m^2) with the following steps :

1) variography analysis and auto-adjustment of a model to the experimental variogram (nugget effect and isotropic exponential model) ;
2) geostatistical interpolation of data using the variogram model. The implemented interpolation method used was a block kriging model with a 200 m sliding neighbourhood.

These stages were carried out with R Software (Team 2014), using the RGeostats package developed by the école des Mines of Paris (http://rgeos.free.fr/). Fig. 28 summarizes these different steps.

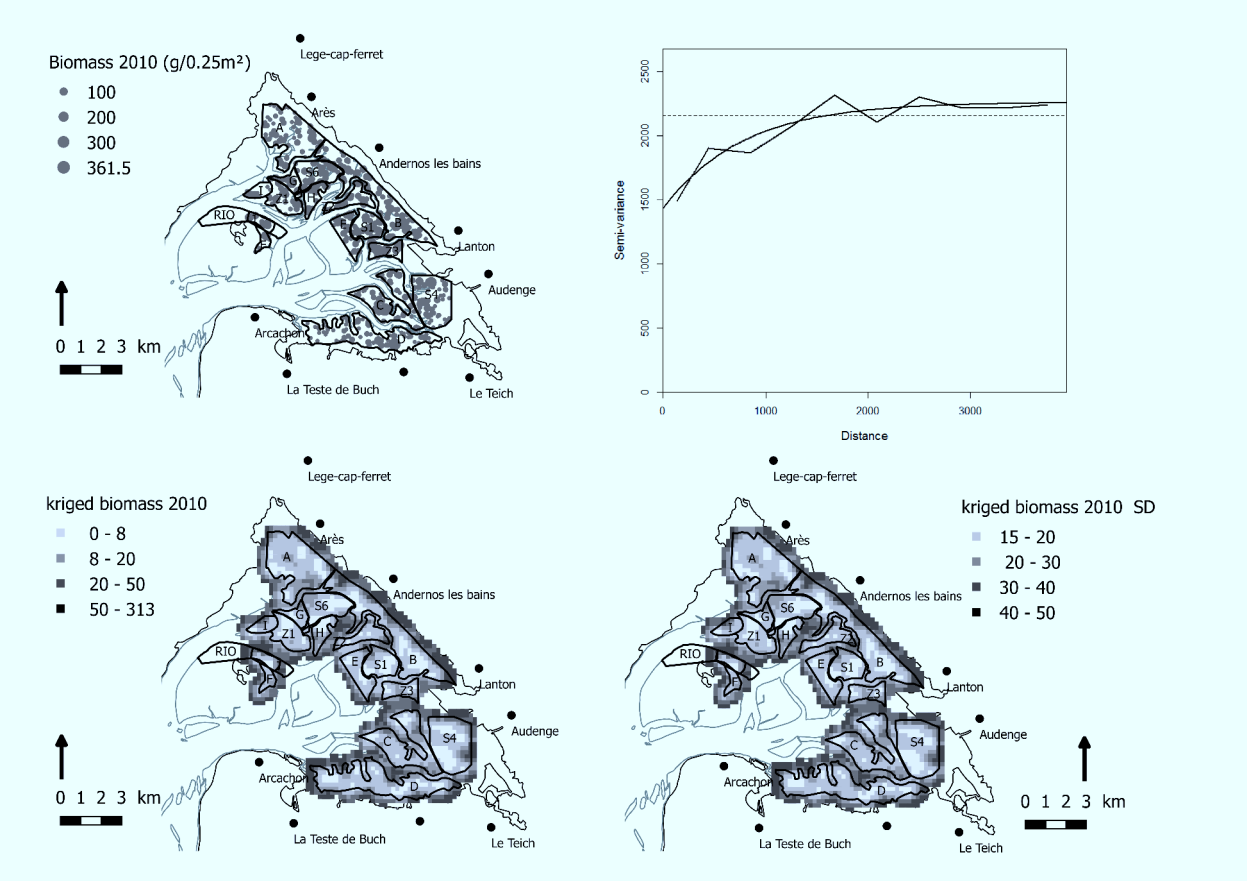


Fig. 28 : Example of kriging procedure for biomass in 2010. Top-left - Survey samples positions and results for biomass, top-right - experimental semi-variogram (non-smoothed line) and modelled associated semi-variogram (smoothed line), bottom-left - kriging results and bottom-right - standard deviation associated to the kriging method.

Spatial distribution of abundance were interpolated from the data collected during the six survey campaigns (2003, 2006, 2008, 2010, 2012 and 2014) for both of the two estimates (biomass and abundance). These populations are herein called “semi-virtual populations” in the sense they represent the real phenomenon with an associated standard deviation and are used to compare the efficiency of sampling designs.

The sampling designs to be compared : StRS vs GRTS vs BAS

The basic survey design used in a 6 years survey in Arcachon Bay (stratified simple random sampling - StRS) is compared to two spatially balanced survey designs : Generalized Random Tessellation Sampling (GRTS) known for its high efficiency, and the recent one Balanced Acceptance Sampling (BAS) whom efficiency is not fully assessed yet.

SRS

It is common ground that probability based survey designs are not enough used in ecology studies (Smith, Anderson, et Pawley 2017), but when used, simple random sampling (SRS) is the most common of them, certainly due to its ease of use and its flexibility. Additional samples can be easily added to the survey design a posteriori. The main disadvantage of SRS is that, sometimes, it exists cluster of samples or areas devoid of samples in the survey scheme (Stevens Jr et Olsen 2004 ; Christianson et Kaufman 2016). A recent study (Christianson et Kaufman 2016) indeed underlines that random sampling can have particular spatial arrangement that over- or under-samples certain regions of the studied area and leads to larger survey error. This design can fail to detect spatial patterns so it can be inefficient for patchily distributed resource study (Yu et al. 2012). A variant of SRS is the spatial stratified SRS technique (called stratified random sampling - StRS) which involves dividing the study area in strata and then randomly sample within each stratum. For a better performance, strata must be created in order to be relatively homogeneous among themselves (Yoccoz, Nichols, et Boulinier 2001) according to the studied phenomenon (Zhao et al. 2016). The relatively new approach of spatially balanced survey designs appears to often create more efficient, flexible and rigorous monitoring designs (Theobald et al. 2007). Some studies demonstrate that using a spatially balanced design can be an advantage when the studied variable has spatial trend (Stevens Jr et Olsen 2004 ; Theobald et al. 2007 ; Anton Grafström, Lundström, et Schelin 2012 ; Anton Grafström 2012 ; Anton Grafström et Lundström 2013 ; Robertson et al. 2013) but none have tested it yet on spatially structured populations with temporal variations of lagoons areas. Hence, for this study, we choose GRTS and BAS designs as examples of spatially balanced survey designs.

GRTS

GRTS survey design is an unified survey approach adapted to the environmental monitoring at large scale and over long term (Stevens et Olsen 1999, 2003, 2004). It uses a spatially balanced algorithm instead of a simple random one. This spatially balanced algorithm orders area statistical units in a line and then runs a reverse sequence function before selecting a sample with a systematic design. When the selected samples are reassigned to their original place in the area, they are spatially balanced. This leads GRTS to resemble in some ways the systematic survey strategy, despite actually surpassing it. This survey strategy is a true spatially balanced probabilistic design even if the distribution of samples is constrained in space. So, it allows design-based inferences to the entire study area. In addition, it provides the advantages that any point (site) in the target population is not too far from another sampling point, and very few sample points are found to be close. It allows freeing from one of the main disadvantage of simple random survey. As with SRS (and all others designs), the main area can be divided in homogeneous strata to be individually sampled. GRTS survey

was used in several studies, for example to determine bull trout population status through counts in Basins of the Columbia River Plateau (Katz et al. 2013), or to develop survey ArcGIS tools through a case study of forest biodiversity survey in Hunan Province(Li, Xu, et Zhou 2012). In this sense, GRTS is very relevant when it comes to draw units (sites) sampling of natural resources in the space because it allows to select spatially balanced samples. It was recently proved that clam monitoring in Arcachon bay can be GRTS considerably improved by GRTS in the basis of one-year data study [kermorvant_optimization_2017-1].

BAS

BAS is a newly developed spatially balanced sampling design (Robertson et al. 2013, 2017). It is an extension of the idea of GRTS where a Halton sequence is used to spread the samples across the study area. Study on BAS shown that it achieves a little better spatial balance than GRTS, but its main advantage is that it is faster to execute (Robertson et al. 2013) on computer. BAS could also be used in an area divided in strata. Very few scientific studies have already used it ; one showed that BAS is superior to SRS in terms of spatial spread and precision on a crab population (Abi et al. 2017). One report already used a two dimensional BAS to survey bats in Bighorn Canyon National Recreation Area (BICA) (Keinath et NRA 2016) and show good results. We choose BAS as an example of spatially balanced sampling because it has the huge extra-advantage to balance the samples in three or more dimensions (representing for example latitude, longitude and altitude). This ability is particularly appropriate to monitor natural resources. A three dimensions survey is relevant for monitoring in water bodies (integrating deepness), and a four or more dimensions study design will be able to integrate information such as ecological threats, time intervals, species population structure, environmental data…(Brown, Robertson, et McDonald 2015). If we found that BAS perform well for manila clam's monitoring program in Arcachon bay with only two dimensions (x and y coordinates), we could expect that it will perform even better with more dimensions.

Defining optimal sample size by sampling design

The first sample size which realizes the targeted accuracy is called "optimal sample size". To remove all possible hazardous bias in results, we use 1000 simulations of sampling efforts per year, per stratum and per estimates (biomass and abundance). It means that, for each year, each stratum and each estimate, 1000 simulations of 1 then 2 then 3. then n stations selection were performed for both sampling designs. Accuracy on the results increases when the sample size increases.

We define then an optimal sample size for the whole bay according to the following steps :

1. For each stratum i and each year j, let note $n_{i,j}^{B}$ and $n_{i,j}^{A}$ the optimal sizes for biomass and abundance respectively. Then keep $n_{i,j}^{opt} = \max(n_{A}^{i,j}, n_{B}^{i,j})$ which represents the selected optimal size to study both the biomass and the abundance ;
2. For each year j, let $n_{j}^{opt} = \sum_{i} n_{i,j}^{opt}$ be the total optimal sample size for this year.

To be constant throughout the years, optimal sample sizes per campaign are compared only for the strata which were sampled all the studied years (thus excluding RIO, I and J strata).

This method allows choosing a level of accuracy that we want to achieve when sampling the semi-virtual populations. In this study, we set the level of accuracy in biomass and abundance estimates to 10% (- see Supplementary materials for 5% and 20% levels results). So here, the optimal number of samples will be the minimal number of samples needed by a sampling design to achieve at least 10% of accuracy in biomass and abundance estimates.

Optimal samples sizes are firstly compared by years to detect temporal variation. Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are calculated with R software for both of the sampling designs. They allow to know if a value at a t time is correlated to the previous value t-1, and involve detecting temporal correlation in our optimal sample sizes. If these functions show that there is no temporal correlation in optimal samples sizes obtained throughout the 6 studied years, this means that the assessed sampling design is able to deal with temporal variation of the clam distribution. In this case we could set a fixed number of samples to do by years for future campaigns to achieve 10% of accuracy in results. Secondly, using a non parametric paired Wilcoxon-Mann-Whitney statistic test, optimal sample sizes are compared between strata to know if there are differences between both sampling designs results at this scale. This information tells if one or other sampling design perform better than the others.

Optimizing the design and assessing the monitoring cost

To optimize the survey design, it is necessary to found the number of samples which has to be done in each stratum during future survey campaigns, a quantity named hereafter "total sample size". We consider two alternatives to give the total sample size (both of BAS, GRTS and StRS) :

- "mean Nopt" : for each stratum, the mean on the six studied years optimal sample sizes are summed to have a total sample size in the whole bay ;
- "max Nopt" : retained total sample size is the maximal optimal sample size obtained for each stratum between the six studied years and sums them.

A total of 1000 simulations are used to assess and compare these two total sample sizes for the virtual populations (one virtual population for each year of data). This gives the accuracy that would have been reached if these total sample size had applied in the field since 2003.

The total survey cost (TSC) is defined and computed using two parameters :

- (V) a survey variable cost dependent on the sample size (the costs of boat and grab rents, scientists and fishermen participation costs, etc.), established on expert knowledge at V= 80 € for this study.
- (F) a survey fixed cost (the costs of meetings to prepare the survey, material, data treatment, meetings to present the results to fishermen and administration, etc.), established on expert knowledge at F = 12 500 € for this study.

The overall cost of the simulated designs is then calculated as follows :

(0) $TSC = F + n\ x\ V$,

where n is the number of samples.

Results

Total optimal sample sizes required to reach an accuracy of 10% per year in the whole bay differed between the assessed designs ; StRS requires almost twice GRTS or BAS optimal sample size for all the studied years (Fig. 29).

Fig. 29 : Total optimal sample size for the studied years in the whole Arcachon Bay for StRS, GRTS and BAS designs to achieve 10 percent of accuracy - Values are given in Supplementary information.

BAS and GRTS designs show comparable results for annual optimal sample size. Some variations are noticeable between years but all of them have a comparable optimal sample size with both designs. Tests of stationarity made on the three time series using autocorrelation and partial autocorrelation functions (see graphs in supplementary materials) indicate that all functions are stationary.

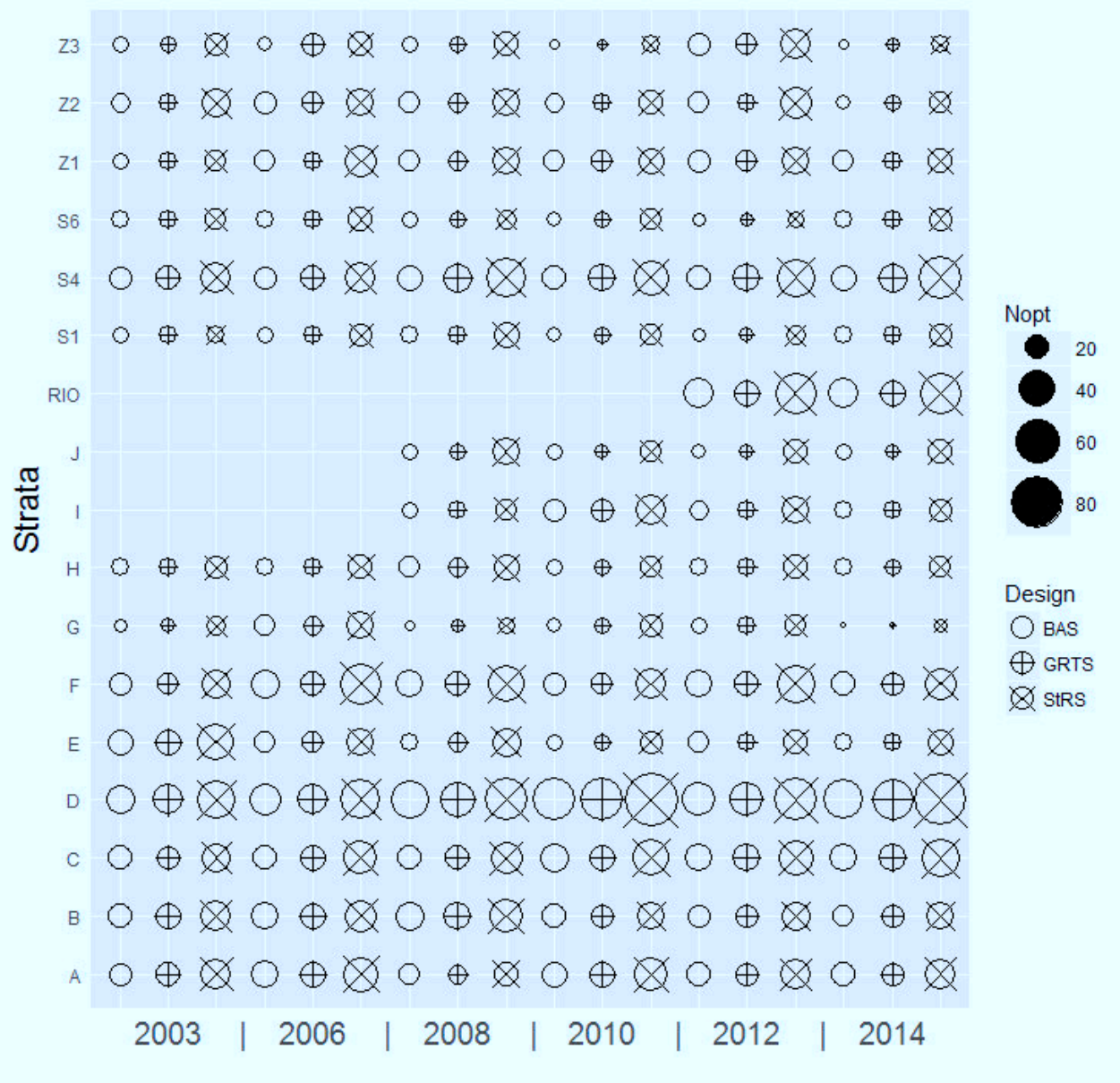


FIG. 30 : Optimal sample size (Nopt) obtained by our methodology for the 17 strata for the 6 surveys. Bubbles sizes are proportional to optimal sample sizes. Empty bubbles show results for BAS design the ones with a '+' inside for GRTS and the ones with an 'x', StRS.

Graphical comparison of optimal samples sizes for all strata is made using bubble plot (Fig. 30) : as StRS bubbles are always bigger than GRTS and BAS ones, StRS always needs higher number of samples than GRTS and BAS to achieve 10% of accuracy in the results at the stratum scale (paired Wilcoxon tests between GRTS and StRS and then between BAS and StRS indicate that StRS is significantly different from GRTS and BAS ; p-values < 2.2e-16). GRTS and BAS results are similar (paired Wilcoxon tests indicate a non-significant difference ; p-value = 0.07909), for some strata GRTS needs fewer samples than BAS and sometimes it is BAS that needs fewer than GRTS. Overall, optimal sample sizes found for

same strata are more or less equal among years.

Table 1 : "mean Nopt" and "max Nopt" total sample sizes with both of the survey designs (BAS, GRTS and StRS) and the approximation of their cost if they would be used on field.

	MeanNopt		Max Nopt	
Design	Total sample size	Overall cost (in €)	Total sample size	Overall cost (in €)
BAS	292	35 860	363	41 540
GRTS	287	35 460	347	40 260
StRS	482	51 060	633	63 140

There are big differences between BAS / GRTS and StRS total sample size and between prices (Table 1). BAS and GRTS total sample sizes and overall cost are widely lower than StRS ones. BAS would cost 30% and GRTS would cost 31% less than StRS if the Mean Nopt" sample size would be used (and respectively 33% and 35% for "max Nopt").

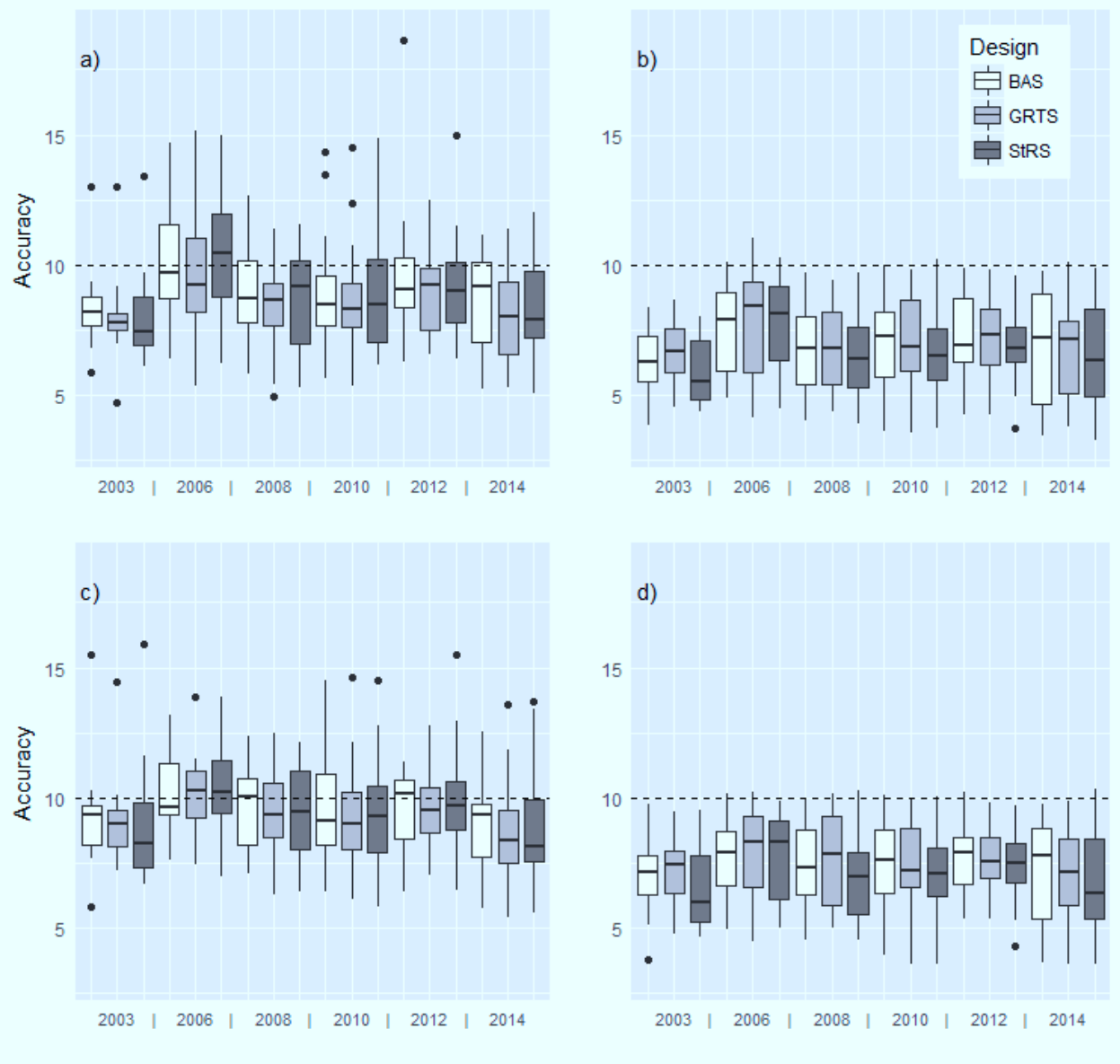


FIG. 31 : Box-plots of reached accuracy for all strata for : a) biomass with the 'mean Nopt' total sample size, b) biomass with the 'max Nopt' total sample size, c) abundance with the 'mean Nopt' total sample size and d) abundance with the 'max Nopt' total sample size - 1000 simulations.

If "max Nopt" and "mean Nopt" sample sizes are used, almost all the median are under the targeted level of 10% of accuracy (fig. 31). With "max Nopt" median level of accuracy are far under this targeted level. Both of these sample sizes are associated with few inter-annual variations in accuracy results.

Discussion :

There are two critical aspects to take into account when surveying an animal population : spatial variation in animal abundance and species detectability (MacKenzie 2006 ; Royle et Dorazio 2008). Clams do not have the capacity to escape during the survey and all the sediment of all the samples had been sifted : hence, one can assume that detectability denoted "p" is almost maximal (p = 1). Thus, it can be considered that our data only account for clam spatial variation and not for survey method non-detection. The multi-seasonal database allows to also considering clam's patches temporal variation.

This simulation study convincingly demonstrates that GRTS and BAS perform nearly twice better than StRS for Arcachon Bay manila clam long term monitoring, confirming and extending thus the results published for the GRTS-only design and on the basis of one-year study (Kermorvant et al. 2017). For all the studied years, GRTS and BAS always need fewer sample points to achieve the same accuracy than StRS. This means that, years after years, these spatially sampling designs perform better in presence on spatial variation of clam's estimates and so are more cost-effective than simple random sampling. There is statistically no difference between using GRTS or BAS ; sometimes GRTS needs fewer samples and sometimes it is BAS. Otherwise, we have shown that GRTS, BAS and StRS achieve good performances with few inter-annual variations in accuracy for this population. This indicates that temporal variation of clam's abundance and biomass does not have any noticeable effect on the performance of these sampling designs. This characteristic allows us providing "total sample sizes" which integrates both spatial and temporal heterogeneity of manila clam population. In terms of field conveniences, it indicates that we can set a same number of samples by strata for all the future campaigns and reach the targeted accuracy in population estimates.

Three accuracy levels (i.e. 5%, 10% and 20%) were assessed in our simulation study but only the 10% level is presented in this paper to avoid overloading figures. Results for 5% and 20% are similar and the patterns are consistent with this 10% level (see Supplementary Information). The ability of GRTS and BAS to provide a spatially balanced scheme certainly has a role in their capacity to deal correctly with the spatially structured population of manila clams. Indeed, Christianson et Kaufman (2016) showed that sample spacing is a key factor for a design which has to bring a good estimation in presence of spatial variation. Also, the efficiency of a survey strategy is highly related to its space-filling characteristics (Rajabi et Ataie-Ashtiani 2014). A recent study, although limited to comparing two designs and not incorporating advanced survey techniques, highlighted that systematic survey designs could be superior than simple random design for selecting samples in clustered populations (McGarvey, Burch, et Matthews 2016). Several other works show also that there is advantage in using a spatially balanced design when the response has spatial trend too (Stevens Jr et Olsen 2004 ; Theobald et al. 2007 ; Anton Grafström, Lundström, et Schelin 2012 ; Anton Grafström 2012 ; Anton Grafström et Lundström 2013 ; Robertson et al. 2013). Results of our study further highlight the importance of taking into account spatial and temporal variations in the choice of sampling design for the optimisation of monitoring programs of spatial and temporal structured populations.

In a practical point of view, if StRS is still used during the future campaign in Arcachon bay with our sample size "mean Nopt", the price of the field survey will be the same than before ($\approx$ 50 000 €) ; but with 10% of accuracy reached in both strata instead of an accuracy highly variable between both of them. The use of GRTS or BAS could advantageously reduce the price of future surveys of approximately 30 % (dropping the total cost to $\approx$ 14 000 €). The price depends also on the used sample size : "mean Nopt" costs 5 000 € less for BAS and GRTS than "max Nopt".

The main goal of our study was to prove that local monitoring programs could be optimised. For this, we had to found the more performing sampling design for manila clam's monitoring, and to propose a sample size which can be applied in all the futures field surveys of clam monitoring. Even if using "max Nopt" ensures an accuracy of the results far lower than 10%, we suggest using instead "mean Nopt" with GRTS or BAS. The main rationale behind this proposal lies in the fact that i/ "mean Nopt" is sufficient to achieve an accuracy of nearly 10% for all strata, and ii/ it is less expensive than "max Nopt". Whilst we succeed at demonstrating that local monitoring programs could be optimised, this study remains a simulation study where the percentage of accuracy obtained is the mean of 1000 simulations (with an associated variance). In the field, it will obviously not be possible to lead 1000 replicates of the chosen sample size, and consequently (due to the random property of sampling designs) the real accuracy might not be exactly 10%.

We know that accuracy will not be exactly 10% for all the strata and all the years but this optimisation solution appears as the most logical trade-off between lowering the campaign costs and keeping a good accuracy in the results. Nonetheless, to stick to the reality of true needs, these results will have to be explained to professional clam fishermen. Even if they are already aware of this methodological reflexion, it will be mandatory to detail the pros and cons and make sure that they are convinced of the reliability of the survey if we make changes. For a successful collaboration, it is indeed essential that they understand and approve these methodological improvements. Thanks to GRTS or BAS, campaign costs can be lowered under 50 000€ keeping a "good" accuracy in clams abundance and biomass, allowing for a continuous assessment of manila clam stock in Arcachon bay, and avoiding thus any other problematic break in the temporal series of this data set.

Nonetheless, despite great potential of GRTS in spatially balanced surveys of natural resources, it is restricted to a two-dimensional space and has no obvious extension to a higher dimension space. Theobald et al. (2007) and Anton Grafström (2012) proposed several refinement or variation on GRTS, respectively called reversed randomized quadrant-recursive raster (RRQRR) theoretically possible in three-dimensional space, and spatially correlated Poisson survey (SCPS). BAS was also developed to allow a spatial balancing of samples in a multidimensional scale. Our study is the first ever in which BAS was used to optimise a monitoring design, demonstrating also that it performs as well as GRTS. Its promising results in two dimensions and its ability to perform spatial balance in more than two dimensions let us think that BAS will be very more relevant than GRTS for future surveys optimisation.

Conclusion

This study highlights that spatially balanced designs performs better than the simple random sampling to deal with the patchily distribution of manila clam population in Arcachon bay. But, we have shown that they also behave appropriately with the temporal variation of manila clam spatial structure. We argue that it is now possible to define a sample size by strata that could be done for all future campaigns to achieve a given level of accuracy in the results. Results show that, even if spatially balanced designs are more cost effective, trade-off between achieving accuracy and survey price are always to be considered. We convincingly emphasize a positive effect of adopting such advanced spatially balanced design in avoiding any potential temporal breaks following money short-coming (such as the one occurring during 2014 when no manila clam monitoring campaign in Arcachon bay could be done because of fund lack). In the past, one field campaign used to cost at least 50 000 € and given strata size-dependent sampling effort, accuracy on the results proved to be different between strata. Now, benefiting from the outputs of our study, we advocate that changing the sampling design and using a non strata size-dependent sampling effort, the campaign should potentially cheaper, with a gain of 14 000 € (i.e. - 30%) if a mean of 10 % of accuracy in each strata is to be achieved. This would undeniably lead to provide a rigorous dynamic image of the clam population in the whole bay and help thus fishermen to define good sustainable management for their resource, on a continuous basis.

With this work, we hope demonstrating efficiently that not only long term monitoring programs can be optimised but this change in sampling procedure is easily transferable to be reproducible in other similar contexts where resources have to be surveyed on a continuous dynamic basis. If using spatially balanced sampling design can reduce by 30% the price of Arcachon bay manila clams monitoring, using it elsewhere in other similar socio-economic context should also profitably reduce the long term monitoring costs. To extend the benefits of this study to a larger scale (European for example) the concept of master sample (van Dam-Bates, Gansell, et Robertson 2018) could be assessed.

Eventually, to be totally generic, it remains necessary to test other spatially balanced sampling designs such as local pivotal method (Anton Grafström 2012). The main reason why we choose to assess performances of BAS, and why we believe in its potential for the future, is because it permits to extend the spatial balancing of the samples to more than two dimensions. BAS performance was never tested to optimise a monitoring programs, and now that we know that BAS is relevant for this ones in two dimensions (x and y coordinates), we have to found one (or more) explicative variable(s) (for example sedimentology or water circulation in the bay) of manila clam dispersion in Arcachon Bay and use it as a third (or more) dimension. This could lead to an even more cost-effective sampling design.

Acknowledgments

Data were collected in the context of the Manila clam monitoring program financed by the French institute Ifremer, European grants (IFOP - instrument Financier d'Orientation de la Pêche ; and FEP - Fonds Européens pour la Pêche), the French government (MEEM - Ministère de l'Environnement, de l'Energie et de la Mer), Aquitaine Regional Council,

Gironde General Council and professional fishing organizations. This work was supported by Ifremer's scientific direction within the framework of site policy grants (DESCARTES 2 project) aiming to reinforce collaborations between Ifremer and regional academic partners. This work was also supported by "Communauté d'Agglomération Pays Basque - Euskal Hirigune Elkargoa" through a thesis grant.

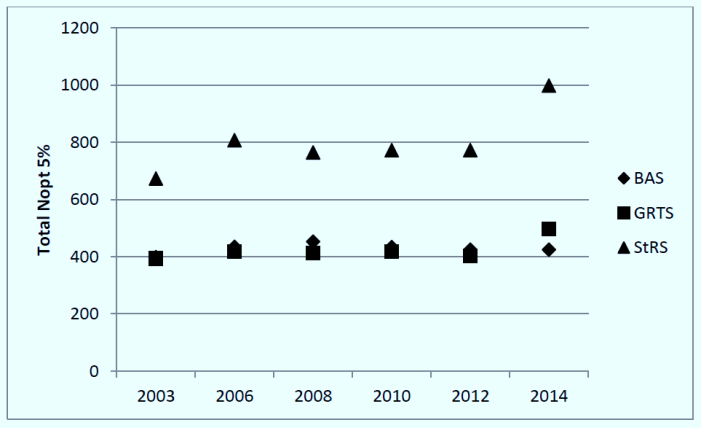
Supplementary informations

10%	2014			2012			2010			2008			2006			2003		
Strata	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS
A	19	18	30	19	18	29	21	22	34	16	14	20	23	22	37	17	19	30
B	16	17	23	18	17	25	19	18	23	26	24	35	24	22	31	20	21	29
C	24	24	44	23	23	36	25	21	40	20	20	31	19	21	34	20	18	27
D	49	50	84	35	34	57	56	55	88	43	38	56	32	29	47	25	30	43
E	12	12	21	15	13	19	11	11	17	13	14	26	16	15	23	22	22	40
F	20	17	32	24	20	44	17	15	30	24	20	41	26	20	52	18	15	26
G	7	7	9	11	12	16	10	11	18	8	9	11	15	14	23	9	10	14
H	12	11	16	13	13	20	11	11	16	15	14	22	12	12	20	12	12	17
I	12	12	17	14	13	23	17	17	29	11	12	16						
J	11	10	20	10	10	19	11	10	16	11	11	23						
RIO	28	22	52	28	22	52												
S1	12	12	17	9	10	14	10	11	16	12	12	21	11	12	17	11	12	13
S4	22	26	56	20	24	43	20	23	39	21	27	47	17	20	29	17	19	30
S6	12	13	17	9	9	11	10	11	16	11	11	14	12	13	19	12	12	16
Z1	16	13	19	18	15	24	16	16	22	15	14	24	16	13	31	11	13	16
Z2	10	11	16	15	13	31	14	12	19	16	14	23	18	16	24	14	13	26
Z3	8	9	12	17	15	30	8	8	12	11	11	20	10	18	19	11	11	17
TOTAL	290	284	485	298	281	493	276	272	435	273	265	430	251	247	406	219	227	344
TOTAL (- I, J and RIO)	239	240	396	246	236	399	248	245	390	251	242	391	251	247	406	219	227	344

Supplementary information 1 : Optimal samples sizes obtained for a level of 10% of accuracy for both strata all the years

5%	2014			2012			2010			2008			2006			2003		
Strata	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS
A	36	28	60	33	28	53	37	35	67	27	24	38	37	38	73	33	32	57
B	32	29	45	32	30	49	34	33	46	47	41	65	44	37	62	36	37	58
C	41	39	106	38	39	71	40	37	74	33	35	62	32	37	71	33	31	53
D	82	86	143	61	62	115	97	98	180	79	66	108	56	51	97	47	54	86
E	21	18	35	26	24	36	19	19	33	24	25	55	25	26	45	40	36	79
F	35	29	53	41	35	84	32	26	61	44	34	79	46	32	103	31	25	51
G	11	77	174	20	20	31	19	17	34	15	15	21	26	24	46	16	16	28
H	22	21	44	24	23	39	18	17	32	28	25	43	20	18	39	22	22	32
I	18	39	73	24	22	47	31	32	58	20	18	33						
J	19	26	44	18	17	35	18	17	28	18	19	47						
RIO	48	22	35	48	38	100												
S1	22	44	73	14	17	27	18	37	77	21	20	41	20	21	33	20	22	27
S4	40	38	102	35	37	84	34	18	31	41	42	94	29	33	56	33	32	58
S6	22	22	39	16	17	22	18	26	45	20	20	27	18	21	36	21	21	31
Z1	29	17	26	31	25	48	29	21	39	25	23	48	27	21	64	21	21	32
Z2	18	27	64	25	23	58	24	15	23	30	25	47	36	27	46	25	23	52
Z3	14	28	60	29	24	56	15	35	67	19	19	37	19	33	37	21	22	30
TOTAL	510	562	1116	515	481	955	483	467	859	491	451	845	435	419	808	399	394	674
TOTAL (- I, J and RIO)	425	497	999	425	404	773	434	418	773	453	414	765	435	419	808	399	394	674

Supplementary information 2 : Optimal samples sizes obtained for a level of 5% of accuracy for both strata all the years



Supplementary information 3 : Plot of total optimal sample sizes obtained by sampling design and by years for 5% of accuracy

20%	2014			2012			2010			2008			2006			2003		
Strata	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS	BAS	GRTS	StRS
A	12	10	10	11	11	14	12	10	16	9	8	10	14	13	19	11	11	15
B	10	9	6	10	10	13	10	13	19	13	13	17	13	13	16	10	11	15
C	14	13	28	13	13	19	14	30	43	11	11	15	11	12	19	11	11	14
D	27	27	8	21	21	30	32	6	43	23	21	28	17	18	24	15	16	23
E	7	7	8	8	7	9	6	9	15	8	9	14	9	8	11	12	12	20
F	11	10	16	13	12	22	10	13	17	14	12	21	14	12	27	9	9	13
G	5	24	42	6	7	8	7	7	9	5	5	6	8	8	12	6	6	7
H	7	7	11	8	8	10	6	6	8	8	8	11	7	7	10	7	6	9
I	7	13	22	8	8	12	10	10	14	7	6	9						
J	7	8	12	6	6	9	6	6	8	7	7	12						
RIO	16	13	27	16	13	26												
S1	7	6	9	5	6	7	7	7	8	7	8	11	7	7	9	6	6	7
S4	13	16	8	11	14	21	12	13	20	14	17	23	10	12	14	11	11	15
S6	8	13	27	5	5	6	6	6	8	6	6	7	7	8	10	7	7	8
Z1	9	7	10	9	8	12	9	9	12	9	8	12	10	10	15	7	7	8
Z2	6	6	5	9	9	15	8	7	10	9	8	12	9	10	12	8	7	13
Z3	5	9	17	11	9	15	5	5	6	7	8	10	6	10	10	6	7	14
TOTAL	171	198	266	170	156	234	160	157	256	157	155	218	142	148	208	142	127	181
TOTAL (- I, J and RIO)	141	164	205	140	140	201	144	141	234	143	142	197	142	148	208	142	127	181

Supplementary information 4 : Optimal samples sizes obtained for a level of 20% of accuracy for both strata all the years

Supplementary information 5 : Plot of total optimal sample sizes obtained by sampling design and by years for 20% of accuracy

Résultats issus de ce chapitre :

- Le manque de financements est un problème majeur en écologie, en particulier à l'échelle locale. Il en résulte un compromis entre le coût et l'efficacité des suivis environnementaux.
- De nouveaux protocoles d'échantillonnage avancés, le GRTS et le BAS, ont la spécificité d'être spatialement équilibrés. Le GRTS est déjà connu pour ses performances. Jusqu'ici le BAS n'avait jamais été testé sur une population réelle.
- La méthode développée dans cette thèse est ici utilisée sur les résultats issus de toutes les campagnes d'échantillonnage de la palourde dans le bassin d'Arcachon (6 campagnes) pour optimiser ce suivi dans le temps et dans l'espace.
- Les résultats confirment que les deux protocoles spatialement équilibrés (GRTS, BAS) ont une meilleure performance que l'aléatoire simple utilisé en routine pour ce suivi. Ils permettront certainement de baisser le prix de la campagne, pour une même précision finale dans les estimations d'abondance et d'effectifs. Entre eux, ils ont une performance comparable.
- Le nombre d'unités statistiques nécessaires aux protocoles pour atteindre une précision fixée dans les résultats est constant au cours du temps.
- La méthode permet une optimisation spatiale et temporelle de ce suivi.

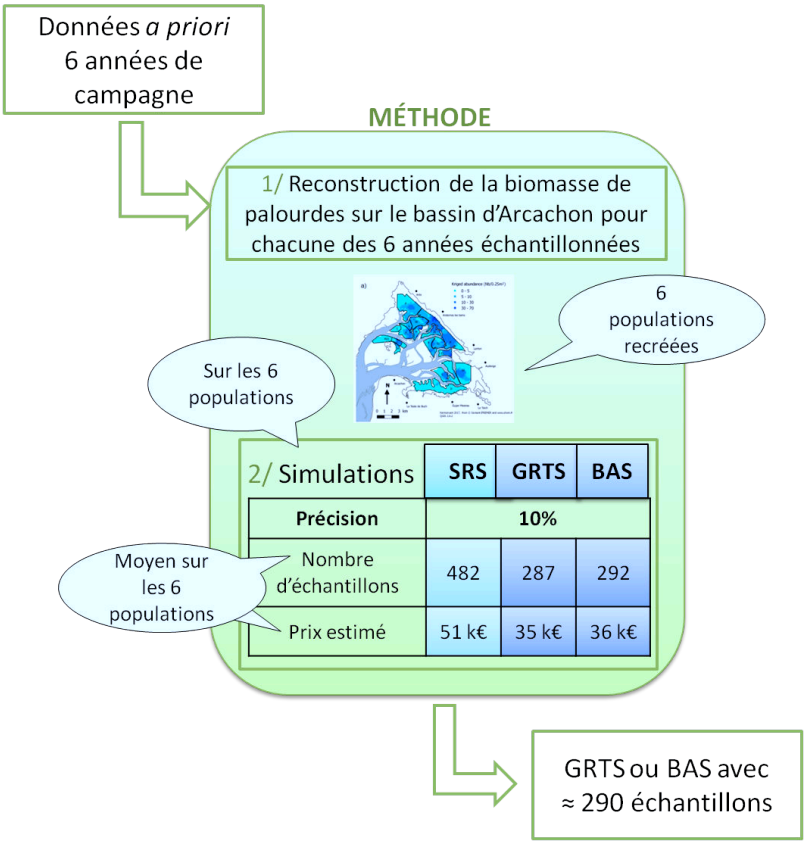


FIG. 32 : Conclusions du CHAPITRE V

DISCUSSION GENERALE

Rappel du contexte

Dans un contexte où le coût financier est bien souvent le facteur limitant au déploiement des suivis environnementaux, les environnementalistes ont l'habitude de favoriser un placement des unités statistiques moins couteux mais basé sur du jugement, du pseudo-aléatoire ou de l'aveuglette. Malheureusement, ce type de procédures n'utilisant pas de processus de sélection aléatoire des unités statistiques ne permettent pas de certifier que les résultats seront précis et sans biais. Ils peuvent même mener à des résultats complètement erronés et donc des suivis inutiles.

En plus de certifier des résultats corrects, préparer une procédure d'échantillonnage en amont de la phase de terrain permet, par exemple, de savoir où et dans quel ordre vont être récoltées les unités statistiques mais aussi quelle précision dans les résultats peut être atteinte avec le budget alloué au suivi. Ces deux dernières décennies, plusieurs outils probabilistiques de sélection des unités statistiques ont étés développés. Ils sont tous plus ou moins efficaces en fonction de la population étudiée. Plus un protocole est efficace, moins il nécessite d'unités statistiques dans l'échantillon pour atteindre des résultats précis, et donc *in extenso*, moins il est coûteux à mettre en place. Autrement dit, si l'utilisateur a un nombre fixe d'unités statistiques à effectuer sur le terrain, utiliser un protocole plus efficace que les autres lui permettra d'atteindre des résultats plus précis avec ce nombre d'unités statistiques dans l'échantillon. Lors de la planification de l'échantillonnage, il est important, voire indispensable, de bien choisir le protocole d'échantillonnage ainsi que le nombre d'unités statistiques qui vont être réalisés sur le terrain. Mais peu d'outils de planification sont disponibles jusqu'à présent.

L'objectif de cette thèse était de développer, puis de tester sur des exemples concrets, une méthode générale qui permettra de choisir un protocole d'échantillonnage performant couplé à un nombre d'unités statistiques à échantillonner adapté. En plus de garantir des résultats précis et sans biais, la procédure retenue permettra de réduire le coût du suivi.

La discussion générale s'articulera autour de 3 axes :

- Rappel des principaux résultats,
- Discussion sur la méthode,
- Perspectives.

Principaux résultats

Les protocoles d'échantillonnage probabilistes sont les outils qui permettent de définir l'ordre et l'emplacement des unités statistiques à récolter dans la population que l'on cherche à étudier. La composante aléatoire qui les caractérise garantit la fiabilité et la robustesse des données. Dans cette grande famille que sont les protocoles d'échantillonnage probabilistes, certains incluent aussi une composante spatiale. Ces derniers, appelés protocoles d'échantillonnage probabilistes spatialement équilibrés (SBS), ont la capacité d'être particulièrement performants lorsque la variable que l'on cherche à estimer par l'échantillonnage présente une tendance spatiale.

Le premier chapitre de cette thèse a eu pour principal résultat la mise en lumière de ces

protocoles innovants. Un objectif sous-jacent étant d'essayer d'enrayer, ou au moins de limiter, l'utilisation des protocoles non-probabilistes. Les scientifiques de l'environnement ont beaucoup à gagner à utiliser les SBS, tant au niveau du coût total de leurs suivis qu'au niveau de la précision de leurs estimations. Une autre particularité intéressante des SBS est que certains d'entre eux peuvent, en plus d'équilibrer les unités statistiques dans l'espace, l'équilibrer dans d'autres dimensions comme le temps ou la profondeur, ou bien dans des dimensions plus abstraites comme de la couverture des sols ou de la variance plus ou moins hétérogène. C'est cette particularité qui fait que les SBS seront certainement des protocoles d'échantillonnage de plus en plus utilisés dans les années à venir. Un autre résultat majeur du premier chapitre réside dans l'édition des scripts de l'utilisation des SBS sous le logiciel gratuit R. Ils serviront à la diffusion de ces protocoles avancés en facilitant leur prise en main pour les utilisateurs qui souhaiteraient mettre en place un suivi.

Néanmoins, il existe plusieurs protocoles d'échantillonnage probabilistes mais aucun ne peut être qualifié de "meilleur que les autres" pour tous les cas d'études. Pourtant, la personne en charge de la planification de l'échantillonnage doit en choisir un. La méthodologie générale présentée au chapitre 2 permet de choisir entre différents protocoles d'échantillonnage probabilistes en déterminant le plus performant entre eux pour chaque population testée. Puisque la performance est l'aptitude d'un protocole d'échantillonnage à fournir des résultats précis d'estimations avec le moins d'unités statistiques possibles ; un nombre d'unités statistiques à échantillonner est aussi renvoyé par notre méthode. Les trois derniers chapitres de ce rapport de thèse s'appliquent à démontrer que la méthode peut être utilisée pour mettre en place un suivi efficace, pour optimiser spatialement un suivi lorsque des données récoltées une saison par un protocole probabiliste sont déjà disponibles et pour optimiser spatialement et temporellement un suivi lorsque des données récoltées avec un protocole probabiliste sont disponibles sur plusieurs saisons.

Les résultats principaux issus de ce manuscrit sont :

- La description et l'aide à l'utilisation de protocoles d'échantillonnage avancés.
- Le développement d'une méthode statistique permettant la mise en place de suivis efficaces ainsi que l'optimisation de suivis déjà en place.
- L'illustration de l'efficacité de cette méthode à travers trois exemples de suivis environnementaux ; lors d'une mise en place, d'une optimisation à partir d'une seule saison de données et d'une optimisation à partir de plusieurs saisons de données.

Un résultat commun aux trois exemples est que les protocoles d'échantillonnage spatialement équilibrés (SBS) sont toujours autant performants, voire beaucoup plus performants, que le protocole en aléatoire simple (SRS). Ils ont permis de proposer un suivi efficace pour le moustique tigre *Aedes albopictus* sur l'agglomération Bayonne-Anglet-Biarritz et d'optimiser le suivi de la palourde dans le bassin d'Arcachon. Dans ce dernier exemple, le nombre d'unités statistiques nécessaires pour réaliser des estimations de population assez précises pour permettre une gestion durable de la ressource peut être abaissé en utilisant un SBS au lieu d'un SRS. Cela a un impact direct le coût total du suivi qui, de ce fait, est réduit de 30%.

Discussion sur la méthode

Dans le domaine des sciences environnementales, deux grandes théories pour l'échantillonnage sont souvent opposées : le *design-based* et le *model-based*. Nous avons développé une méthode qui utilise conjointement ces deux théories. Le *model-based* est employé pour reproduire le plus fidèlement possible la population statistique d'intérêt. Une fois cette population reconstruite, il est aisé d'y tester différentes procédures d'échantillonnage avec la méthode *design-based*. L'écologie virtuelle est un outil de plus en plus appliqué pour l'évaluation de méthodes nouvellement développées (méthodes d'échantillonnage, d'analyses statistiques ou d'outils pour la modélisation). Elle est intuitive et puissante puisque toutes les variables sont contrôlables par l'expérimentateur (Zurell et al. 2010). Dans le contexte actuel du fort coût financier et logistique des suivis de populations écologiques, la possibilité de pouvoir tester des procédures d'échantillonnage sur le terrain est quasiment inenvisageable. Les simulations sont donc une bonne alternative pour tester puis évaluer les protocoles d'échantillonnage (Conn et al. 2016), avant de les appliquer sur le terrain (Byers et al. 2002 ; Albert et al. 2010). Un grand nombre de combinaisons protocole/nombre d'unités statistiques/précision atteinte dans l'estimation finale peut être testé à faible coût.

La première partie de la procédure séquentielle développée, utilisant le *model-based*, doit être réalisée avec précaution. Même s'ils permettent de prendre en compte des données récoltées avec un protocole probabiliste, des données récoltées sans protocole, ou même fonctionnant sans données (modèles bayésiens), les modèles ont tous des conditions d'application bien particulières. Si ces conditions ne sont pas respectées, alors les résultats issus de ces modèles – soit la population statistique sur laquelle les simulations sont basées – peuvent être biaisés et donc ne pas représenter de façon fiable la population d'intérêt. Dans notre cas, cela peut avoir une incidence non négligeable sur les résultats issus de la deuxième partie de la méthode. Si la population statistique est représentative de la population naturelle alors les résultats de la procédure d'échantillonnage testée imiteront parfaitement les résultats qui auraient été obtenus si cette procédure avait réellement été réalisée sur le terrain. Plus la population statistique est différente de la population naturelle, plus les résultats de simulation seront biaisés par rapport à la réalité de terrain.

Pour le cas de l'optimisation du suivi de la palourde dans le bassin d'Arcachon, la variabilité à petite échelle a été omise dans la première partie de cette méthode. En utilisant la méthode géostatistique du krigeage ordinaire, une seule valeur de biomasse ou d'abondance est estimée par unité statistique de 200m^2 sur l'ensemble de la population statistique. Cela a très certainement pour effet de lisser la variabilité à petite échelle (i.e. <200m^2). Or, le nombre d'unités à prélever est très dépendant de la variabilité de la variable étudiée. Cette variabilité à petite échelle étant, d'après les dires d'experts, très importante pour la palourde, il serait opportun de la prendre en compte dans un futur travail, avant de réellement baisser le nombre d'unités d'échantillonnage pour ce suivi.

Dans la seconde partie de la procédure, des processus d'échantillonnage *design-based* sont simulés sur la population statistique recréée par modélisation. Un grand nombre de simulations (>1000) sont faites pour chaque effort d'échantillonnage (nombre d'unitées statistiques sélectionnées) avec chaque protocole. Ceci car les résultats issus du *design-based* permettent

d'estimer des paramètres non-biaisés, en moyenne. Donc, lorsque la moyenne des estimations calculées par l'ensemble des simulations est assez précise (critère de sélection du nombre d'unités statistiques de notre méthode), cela ne signifie pas que toutes les valeurs d'estimations le sont. Lorsque l'on applique cela sur le terrain, et de part le caractère aléatoire des protocoles probabilistes, il existe une probabilité infime (mais non-nulle) que l'échantillon sélectionné renvoie des résultats moins précis que désiré. Cet effet indésirable est inhérent à la méthode *design-based*, nous ne pouvons qu'essayer de le diminuer en augmentant le nombre d'unités statistiques dans l'échantillon.

Le travail effectué pendant cette thèse se situe entre l'écologie et les statistiques, l'utilisation de la méthode nécessite des compétences dans les deux domaines. De bonnes connaissances de la population, de son écologie, de son habitat, de sa dispersion permettent de sélectionner des variables exogènes pertinentes pour la modélisation de sa distribution spatiale. Mais, pour créer le modèle et l'utiliser pour de la prédiction, il est indispensable d'avoir des compétences statistiques. Cela est également le cas pour la simulation des procédures d'échantillonnage puis le calcul des estimateurs. La nécessité d'être compétent en écologie et en statistiques peut être un frein à l'utilisation de cette méthode par des non spécialistes.

Perspectives

Des perspectives d'amélioration ou d'adaptation de la méthode peuvent être évoquées. Nous avons voulu cette méthode très générale pour qu'elle puisse être adaptable et adaptée à la pluspart des cas d'étude. La palourde était une espèce relativement confortable pour tester l'optimisation des suivis environnementaux par la méthode. La probabilité de détection de ce bivalve est quasiment égale à 1 avec la méthode d'échantillonnage utilisée (benne Hamon). Mais il est possible d'adapter la méthode à une espèce dont la probabilité de détection serait inférieure à 1, c'est-à-dire une espèce pour laquelle l'opérateur terrain ne détecterait pas automatiquement l'espèce alors qu'elle est effectivement présente. Pour l'adapter dans le cas d'un suivi de présence/absence, il suffirait de rajouter une loi de Bernouilli – détecte, ne détecte pas – lors du tirage des unités statistiques. Chaque unité statistique de la population contenant une valeur « présence » aurait donc une probabilité p (égale à la probabilité de détection de l'espèce) de garder sa valeur « présence » (l'opérateur a détecté l'espèce) et une probabilité $q = 1 - p$ de prendre une valeur « absence » (l'opérateur n'a pas détecté l'espèce alors qu'elle était présente). Autre exemple ; dans le cas d'un suivi d'abondance, pour lequel le dénombrement des individus d'une espèce à probabilité de détection inférieure à 1 est nécessaire ; l'opérateur peut détecter certains individus, mais pas nécessairement tous les individus. Si 10 individus issus d'une espèce dont la probabilité de détection est de 0.8 sont réellement présents sur une unité statistique ; alors seulement 8 individus seront détectés. Il faut donc, pour être le plus représentatif possible de la réalité pendant la phase de simulation, pondérer les valeurs des unités statistiques sélectionnées dans la population statistique par la probabilité de détection. On peut aussi imaginer une phase de revisite des sites qui permettra d'approcher la vraie valeur de l'unité statistique. D'ailleurs, Mckann, Gray, et Thogmartin (2013) expliquent qu'une recette simple pour les procédures d'échantillonnage est d'estimer la probabilité de détection puis ensuite de déterminer le nombre de sites nécessaires à l'obtention d'une probabilité *Psite* (probabilité de détecter l'espèce au moins une fois pendant

les multiples visites sur un site) proche de 90%. Ceci permettra de trouver le nombre d'unités statistiques à échantillonner, voire un nombre de visites, pour atteindre une précision voulue dans l'estimation finale tout en s'affranchissant de la probabilité de détection.

Lorsque deux des trois paramètres que sont le protocole d'échantillonnage, le nombre d'unités statistiques et la précision finale des résultats sont connus, le troisième peut être estimé. Nous avons développé, testé et utilisé la méthode pour définir un nombre d'unités statistiques lorsque la précision finale nécessaire et le protocole d'échantillonnage (puisqu'ils sont testés un par un puis comparés entre eux) étaient connus. Mais une adaptation de la méthode peut être faite si l'utilisateur désire savoir quelle précision il va pouvoir atteindre au maximum avec un nombre donné d'unités statistiques.

D'autres perspectives intéressantes de ce travail sont les perspectives d'utilisation de la méthode. Un de ses atouts est de permettre la détermination du protocole d'échantillonnage et du nombre d'unités statistiques pour atteindre la précision voulue dans l'estimation du paramètre fixé. Et cela en ayant la possibilité d'évaluer le coût financier d'un tel suivi. Actuellement, les scientifiques cherchent la plupart du temps des fonds de financement pour leurs suivis avant d'avoir exatement défini le nombre d'unités statistiques et le protocole qui va être utilisé. Ils ne savent donc pas, au préalable, si les résultats du suivi qu'ils cherchent à mettre en place seront suffisamment précis pour permettre de répondre à la problématique initiale. Par exemple, pour détecter un changement significatif d'abondance moyenne d'une espèce entre deux années, il faut que le test statistique utilisé puisse détecter ce changement et ne renvoie pas un faux-négatif (erreur de type II). Si les valeurs estimées d'abondance moyenne chaque année ont une précision médiocre, le test risque de ne pas détecter un changement entre deux années alors qu'il y en a effectivement eu un. Avec la méthode développée dans ce manuscrit, les scientifiques peuvent déterminer le coût de leur suivi pour que les résultats soient assez précis et permettent de répondre à la problématique initiale, et ce avant de demander les financements pour la phase de terrain.

Une autre perspective d'utilisation de la méthode est l'évaluation de procédures d'échantillonnage déjà en place depuis plus ou moins longtemps. Les données déjà disponibles sur la population étudiée peuvent être utilisées pour reconstruire la population statistique ; puis la procédure d'échantillonnage avec le nombre d'unités statistiques et le protocole (si le suivi est basé sur la théorie du *design-based*) utilisés en routine pourraient être simulés. Cette utilisation peut conforter un utilisateur, gestionnaire, scientifique ou financeur sur la précision atteinte dans les résultats d'estimation, et donc dans la capacité du suivi à répondre à la problématique de base.

On peut aisément imaginer que cette méthode soit utilisée par des commanditaires de suivis, lors de la mise en place de ces derniers. Les résultats issus de la méthode seraient directement implémentés dans le cahier des charges du suivi, pour guider les structures sous-traitant l'échantillonnage ; soit en leur fournissant le protocole à utiliser et le nombre d'unités statistiques à échantillonner, soit directement une liste de points géo-référencés.

CONCLUSIONS

Cette thèse a permis de développer et de tester une méthode pour choisir le protocole d'échantillonnage le plus optimal ainsi que le nombre d'unités statistiques devant être récoltés sur le terrain lors de suivis environnementaux. Ce protocole retenu/choisi et ce nombre d'unités statistiques permettent d'atteindre une précision voulue dans les résultats d'estimation de population pour pouvoir répondre à la problématique du suivi. Cette méthode a été testée pour la mise en place d'un suivi (CHAPITRE III), pour l'optimisation d'un suivi où une seule saison de données est disponible (CHAPITRE IV), puis pour l'optimisation d'un suivi où plusieurs saisons de données sont disponibles (CHAPITRE V). Nous espérons que cette thèse pourra aider les scientifiques et les gestionnaires dans la mise en place de leurs suivis ou de leur l'optimisation. Nous espérons par ailleurs qu'elle contribuera à l'enrayement de l'utilisation de protocoles non-probabilistes, en donnant accès à des protocoles qui garantissent un échantillon sans biais et des résultats non discutables. Nous escomptons également promouvoir les protocoles d'échantillonnage spatialement équilibrés, qui ont montré une haute performance et qui permettront une baisse des coûts des campagnes scientifiques tout en gardant des résultats précis.

REFERENCES

Abi, Naeimeh, Mohammad Moradi, Mohammad Salehi, Jennifer Brown, Jassim A Al-Khayat, et Elena Moltchanova. 2017. « Application of Balanced Acceptance Sampling to an Intertidal Survey ». *Journal of Landscape Ecology* 10 (1) : 96-107.

Abreu, Filipe Vieira Santos de, Maira Moreira Morais, Sérvio Pontes Ribeiro, et Álvaro Eduardo Eiras. 2015. « Influence of Breeding Site Availability on the Oviposition Behaviour of Aedes Aegypti ». *Memórias do Instituto Oswaldo Cruz* 110 (5) : 669-76.

Albert, Cécile H, Nigel G Yoccoz, Thomas C Edwards, Catherine H Graham, Niklaus E Zimmermann, et Wilfried Thuiller. 2010. « Sampling in Ecology and Evolution–Bridging the Gap between Theory and Practice ». *Ecography* 33 (6) : 1028-37.

Armonies, W. 1996. « Changes in Distribution Patterns of 0-Group Bivalves in the Wadden Sea : Byssus-Drifting Releases Juveniles from the Constraints of Hydrography ». *Journal of Sea Research* 35 (4) : 323-34.

Astorga, Marcela P. 2014. « Genetic Considerations for Mollusk Production in Aquaculture : Current State of Knowledge ». *Frontiers in genetics* 5 : 435.

Auby, Isabelle. 1993. « Evolution de La Richesse Biologique Du Bassin d'Arcachon ». Rapport de contrat 15144. France : Ifremer. http://archimer.ifremer.fr/doc/00040/15144/.

Auby, Isabelle, et Pierre-Jean Labourg. 1996. « Seasonal Dynamics of *Zostera Noltii Hornem.* in the Bay of Arcachon (France) ». *Journal of Sea Research* 35 (4) : 269-77.

Bald, J., A. Borja, I. Muxika, J. Franco, et V. Valencia. 2005. « Assessing Reference Conditions and Physico-Chemical Status According to the European Water Framework Directive : A Case-Study from the Basque Country (Northern Spain) ». *Marine Pollution Bulletin* 50 (12) : 1508-22. https://doi.org/10.1016/j.marpolbul.2005.06.019.

Bald, Juan, et Angel Borja. 2005. « Estudio Del Estado de Los Recursos de Almeja y Berberecho En Los Estuarios Mundaka, Plentzia y Txingudi (1998-2004)). *Inf. Tec. Gob. Vas.* » 105 : 76.

Bald, Juan, et Ángel Borja. 2001. « Estudio de Los Recursos de Almeja y Berberecho En Mundaka y Plentzia (1998-2000). *Inf. Tec. Gob. Vas.* » 93 : 80.

Barabesi, L, et S Franceschi. 2011. « Sampling Properties of Spatial Total Estimators under Tessellation Stratified Designs ». *Environmetrics* 22 (3) : 271-78.

Beilhe, Leila Bagny, Stéphane Arnoux, Hélène Delatte, Gilles Lajoie, et Didier Fontenille. 2012. « Spread of Invasive Aedes Albopictus and Decline of Resident Aedes Aegypti in Urban Areas of Mayotte 2007–2010 ». *Biological invasions* 14 (8) : 1623-33.

Benedetti, Roberto, et Federica Piersimoni. 2017. « A Spatially Balanced Design with Probability Function Proportional to the within Sample Distance ». *Biometrical Journal* 59 (5) : 1067-84.

Benedetti, Roberto, Federica Piersimoni, et Paolo Postiglione. 2015. *Sampling Spatial Units for Agricultural Surveys*. Springer.

———. 2017. « Spatially Balanced Sampling : A Review and a Reappraisal ». *International Statistical Review* 85 (3) : 439-54. https://doi.org/https://doi.org/10.1111/insr.12216.

Bernhardt, Emily S., Margaret A. Palmer, J. D. Allan, G. Alexander, Katie Barnas, Shane Brooks, J. Carr, Stephen Clayton, Cliff Dahm, et Jennifer Follstad-Shah. 2005. *Synthesizing US River Restoration Efforts*. American Association for the Advancement of Science.

Berthou, Patrick, J Huet, P Noel, Michele Jezequel, et Spyros Fifas. 1997. « Etude de La Pêcherie de Palourdes Du Golfe Du Morbihan ». Rapport 38340. France : Ifremer. http://archimer.ifremer.fr/doc/00272/38340/.

Bertignac, Michel, Isabelle Auby, J Foucard, S Martin, X De Montaudouin, et PG Sauriau. 2001. « Evaluation Du Stock de Palourdes Du Bassin d'Arcachon ». Rapport de contrat 21658. France : Ifremer. http://archimer.ifremer.fr/doc/00105/21658/.

Bidegain, Gorka, Javier Francisco Bárcena, Andrés García, et José Antonio Juanes. 2015. « Predicting Coexistence and Predominance Patterns between the Introduced Manila Clam (*Ruditapes Philippinarum*) and the European Native Clam (*Ruditapes Decussatus*) ». *Estuarine, Coastal and Shelf Science* 152 : 162-72.

Bivand, R, E Pebesma, et V Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer. Vol. 10. New york. http://www.asdar-book.org/.

Board, Oklahoma Water Resources. 2013. « The Statewide Stream/River Probabilistic Monitoring Network -Final Report ». 3800 N. Classen, Oklahoma City, Oklahoma 73118.

Bondesson, Lennart, et Daniel Thorburn. 2008a. « A List Sequential Sampling Method Suitable for Real-time Sampling ». *Scandinavian Journal of Statistics* 35 (3) : 466-83.

———. 2008b. « A List Sequential Sampling Method Suitable for Real-time Sampling ». *Scandinavian Journal of Statistics* 35 (3) : 466-83.

Borja, Ángel, et Juan Bald. 2000. « Estado de Los Recursos Marisqueros Del País Vasco En 1998-1999 (Con Especial Atención a Almeja y Berberecho). *Informes Técnicos (Departamento de Agricultura y Pesca, Gobierno Vasco)* » 86 : 78.

Boubidi, Saïd Chaouki. 2016. « Surveillance et Contrôle Du Moustique Tigre, Aedes Albopictus (Skuse, 1894) à Nice, Sud de La France ».

Bouchet, Jean-Marie, Jean-Pierre Deltreil, Francois Manaud, Daniele Maurer, et Gilles Trut. 1997. « Etude Intégrée Du Bassin d'Arcachon-Synthèse ». Rapport 19399. France : Ifremer. http://archimer.ifremer.fr/doc/00083/19399/.

Brown, J.A., B.L. Robertson, et T. McDonald. 2015. « Spatially Balanced Sampling : Application to Environmental Surveys ». *Spatial Statistics conference 2015* 27 : 6-9. https:

//doi.org/10.1016/j.proenv.2015.07.108.

Brown, Jennifer A. 2003. « Designing an Efficient Adaptive Cluster Sample ». *Environmental and Ecological Statistics* 10 (1) : 95-105.

Buckner, Eva A, Mark S Blackmore, Stephen W Golladay, et Alan P Covich. 2011. « Weather and Landscape Factors Associated with Adult Mosquito Abundance in Southwestern Georgia, USA ». *Journal of Vector Ecology* 36 (2) : 269-78.

Byers, James E, Sarah Reichard, John M Randall, Ingrid M Parker, Carey S Smith, WM Lonsdale, IAE Atkinson, TR Seastedt, Mark Williamson, et E Chornesky. 2002. « Directing Research to Reduce the Impacts of Nonindigenous Species ». *Conservation Biology* 16 (3) : 630-40.

Caill-Milly, Nathalie, Joachim Bobinet, Muriel Lissardy, Gilles Morandeau, et Florence Sanchez. 2008. « Campagne d'évaluation Du Stock de Palourdes Du Bassin d'Arcachon-Année 2008 ». Rapport de contrat 17800.

Caill-Milly, Nathalie, Noëlle Bru, Kélig Mahé, Catherine Borie, et Frank D'Amico. 2012. « Shell Shape Analysis and Spatial Allometry Patterns of Manila Clam (*Ruditapes Philippinarum*) in a Mesotidal Coastal Lagoon ». *Journal of Marine Biology* 2012 : 11.

Caill-Milly, Nathalie, Marie-Noëlle de Casamajor, Muriel Lissardy, Florence Sanchez, et Gilles Morandeau. 2003. « Évaluation Du Stock de Palourdes Du Bassin d'Arcachon– Campagne 2003 ». Rapport de contrat 17801.

Caill-Milly, Nathalie, Benoît Duclercq, et Gilles Morandeau. 2006. « Campagne d'évaluation Du Stock de Palourdes Du Bassin d'Arcachon-Année 2006 ». Rapport 2218. France : Ifremer. http://archimer.ifremer.fr/doc/00000/2218/.

Carlton, James T, Janet K Thompson, Laurence E Schemel, et Frederic H Nichols. 1990. « Remarkable Invasion of San Francisco Bay (California, USA), by the Asian Clam *Potamocorbula Amurensis*. I. Introduction and Dispersal ». *Marine Ecology Progress Series* 66 : 81-94.

Carvalho, Silvia B, João Gonçalves, Antoine Guisan, et João P Honrado. 2016. « Systematic Site Selection for Multispecies Monitoring Networks ». *Journal of Applied Ecology* 53 (5) : 1305-16.

Chen, Zhuo. 2018. « Statistical Analysis and Sampling Standards for Maintenance Management Quality Assurance (MMQA) ».

Chiarucci, A., N. J. Enright, G. L. W. Perry, B. P. Miller, et B. B. Lamont. 2003. « Performance of Nonparametric Species Richness Estimators in a High Diversity Plant Community ». *Diversity and Distributions* 9 (4) : 283-95. https://doi.org/10.1046/j.1472-4642.2003.00027.x.

Choy, Samantha Low, Rebecca O'Leary, et Kerrie Mengersen. 2009. « Elicitation by Design in Ecology : Using Expert Opinion to Inform Priors for Bayesian Statistical Models ». *Ecology* 90 (1) : 265-77.

Christianson, Danielle S, et Cari G Kaufman. 2016. « Effects of Sample Design and Landscape Features on a Measure of Environmental Heterogeneity ». *Methods in Ecology and Evolution* 7 (7) : 770-82. https://doi.org/10.1111/2041-210X.12539.

Cochran, William Gemmell. 1977. *Sampling Techniques : 3d Ed.* Wiley.

Conn, Paul B, Erin E Moreland, Eric V Regehr, Erin L Richmond, Michael F Cameron, et Peter L Boveng. 2016. « Using Simulation to Evaluate Wildlife Survey Designs : Polar Bears and Seals in the Chukchi Sea ». *Royal Society Open Science* 3 (1) : 150561. https://doi.org/10.1098/rsos.150561.

Cox, Dennis D., Lawrence H. Cox, et Katherine B. Ensor. 1997. « Spatial Sampling and the Environment : Some Issues and Directions ». *Environmental and Ecological Statistics* 4 (3) : 219-33. https://doi.org/10.1023/A:1018578513217.

Cressie, Noel. 1985. « Fitting Variogram Models by Weighted Least Squares ». *Journal of the International Association for Mathematical Geology* 17 (5) : 563-86.

Cunningham, Ross B, et David B Lindenmayer. 2017. « Approaches to Landscape Scale Inference and Study Design ». *Current Landscape Ecology Reports* 2 (1) : 42-50.

Dam-Bates, Paul van, Oliver Gansell, et Blair Robertson. 2018. « Using Balanced Acceptance Sampling as a Master Sample for Environmental Surveys ». *Methods in Ecology and Evolution* 9 (7) : 1718-26. https://doi.org/https://doi.org/10.1111/2041-210X.13003.

Dang, Cécile. 2009. « Dynamique des populations de palourdes japonaises (*Ruditapes philippinarum*) dans le bassin d'Arcachon : conséquences sur la gestion des populations exploitées ». Thèse de doctorat, Bordeaux 1. http://archimer.ifremer.fr/doc/00000/7382/.

Dang, Cécile, Xavier De Montaudouin, Cindy Binias, Flora Salvo, Nathalie Caill-Milly, Juan Bald, et Philippe Soudant. 2013. « Correlation between Perkinsosis and Growth in Clams Ruditapes Spp. » *Diseases of aquatic organisms* 106 (3) : 255-65.

Davidson, Katie, et Sarah E. Dudas. 2016. « Microplastic Ingestion by Wild and Cultured Manila Clams (*Venerupis Philippinarum*) from Baynes Sound, British Columbia ». *Archives of Environmental Contamination and Toxicology* 71 (2) : 147-56. https://doi.org/10.1007/s00244-016-0286-4.

Defeo, Omar. 2003. « Marine Invertebrate Fisheries in Sandy Beaches : An Overview ». *Journal of Coastal Research*, 56-65. http://www.jstor.org/stable/40928749.

———. 2011. « Sandy Beach Fisheries as Complex Social-Ecological Systems : Emerging Paradigms for Research, Management and Governance ». In, 111-12.

Delatte, Helene, Amelie Desvars, Anthony Bouétard, Séverine Bord, Geoffrey Gimonneau, Gwenaël Vourc'h, et Didier Fontenille. 2010. « Blood-Feeding Behavior of Aedes Albopictus, a Vector of Chikungunya on La Réunion ». *Vector-Borne and Zoonotic Diseases* 10 (3) : 249-58.

De Montaudouin, Xavier, Isabelle Kisielewski, Guy Bachelet, et Céline Desclaux. 2000. « A Census of Macroparasites in an Intertidal Bivalve Community, Arcachon Bay, France ».

Oceanologica Acta 23 (4) : 453-68. https://doi.org/10.1016/S0399-1784(00)00138-9.

De Montaudouin, X, M Lucia, C Binias, M Lassudrie, M Baudrimont, A Legeay, N Raymond, F Jude-Lemeilleur, C Lambert, et N Le Goic. 2015. « Why Is Asari (= Manila) Clam *Ruditapes Philippinarum* Fitness Poor in Arcachon Bay : A Meta-Analysis to Answer ? » *Estuarine, Coastal and Shelf Science* 179 : 226-35. http://dx.doi.org/10.1016/j.ecss.2015.09.009.

Denadai, Márcia R, A Cecília Z. Amaral, et Alexander Turra. 2005. « Along-and Across-shore Components of the Spatial Distribution of the Clam *Tivela Mactroides* (Born, 1778)(Bivalvia, Veneridae) ». *Journal of Natural History* 39 (36) : 3275-95.

Dickson, Maria Michela, Roberto Benedetti, Diego Giuliani, et Giuseppe Espa. 2014. « The Use of Spatial Sampling Designs in Business Surveys ». *Open Journal of Statistics* 4 (05) : 345.

Disease Prevention and Control, European Centre for. 2013. « Environmental Risk Mapping : *Aedes Albopictus* in Europe. » Stockholm.

Dobrowski, Solomon Z, James H Thorne, Jonathan A Greenberg, Hugh D Safford, Alison R Mynsberge, Shawn M Crimmins, et Alan K Swanson. 2011. « Modeling Plant Ranges over 75 Years of Climate Change in California, USA : Temporal Transferability and Species Traits ». *Ecological Monographs* 81 (2) : 241-57.

Dormann, Carsten, Jana McPherson, Miguel Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, et al. 2007. « Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data : A Review ». *Ecography* 30 (5) : 609-28. https://doi.org/10.1111/j.2007.0906-7590.05171.x.

Draper, Norman R, et Harry Smith. 1998. *Applied Regression Analysis*. John Wiley & Sons. Vol. 326. John Wiley & Sons.

Dugan, Jenifer E, et Anton McLachlan. 1999. « An Assessment of Longshore Movement in *Donax Serra Röding* (Bivalvia : Donacidae) on an Exposed Sandy Beach ». *Journal of Experimental Marine Biology and Ecology* 234 (1) : 111-24. https://doi.org/10.1016/S0022-0981(98)00145-2.

Duncan, Greg J, et Graham Kalton. 1987. « Issues of Design and Analysis of Surveys across Time ». *International Statistical Review/Revue Internationale de Statistique*, 97-117.

Eiras, Alvaro E., Marcelo C. Resende, José L. Acebal, et Kelly S. Paixão. 2018. « New Cost-Benefit of Brazilian Technology for Vector Surveillance Using Trapping System ». In *From Local to Global Impact of Mosquitoes*. IntechOpen.

Elith, Jane, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, et Colin J Yates. 2011. « A Statistical Explanation of MaxEnt for Ecologists ». *Diversity and distributions* 17 (1) : 43-57.

Ene, Liviu Theodor, Erik Næsset, et Terje Gobakken. 2016. « Simulation-Based Assessment of Sampling Strategies for Large-Area Biomass Estimation Using Wall-to-Wall and Partial Coverage Airborne Laser Scanning Surveys ». *Remote Sensing of Environment* 176 : 328-40.

Eskildsen, Anne, Peter C le Roux, Risto K Heikkinen, Toke T Høye, W Daniel Kissling, Juha Pöyry, Mary S Wisz, et Miska Luoto. 2013. « Testing Species Distribution Models across Space and Time : High Latitude Butterflies and Recent Warming ». *Global Ecology and Biogeography* 22 (12) : 1293-1303.

Fay, R. W., et DONALD A. Eliason. 1966. « A Preferred Oviposition Site as a Surveillance Method for Aedes Aegypti ». *Mosq news* 26 (4) : 531-35.

Field, Scott A, Andrew J Tyre, et Hugh P Possingham. 2005. « Optimizing Allocation of Monitoring Effort under Economic and Observational Constraints ». *The Journal of Wildlife Management* 69 (2) : 473-82.

Figurska, Małgorzata, Maciej Stańczyk, et Kamil Kulesza. 2008. « Humans Cannot Consciously Generate Random Numbers Sequences : Polemic Study ». *Medical hypotheses* 70 (1) : 182-85.

Fiske, Ian, et Richard Chandler. 2011. « Unmarked : An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance ». *Journal of Statistical Software* 43 (10) : 1-23.

Fontaine, Johnny RJ, Ype H Poortinga, Luc Delbeke, et Shalom H Schwartz. 2008. « Structural Equivalence of the Values Domain across Cultures : Distinguishing Sampling Fluctuations from Meaningful Variation ». *Journal of Cross-Cultural Psychology* 39 (4) : 345-65.

Foster, Scott D. 2016. *MBHdesign : Spatial Designs for Ecological and Environmental Surveys* (version 1.0.61). https://CRAN.R-project.org/package=MBHdesign.

Foster, Scott D, Geoffrey R Hosack, Nicole A Hill, Neville S Barrett, et Vanessa L Lucieer. 2014. « Choosing between Strategies for Designing Surveys : Autonomous Underwater Vehicles ». *Methods in Ecology and Evolution* 5 (3) : 287-97.

Foster, Scott D, Geoffrey R Hosack, Emma Lawrence, Rachel Przeslawski, Paul Hedge, M Julian Caley, Neville S Barrett, Alan Williams, Jin Li, et Tim Lynch. 2017. « Spatially-Balanced Designs That Incorporate Legacy Sites ». *Methods in Ecology and Evolution*.

Fournier, Auriel MV, Easton R White, et Stephen B Heard. 2019. « Site-selection Bias and Apparent Population Declines in Long-term Studies ». *Conservation Biology*.

Gaus, I, DG Kinniburgh, JC Talbot, et R Webster. 2003. « Geostatistical Analysis of Arsenic Concentration in Groundwater in Bangladesh Using Disjunctive Kriging ». *Environmental geology* 44 (8) : 939-48.

Gibson, Laurie D, et Charles T Lenzmeier. 1981. « A Hierarchical Pattern Extraction System for Hexagonally Sampled Images ». DTIC Document.

Goldsmith, Frank Barrie. 2012. *Monitoring for Conservation and Ecology*. Vol. 3. Springer Science & Business Media.

Gosling, Elizabeth. 2008. *Bivalve Molluscs : Biology, Ecology and Culture*. Wiley-Blackwell.

Goulletquer, Philippe, et Cédric Bacher. 1988. « Empirical Modelling of the Growth of *Ruditapes Philippinarum* by Means of Non Linear Regression on Factorial Coordinates ».

Aquatic living resources 1 (03) : 141-54.

Grafström, A, et J Lisic. 2016. « BalancedSampling : Balanced and Spatially Balanced Sampling [Online]. R Package Version 1.5. 2 ».

Grafström, Anton. 2012. « Spatially Correlated Poisson Sampling ». *Journal of Statistical Planning and Inference* 142 (1) : 139-47. https://doi.org/10.1016/j.jspi.2011.07.003.

Grafström, Anton, et Niklas LP Lundström. 2013. « Why Well Spread Probability Samples Are Balanced ». *Open Journal of Statistics* 3 (1) : 36-41.

Grafström, Anton, Niklas LP Lundström, et Lina Schelin. 2012. « Spatially Balanced Sampling through the Pivotal Method ». *Biometrics* 68 (2) : 514-20.

Grafström, Anton, et Lina Schelin. 2014. « How to Select Representative Samples ». *Scandinavian Journal of Statistics* 41 (2) : 277-90.

Grafström, Anton, et Yves Tillé. 2013. « Doubly Balanced Spatial Sampling with Spreading and Restitution of Auxiliary Totals ». *Environmetrics* 24 (2) : 120-31.

Grafström, Anton, Xin Zhao, Martin Nylander, et Hans Petersson. 2017. « A New Sampling Strategy for Forest Inventories Applied to the Temporary Clusters of the Swedish National Forest Inventory ». *Canadian Journal of Forest Research* 47 (9) : 1161-7.

Grard, Gilda, Mélanie Caron, Illich Manfred Mombo, Dieudonné Nkoghe, Statiana Mboui Ondo, Davy Jiolle, Didier Fontenille, Christophe Paupy, et Eric Maurice Leroy. 2014. « Zika Virus in Gabon (Central Africa)–2007 : A New Threat from Aedes Albopictus ? » *PLoS neglected tropical diseases* 8 (2) : e2681.

Gray, Charles A. 2016a. « Assessment of Spatial Fishing Closures on Beach Clams ». *Global Ecology and Conservation* 5 (janvier) : 108-17. https://doi.org/10.1016/j.gecco.2015.12.002.

———. 2016b. « Tide, Time and Space : Scales of Variation and Influences on Structuring and Sampling Beach Clams ». *Journal of Experimental Marine Biology and Ecology* 474 (janvier) : 1-10. https://doi.org/10.1016/j.jembe.2015.09.013.

Gray, Charles A., Daniel D. Johnson, Darren Reynolds, et Douglas Rotherham. 2014. « Development of Rapid Sampling Procedures for an Exploited Bivalve in the Swash Zone on Exposed Ocean Beaches ». *Fisheries Research* 154 (juin) : 205-12. https://doi.org/10.1016/j.fishres.2014.02.027.

Gregoire, Timothy G. 1998. « Design-Based and Model-Based Inference in Survey Sampling : Appreciating the Difference ». *Canadian Journal of Forest Research* 28 (10) : 1429-47.

Guillera-Arroita, Gurutzeta, et José J. Lahoz-Monfort. 2012. « Designing Studies to Detect Differences in Species Occupancy : Power Analysis Under Imperfect Detection ». *Methods in Ecology and Evolution* 3 (5) : 860-69. https://doi.org/10.1111/j.2041-210X.2012.00225.x.

Guillera-Arroita, Gurutzeta, José J Lahoz-Monfort, Jane Elith, Ascelin Gordon, Heini Kujala, Pia E Lentini, Michael A McCarthy, Reid Tingley, et Brendan A Wintle. 2015. « Is My Species Distribution Model Fit for Purpose ? Matching Data and Models to Applications ». *Global Ecology and Biogeography* 24 (3) : 276-92.

Guillera-Arroita, Gurutzeta, Martin S Ridout, et Byron JT Morgan. 2010. « Design of Occupancy Studies with Imperfect Detection ». *Methods in Ecology and Evolution* 1 (2) : 131-39.

Haining, Robert P. 2003. *Spatial Data Analysis : Theory and Practice*. United KIingdom : Cambridge University Press.

Halton, John H. 1960. « On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals ». *Numerische Mathematik* 2 (1) : 84-90.

Hasel, Austin A. 1938. « Sampling Error in Timber Surveys. » *Journal of Agricultural Research*, 1938.

Hayward, Matt W., Luigi Boitani, Neil D. Burrows, Paul J. Funston, K. Ullas Karanth, Darryl I. MacKenzie, Ken H. Pollock, et Richard W. Yarnell. 2015. « Ecologists Need Robust Survey Designs, Sampling and Analytical Methods ». *Journal of Applied Ecology* 52 (2) : 286-90.

Hengl, Tomislav, Gerard BM Heuvelink, et Alfred Stein. 2004. « A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging ». *Geoderma* 120 (1-2) : 75-93.

Henley, Stephen. 2012. *Nonparametric Geostatistics*. Springer Science & Business Media.

Horvitz, Daniel G, et Donovan J Thompson. 1952. « A Generalization of Sampling without Replacement from a Finite Universe ». *Journal of the American statistical Association* 47 (260) : 663-85.

Hurlbert, Stuart H. 1984. « Pseudoreplication and the Design of Ecological Field Experiments ». *Ecological monographs* 54 (2) : 187-211.

Inglada, Jordi, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, et Isabel Rodes. 2017. « Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series ». *Remote Sensing* 9 (1) : 95.

Jackson, Andrew L., Annette C. Broderick, Wayne J. Fuller, Fiona Glen, Graeme D. Ruxton, et Brendan J. Godley. 2008. « Sampling Design and Its Effect on Population Monitoring : How Much Monitoring Do Turtles Really Need ? » *Biological Conservation* 141 (12) : 2932-41. https://doi.org/10.1016/j.biocon.2008.09.002.

Jacobs, Steven E, William Gaeuman, Matt A Weeber, Stephanie L Gunckel, et Steven J Starcevich. 2009. « Utility of a Probabilistic Sampling Design to Determine Bull Trout Population Status Using Redd Counts in Basins of the Columbia River Plateau ». *North American Journal of Fisheries Management* 29 (6) : 1590-1604.

Jaksons, Peter. 2014. « A New Approach to Adaptive Monitoring ».

James, RJ, et PG Fairweather. 1996. « Spatial Variation of Intertidal Macrofauna on a Sandy Ocean Beach in Australia ». *Estuarine, Coastal and Shelf Science* 43 (1) : 81-107.

Jensen, AC, J Humphreys, RWG Caldow, C Grisley, et PEJ Dyrynda. 2004. « Naturalization of the Manila Clam (*Tapes Philippinarum*), an Alien Species, and Establishment of a Clam

Fishery within Poole Harbour, Dorset ». *Journal of the Marine Biological Association of the UK* 84 (05) : 1069-73.

Jentes, Emily S, Gilles Poumerol, Mark D Gershman, David R Hill, Johan Lemarchand, Rosamund F Lewis, J Erin Staples, Oyewale Tomori, Annelies Wilder-Smith, et Thomas P Monath. 2011. « The Revised Global Yellow Fever Risk Map and Recommendations for Vaccination, 2010 : Consensus of the Informal WHO Working Group on Geographic Risk for Yellow Fever ». *The Lancet infectious diseases* 11 (8) : 622-32.

Juanes, José Antonio, Gorka Bidegain, Beatriz Echavarri-Erasun, Araceli Puente, Ana García, Andrés García, Javier F. Bárcena, César Álvarez, et Gerardo García-Castillo. 2012. « Differential Distribution Pattern of Native *Ruditapes Decussatus* and Introduced *Ruditapes Phillippinarum* Clam Populations in the Bay of Santander (Gulf of Biscay) : Considerations for Fisheries Management ». *Ocean & Coastal Management* 69 (décembre) : 316-26. https://doi.org/10.1016/j.ocecoaman.2012.08.007.

Kalyagina, EE. 1995. « Distribution and Population Structure of Commercial Bivalves *Ruditapes Philippinarum* and *Mya Arenaria* in Bousse Lagoon(Southern Sakhalin) ». *Russ. J. Mar. Biol. 20* 3 (3) : 164-68.

Kang, Su Yun, James M McGree, Christopher C Drovandi, M Julian Caley, et Kerrie L Mengersen. 2016. « Bayesian Adaptive Design : Improving the Effectiveness of Monitoring of the Great Barrier Reef ». *Ecological applications* 26 (8) : 2637-48.

Katz, Jacob, Peter B Moyle, Rebecca M Quiñones, Joshua Israel, et Sabra Purdy. 2013. « Impending Extinction of Salmon, Steelhead, and Trout (*Salmonidae*) in California ». *Environmental Biology of Fishes* 96 (10-11) : 1169-86.

Keinath, Douglas A, et Bighorn Canyon NRA. 2016. « Bat Population Monitoring of Bighorn Canyon National Recreation Area : 2015 Progress Report », 2016.

Kenkel, NC, P Juhász-Nagy, et J Podani. 1990. « On Sampling Procedures in Population and Community Ecology ». *Progress in theoretical vegetation science* 83 : 195-207. https://doi.org/https://doi.org/10.1007/BF00031692.

Kermorvant, Claire, Nathalie Caill-Milly, Noëlle Bru, et Franck D'Amico. 2019. « Optimizing Cost-Efficiency of Long Term Monitoring Programs by Using Spatially Balanced Sampling Designs : The Case of Manila Clams in Arcachon Bay ». *Ecological Informatics* 49 : 32-39. https://doi.org/https://doi.org/10.1016/j.ecoinf.2018.11.005.

Kermorvant, Claire, Nathalie Caill-Milly, Frank D'Amico, Noëlle Bru, Florence Sanchez, Muriel Lissardy, et Jennifer Brown. 2017. « Optimization of a Survey Using Spatially Balanced Sampling : A Single-Year Application of Clam Monitoring in the Arcachon Bay (SW France) ». *Aquatic Living Resources* 30 : 37.

Kincaid, Thomas M, et Anthony R Olsen. 2015. *Spsurvey : Spatial Survey Design and Analysis. R Package Version 3.1.* http://www.epa.gov/nheerl/arm/.

Kincaid, Tom, et Tony Olsen. 2016. *Spsurvey : Spatial Survey Design and Analysis* (version 3.3). https://CRAN.R-project.org/package=spsurvey.

Kingston, P F. 2009. « Grabs for Shelf Benthic Sampling ». In *Encyclopedia of Ocean Sciences (Second Edition)*, Academic Press, 70-79. Oxford : John H. Steele. http://dx.doi.org/10.1016/B978-012374473-9.00667-6.

Kombiadou, Katerina, Florian Ganthy, Romaric Verney, et Aldo Sottolichio. 2014. « Modelling the Effects of *Zostera Noltei* Meadows on Sediment Dynamics : Application to the Arcachon Lagoon ». *Ocean Dynamics* 64 (10) : 1499-1516.

Krige, Daniel G. 1951. « A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand ». *Journal of the Southern African Institute of Mining and Metallurgy* 52 (6) : 119-39.

Lanier, Wendy E., Larissa L. Bailey, et Erin Muths. 2016. « Integrating Biology, Field Logistics, and Simulations to Optimize Parameter Estimation for Imperiled Species ». *Ecological Modelling* 335 (septembre) : 16-23. https://doi.org/10.1016/j.ecolmodel.2016.05.006.

Lazarina, Maria, Athanasios S. Kallimanis, John D. Pantis, et Stefanos P. Sgardelis. 2014. « Linking Species Richness Curves from Non-Contiguous Sampling to Contiguous-Nested SAR : An Empirical Study ». *Acta Oecologica* 61 (novembre) : 24-31. https://doi.org/10.1016/j.actao.2014.10.001.

Le, Thanh Cuong, Hyun-Sil Kang, Hyun-Ki Hong, Kwang-Jae Park, et Kwang-Sik Choi. 2015. « First Report of *Urosporidium Sp.*, a Haplosporidian Hyperparasite Infecting Digenean Trematode *Parvatrema Duboisi* in Manila Clam, *Ruditapes Philippinarum* on the West Coast of Korea ». *Journal of invertebrate pathology* 130 : 141-46. https://doi.org/https://doi.org/10.1016/j.jip.2015.08.004.

Le Cacheux, Paul. 1955. « Analyse Statistique de La Forêt Tropicale En Vue de Son Utilisation Pour La Production de Cellulose ». *Journal d'agriculture tropicale et de botanique appliquée* 2 (1) : 1-17. https://doi.org/https://doi.org/10.3406/jatba.1955.2198.

Legendre, Pierre, Mark RT Dale, Marie-Josée Fortin, Jessica Gurevitch, Michael Hohn, et Donald Myers. 2002. « The Consequences of Spatial Structure for the Design and Analysis of Ecological Field Surveys ». *Ecography* 25 (5) : 601-15. https://doi.org/https://doi.org/10.1034/j.1600-0587.2002.250508.x.

Legg, Colin J, et Laszlo Nagy. 2006. « Why Most Conservation Monitoring Is, but Need Not Be, a Waste of Time ». *Journal of environmental management* 78 (2) : 194-99. https://doi.org/https://doi.org/10.1016/j.jenvman.2005.04.016.

Le Treut, Yannig. 1986. « La Palourde. Anatomie, Biologie, Elevage, Peche, Consommation, Inspection Sanitaire. » Thèse de doctorat, Université de Nantes : Ecole Nationale Vétérinaire.

Levy, Paul S, et Stanley Lemeshow. 2013. *Sampling of Populations : Methods and Applications*. New Jersey : John Wiley & Sons.

Li, Mingyang, Ting Xu, et Qi Zhou. 2012. « Development of Python-Based ArcGIS Tools for Spatially Balanced Forest Sampling Design ». In, 939-42. Atlantis Press. https://doi.org/https://doi.org/10.2991/citcs.2012.109.

Liberts, Mārtiņš. 2013. « The Cost Efficiency of Sampling Designs » 14 (1) : 7-30.

Likens, Gene, et David Lindenmayer. 2018. *Effective Ecological Monitoring*. CSIRO publishing.

Lohr, Sharon. 2009. *Sampling : Design and Analysis*. Boston : Nelson Education.

MacKenzie, Darryl I. 2006. *Occupancy Estimation and Modeling : Inferring Patterns and Dynamics of Species Occurrence*. Elsevier. London : Academic Press.

MacKenzie, Darryl I, et J Andrew Royle. 2005. « Designing Occupancy Studies : General Advice and Allocating Survey Effort ». *Journal of Applied Ecology* 42 (6) : 1105-14. https://doi.org/https://doi.org/10.1111/j.1365-2664.2005.01098.x.

Manica, Mattia, Roberto Rosà, Alessandra della Torre, et Beniamino Caputo. 2017. « From Eggs to Bites : Do Ovitrap Data Provide Reliable Estimates of Aedes Albopictus Biting Females ? » *PeerJ* 5 : e2998. https://doi.org/https://doi.org/10.7717/peerj.2998.

Martin, Julien, Wiley M Kitchens, et James E Hines. 2007. « Importance of Well-designed Monitoring Programs for the Conservation of Endangered Species : Case Study of the Snail Kite ». *Conservation Biology* 21 (2) : 472-81. https://doi.org/https://doi.org/10.1111/j.1523-1739.2006.00613.x.

Matheron, Georges. 1963. « Principles of Geostatistics ». *Economic geology* 58 (8) : 1246-66. https://doi.org/https://doi.org/10.2113/gsecongeo.58.8.1246.

McDonald, Trent. 2014. « Sampling Designs for Environmental Monitoring ». In *Introduction to Ecological Sampling*, 145-66. Chapman and Hall/CRC.

———. 2016. *SDraw : Spatially Balanced Samples of Spatial Objects. R Package Version 2.1.8. https ://CRAN.R-project.org/package=SDraw* (version 2.1.3). https://github.com/tmcd82070/SDraw/wiki/SDraw.

McDonald, Trent L. 2003. « Review of Environmental Monitoring Methods : Survey Designs ». *Environmental Monitoring and Assessment* 85 (3) : 277-92. https://doi.org/10.1023/A:1023954311636.

McGarvey, Richard, Paul Burch, et Janet M Matthews. 2016. « Precision of Systematic and Random Sampling in Clustered Populations : Habitat Patches and Aggregating Organisms ». *Ecological Applications* 26 (1) : 233-48. https://doi.org/https://doi.org/10.1890/14-1973.

Mckann, Patrick C, Brian R Gray, et Wayne E Thogmartin. 2013. « Small Sample Bias in Dynamic Occupancy Models ». *The Journal of Wildlife Management* 77 (1) : 172-80. https://doi.org/https://doi.org/10.1002/jwmg.433.

McLachlan, Anton, Jenifer E Dugan, Omar Defeo, AD Ansell, DM Hubbard, E Jaramillo, et PE Penchaszadeh. 1996. « Beach Clam Fisheries ». *Oceanography and marine biology : an annual review* 34 : 163-232.

Medley, Kim A. 2010. « Niche Shifts during the Global Invasion of the Asian Tiger Mosquito, Aedes Albopictus Skuse (Culicidae), Revealed by Reciprocal Distribution Models ». *Global ecology and biogeography* 19 (1) : 122-33. https://doi.org/https://doi.org/10.1111/j.1466-8238.2009.00497.x.

Melià, P, GA De Leo, et M Gatto. 2004. « Density and Temperature-Dependence of Vital Rates in the Manila Clam *Tapes Philippinarum* : A Stochastic Demographic Model ». *Marine Ecology Progress Series* 272 : 153-64. https://doi.org/doi:10.3354/meps272153.

Messer, Jay J, Rick A Linthurst, et W Scott Overton. 1991. « An EPA Program for Monitoring Ecological Status and Trends ». *Environmental Monitoring and Assessment* 17 (1) : 67-78. https://doi.org/https://doi.org/10.1007/BF00402462.

Moore, Alana L, et Michael A McCarthy. 2016. « Optimizing Ecological Survey Effort over Space and Time ». *Methods in Ecology and Evolution* 7 (8) : 891-99. https://doi.org/https://doi.org/10.1111/2041-210X.12564.

Müller, Werner G., Juan M. Rodríguez-Díaz, et María J. Rivas López. 2012. « Optimal Design for Detecting Dependencies with an Application in Spatial Ecology ». *Environmetrics* 23 (1) : 37-45. https://doi.org/10.1002/env.1132.

Neyman, Jerzy. 1934. « On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection ». In *Kotz S., Johnson N.L. (Eds) Breakthroughs in Statistics*, 97 :558-625. Springer Series in Statistics (Perspectives in Statistics). New York : Springer.

Nichols, James D, et Byron K Williams. 2006. « Monitoring for Conservation ». *Trends in ecology & evolution* 21 (12) : 668-73. https://doi.org/https://doi.org/10.1016/j.tree.2006.08.007.

Oliver, MA, et R Webster. 2014. « A Tutorial Guide to Geostatistics : Computing and Modelling Variograms and Kriging ». *Catena* 113 : 56-69. https://doi.org/https://doi.org/10.1016/j.catena.2013.09.006.

Olsen, AR, TM Kincaid, et Q Payton. 2012. « Spatially Balanced Survey Designs for Natural Resources ». In *Design and Analysis of Long-Term Ecological Monitoring Studies*, 126-50. New York : Cambridge University Press.

Olu K, Duperret A, Sibuet M, Foucher JP, et Fiala-Médioni A. 1996. « Structure and Distribution of Cold Seep Communities along the Peruvian Active Margin : Relationship to Geological and Fluid Patterns ». *Marine Ecology Progress Series* 132 : 109-25. http://www.int-res.com/abstracts/meps/v132/p109-125/.

Paillard, Christine, Bassem Allam, et Radouane Oubella. 2004. « Effect of Temperature on Defense Parameters in Manila Clam *Ruditapes Philippinarum* Challenged with *Vibrio Tapetis* ». *Diseases of aquatic organisms* 59 (3) : 249-62. https://doi.org/doi:10.3354/dao059249.

Pantalone, Roberto Benedetti, et Federica Piersimoni. 2019. *Spbsampling : Spatially Balanced Sampling. [Online]. R Package Version 1.2.0.*

Papritz, A, et JR Dubois. 1999. « Mapping Heavy Metals in Soil by (Non-) Linear Kriging : An Empirical Validation ». In *geoENV II—Geostatistics for Environmental Applications*, 429-40. Dordrecht : Springer.

Park, Kyung-Il, et Kwang-Sik Choi. 2001. « Spatial Distribution of the Protozoan Parasite *Perkinsus Sp.* Found in the Manila Clams, *Ruditapes Philippinarum*, in Korea ». *Aquaculture*

203 (1) : 9-22. https://doi.org/https://doi.org/10.1016/S0044-8486(01)00619-6.

Pebesma, E, et R Bivand. 2005. « Classes and Methods for Spatial Data in R. » *R News*, 2005. http://cran.r-project.org/doc/Rnews/.

Peñalver-Alcázar, Miguel, Pedro Aragón, Merel C Breedveld, et Patrick S Fitze. 2016. « Microhabitat Selection in the Common Lizard : Implications of Biotic Interactions, Age, Sex, Local Processes, and Model Transferability among Populations ». *Ecology and evolution* 6 (11) : 3594-3607. https://doi.org/https://doi.org/10.1002/ece3.2138.

Perry, JN, AM Liebhold, MS Rosenberg, J Dungan, M Miriti, A Jakomulska, et S Citron-Pousty. 2002. « Illustrations and Guidelines for Selecting Statistical Methods for Quantifying Spatial Pattern in Ecological Data ». *Ecography* 25 (5) : 578-600. https://doi.org/https://doi.org/10.1034/j.1600-0587.2002.250507.x.

Persaud, Navindra. 2005. « Humans Can Consciously Generate Random Number Sequences : A Possible Test for Artificial Intelligence ». *Medical Hypotheses* 65 (2) : 211-14. https://doi.org/10.1016/j.mehy.2005.02.019.

Peterman, Randall M. 1990. « Statistical Power Analysis Can Improve Fisheries Research and Management ». *Canadian Journal of Fisheries and Aquatic Sciences* 47 (1) : 2-15. https://doi.org/https://doi.org/10.1139/f90-001.

Petitgas, Pierre. 1993. « Use of a Disjunctive Kriging to Model Areas of High Pelagic Fish Density in Acoustic Fisheries Surveys ». *Aquatic Living Resources* 6 (3) : 201-9. https://doi.org/https://doi.org/10.1051/alr:1993021.

———. 2001. « Geostatistics in Fisheries Survey Design and Stock Assessment : Models, Variances and Applications ». *Fish and Fisheries* 2 (3) : 231-49. https://doi.org/https://doi.org/10.1046/j.1467-2960.2001.00047.x.

Pitel, Mathilde, Marie Savina, Spyros Fifas, et Patrick Berthou. 2004. « Evaluations Locales Des Populations de Bivalves Dans Le Golfe de Normand Breton. Résultats de La Campagne BIVALVES2002 - IFREMER », 6-7. http://archimer.ifremer.fr/doc/00000/4609/.

Plus, M, D Maurer, JY Stanisière, et F Dumas. 2006. « Caractérisation Des Composantes Hydrodynamiques d'une Lagune Mésotidale, Le Bassin d'Arcachon. » Rapport 2352. France : Ifremer. http://archimer.ifremer.fr/doc/00000/2352/.

Pranovi, F., G. Franceschini, M. Casale, M. Zucchetta, P. Torricelli, et O. Giovanardi. 2006. « An Ecological Imbalance Induced by a Non-Native Species : The Manila Clam in the Venice Lagoon ». *Biological Invasions* 8 (4) : 595-609. https://doi.org/10.1007/s10530-005-1602-5.

Rajabi, Mohammad Mahdi, et Behzad Ataie-Ashtiani. 2014. « Sampling Efficiency in Monte Carlo Based Uncertainty Propagation Strategies : Application in Seawater Intrusion Simulations ». *Advances in Water Resources* 67 : 46-64. https://doi.org/https://doi.org/10.1016/j.advwatres.2014.02.004.

Reinert, John F, Ralph E Harbach, et Ian J Kitching. 2009. « Phylogeny and Classification of Tribe Aedini (Diptera : Culicidae) ». *Zoological Journal of the Linnean Society* 157 (4) : 700-794. https://doi.org/https://doi.org/10.1111/j.1096-3642.2009.00570.x.

Reiter, PAUL, et MAAAN Colon. 1991. « Enhancement of the Cdc Ovitrap with Hay Infusions for Daily Monitoring of *Aedes Aegypy < \i> Populations* » 7 (1) : 52-55.

Renard, D, N Bez, N Desassis, H Beucher, F Ors, et F Laporte. 2014. « RGeostats : The Geostatistical Package [11.0.1]. MINES ParisTech / ARMINES. » http://cg.ensmp.fr/rgeostats.

Resende, Marcelo Carvalho de, Ivoneide Maria Silva, Brett R Ellis, et Alvaro Eduardo Eiras. 2013. « A Comparison of Larval, Ovitrap and MosquiTRAP Surveillance for Aedes (Stegomyia) Aegypti ». *Memórias do Instituto Oswaldo Cruz* 108 (8) : 1024-30. https://doi.org/http://dx.doi.org/10.1590/0074-0276130128.

Robert, René, et Jean-Pierre Deltreil. 1990. « Élevage de La Palourde Japonaise *Ruditapes Philippinarum* Dans Le Bassin d'Arcachon. Bilan Des Dix Dernières Années et Perspectives de Développement ». Rapport 1652. France : Ifremer. http://archimer.ifremer.fr/doc/00000/1652/.

Robert, R, G Trut, et JL Laborde. 1993. « Growth, Reproduction and Gross Biochemical Composition of the Manila Clam *Ruditapes Philippinarum* in the Bay of Arcachon, France ». *Marine biology* 116 (2) : 291-99. https://doi.org/https://doi.org/10.1007/BF00350019.

Roberts, Kevin A. 1991. « Field Monitoring : Confessions of an Addict ». In *Monitoring for Conservation and Ecology*, 179-211. Dordrecht : Springer.

Robertson, Blair, Jennifer Brown, Trent McDonald, et Peter Jaksons. 2013. « BAS : Balanced Acceptance Sampling of Natural Resources ». *Biometrics* 69 (3) : 776-84. https://doi.org/https://doi.org/10.1111/biom.12059.

Robertson, Blair, Trent McDonald, Chris Price, et Jennifer Brown. 2018. « Halton Iterative Partitioning : Spatially Balanced Sampling via Partitioning ». *Environmental and Ecological Statistics* 25 (3) : 1-19. https://doi.org/https://doi.org/10.1007/s10651-018-0406-6.

Robertson, BL, T McDonald, CJ Price, et JA Brown. 2017. « A Modification of Balanced Acceptance Sampling ». *Statistics & Probability Letters* 129 : 107-12. https://doi.org/0167-7152.

Roche, Benjamin, Lucas Léger, Grégory L'Ambert, Guillaume Lacour, Rémi Foussadier, Gilles Besnard, Hélène Barré-Cardi, Frédéric Simard, et Didier Fontenille. 2015. « The Spread of Aedes Albopictus in Metropolitan France : Contribution of Environmental Drivers and Human Activities and Predictions for a near Future ». *PLoS One* 10 (5) : e0125600.

Rochlin, Ilia, Dominick V Ninivaggi, Michael L Hutchinson, et Ary Farajollahi. 2013. « Climate Change and Range Expansion of the Asian Tiger Mosquito (Aedes Albopictus) in Northeastern USA : Implications for Public Health Practitioners ». *PloS one* 8 (4) : e60874. https://doi.org/https://doi.org/doi:10.7282/T3JD4V1H.

Royle, JA, et RM Dorazio. 2008. « Hierarchical Modelling and Inference in Ecology ». *Á Acad. Press, New York*.

Rudders, David. 2011. « A Simulation Study to Evaluate Sampling Designs for Highly Autocorrelated Populations : With an Application to Sea Scallop Closed Areas », 550-50.

Samson, Dayana M, Reginald S Archer, Temitope O Alimi, Kristopher L Arheart, Daniel

E Impoinvil, Roland Oscar, Douglas O Fuller, et Whitney A Qualls. 2015. « New Baseline Environmental Assessment of Mosquito Ecology in Northern Haiti during Increased Urbanization ». *Journal of Vector Ecology* 40 (1) : 46-58.

Sanchez, Florence, Nathalie Caill-Milly, et Lissardy Muriel De Casamajor Marie-Noelle. 2012. « Campagne d'évaluation Du Stock de Palourdes Du Bassin d'Arcachon. » Rapport de contrat 24114. France : Ifremer.

Sanchez, Florence, Nathalie Caill-Milly, Muriel Lissardy, et Noelle Bru. 2014. « Campagne d'évaluation de Stock de Palourdes Du Bassin d'Arcachon. » Rapport 34383. France : Ifremer.

Sanchez, Florence, Nathalie Caill-Milly, Muriel Lissardy, Marie-Noëlle De Casamajor, et Gilles Morandeau. 2010. « Campagne d'évaluation Du Stock de Palourdes Du Bassin d'Arcachon. » Rapport de contrat 16331. France : Ifremer.

Särndal, Carl-Erik, Ib Thomsen, Jan M Hoem, DV Lindley, O Barndorff-Nielsen, et Tore Dalenius. 1978. « Design-Based and Model-Based Inference in Survey Sampling [with Discussion and Reply] ». *Scandinavian Journal of Statistics*, 27-52.

Schwartz, Shalom H, et Lilach Sagiv. 1995. « Identifying Culture-Specifics in the Content and Structure of Values ». *Journal of cross-cultural psychology* 26 (1) : 92-116.

Sica, Gregory T. 2006. « Bias in Research Studies 1 ». *Radiology* 238 (3) : 780-89.

Simmons, Cameron P, Jeremy J Farrar, Nguyen van Vinh Chau, et Bridget Wills. 2012. « Dengue ». *New England Journal of Medicine* 366 (15) : 1423-32.

Smith, Adam NH, Marti J Anderson, et Matthew DM Pawley. 2017. « Could Ecologists Be More Random ? Straightforward Alternatives to Haphazard Spatial Sampling ». *Ecography* 40 (11) : 1251-5. https://doi.org/10.1111/ecog.02821.

Soberon, Jorge, et A Townsend Peterson. 2005. « Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas » 2. https://doi.org/https://doi.org/10.17161/bi.v2i0.4.

Soudant, Philippe, Christine Paillard, Gwenaelle Choquet, Christophe Lambert, HI Reid, Alain Marhic, Ludovic Donaghy, et TH Birkbeck. 2004. « Impact of Season and Rearing Site on the Physiological and Immunological Parameters of the Manila Clam *Venerupis* (= *Tapes*,= *Ruditapes*) *Philippinarum* ». *Aquaculture* 229 (1) : 401-18.

Stehman, Stephen V, et W Scott Overton. 1994. « 9 Environmental Sampling and Monitoring ». In *Handbook of Statistics*, 12 :263-306. United Kingdom : Elsevier.

———. 1996. « Spatial Sampling ». *Practical handbook of spatial statistics*, 31-63.

Stevens, Don L, et Anthony R Olsen. 1999. « Spatially Restricted Surveys over Time for Aquatic Resources ». *Journal of Agricultural, Biological, and Environmental Statistics*, 415-28.

———. 2003. « Variance Estimation for Spatially Balanced Samples of Environmental Resources ». *Environmetrics* 14 (6) : 593-610.

———. 2004. « Spatially Balanced Sampling of Natural Resources ». *Journal of the American Statistical Association* 99 (465) : 262-78.

Stevens Jr, Don L, et Anthony R Olsen. 2004. « Spatially Balanced Sampling of Natural Resources ». *Journal of the American Statistical Association* 99 (465) : 262-78.

Talley, Drew M, Theresa Sinicrope Talley, et Alexander Blanco. 2015. « Insights into the Establishment of the Manila Clam on a Tidal Flat at the Southern End of an Introduced Range in Southern California, USA ». *PloS one* 10 (3) : e0118891.

Tamayo, David, Irrintzi Ibarrola, Juan Cigarría, et Enrique Navarro. 2015. « The Effect of Food Conditioning on Feeding and Growth Responses to Variable Rations in Fast and Slow Growing Spat of the Manila Clam (*Ruditapes Philippinarum*) ». *Journal of Experimental Marine Biology and Ecology* 471 : 92-103.

Tamura, T. 1970. *Marine Aquaculture*. Washington : National science foundation.

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

Theobald, David M, Don L Stevens Jr, Denis White, N Scott Urquhart, Anthony R Olsen, et John B Norman. 2007. « Using GIS to Generate Spatially Balanced Random Survey Designs for Natural Resource Applications ». *Environmental Management* 40 (1) : 134-46.

Thompson, Steve K. 2012. *Sampling, 3rd Edition*. John Wiley & Sons, Inc.

Tillé, Yves. 2011. *Théorie Des Sondages : Échantillonnage et Estimation En Population Finie : Cours et Exercices Avec Solutions*. Vol. 13. Dunod.

Triantafilis, J, IOA Odeh, Benjamin Warr, et MF Ahmed. 2004. « Mapping of Salinity Risk in the Lower Namoi Valley Using Non-Linear Kriging Methods ». *Agricultural Water Management* 69 (3) : 203-31.

Vanwambeke, Sophie O, Pradya Somboon, Ralph E Harbach, Mark Isenstadt, Eric F Lambin, Catherine Walton, et Roger K Butlin. 2007. « Landscape and Land Cover Factors Influence the Presence of Aedes and Anopheles Larvae ». *Journal of medical entomology* 44 (1) : 133-44.

Vicente, Joana R, Diogo Alagador, Carlos Guerra, Joaquim M Alonso, Christoph Kueffer, Ana S Vaz, Rui F Fernandes, João A Cabral, Miguel B Araújo, et João P Honrado. 2016. « Cost-effective Monitoring of Biological Invasions under Global Change : A Model-based Framework ». *Journal of Applied Ecology* 53 (5) : 1317-29.

Vos, P, E Meelis, et WJ Ter Keurs. 2000a. « A Framework for the Design of Ecological Monitoring Programs as a Tool for Environmental and Nature Management ». *Environmental monitoring and assessment* 61 (3) : 317-44.

———. 2000b. « A Framework for the Design of Ecological Monitoring Programs as a Tool for Environmental and Nature Management ». *Environmental monitoring and assessment* 61 (3) : 317-44.

Walker, Randal L, et Kenneth R Tenore. 1984. « The Distribution and Production of the Hard Clam, *Mercenaria Mercenaria*, in Wassaw Sound, Georgia ». *Estuaries* 7 (1) : 19-27.

Walther, Bruno A, et Joslin L Moore. 2005. « The Concepts of Bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance ». *Ecography* 28 (6) : 815-29.

Wang, Jin-Feng, Cheng-Sheng Jiang, Mao-Gui Hu, Zhi-Dong Cao, Yan-Sha Guo, Lian-Fa Li, Tie-Jun Liu, et Bin Meng. 2012. « Design-Based Spatial Sampling : Theory and Implementation ». *Environmental modelling & software* 40 : 280-88.

Wang, Jin-Feng, A Stein, Bin-Bo Gao, et Yong Ge. 2012. « A Review of Spatial Sampling ». *Spatial Statistics* 2 : 1-14.

Wang, X., et F.J. Hickernell. 2000. « Randomized Halton Sequences ». *Mathematical and Computer Modelling* 32 (7) : 887-99. https://doi.org/10.1016/S0895-7177(00)00178-3.

Webb, JA, SC De Little, KA Miller, MJ Stewardson, ID Rutherfurd, AK Sharpe, L Patulny, et NL Poff. 2015. « A General Approach to Predicting Ecological Responses to Environmental Flows : Making Best Use of the Literature, Expert Knowledge, and Monitoring Data ». *River Research and Applications* 31 (4) : 505-14.

Webster, R, et MA Oliver. 1993. « How Large a Sample Is Needed to Estimate the Regional Variogram Adequately ? » In *Geostatistics Tróia'92*, 155-66. Springer.

Wekell, John C., Erich J. Gauglitz, Harold J. Bamett, Christine L Hatfield, Doug Simons, et Daniel Ayres. 1994. « Occurrence of Domoic Acid in Washington State Razor Clams (*Siliqua Patula*) during 1991-1993 ». *Natural Toxins* 2 (4) : 197-205. https://doi.org/10.1002/nt.2620020408.

Wenger, Seth J, et Julian D Olden. 2012. « Assessing Transferability of Ecological Models : An Underappreciated Aspect of Statistical Validation ». *Methods in Ecology and Evolution* 3 (2) : 260-67.

Wogan, Guinevere OU. 2016. « Life History Traits and Niche Instability Impact Accuracy and Temporal Transferability for Historically Calibrated Distribution Models of North American Birds ». *PLoS One* 11 (3) : e0151024.

Yates, Katherine L, Phil J Bouchet, M Julian Caley, Kerrie Mengersen, Christophe F Randin, Stephen Parnell, Alan H Fielding, Andrew J Bamford, Stephen Ban, et A Márcia Barbosa. 2018. « Outstanding Challenges in the Transferability of Ecological Models ». *Trends in Ecology & Evolution* 33 (10) : 790-802. https://doi.org/https://doi.org/10.1016/j.tree.2018.08.001.

Yoccoz, Nigel G, James D Nichols, et Thierry Boulinier. 2001. « Monitoring of Biological Diversity in Space and Time ». *Trends in Ecology & Evolution* 16 (8) : 446-53.

Yu, Hao, Yan Jiao, Zhenming Su, et Kevin Reid. 2012. « Performance Comparison of Traditional Sampling Designs and Adaptive Sampling Designs for Fishery-Independent Surveys : A Simulation Study ». *Fisheries research* 113 (1) : 173-81.

Zhao, Gang, Holger Hoffmann, Jagadeesh Yeluripati, Specka Xenia, Claas Nendel, Elsa Coucheney, Matthias Kuhnert, Fulu Tao, Julie Constantin, et Helene Raynal. 2016. « Evaluating

the Precision of Eight Spatial Sampling Schemes in Estimating Regional Means of Simulated Yield for Two Crops ». *Environmental Modelling & Software* 80 : 100-112.

Zinger, A. 1963. « Estimations de Variances Avec Échantillonnage Systématique ». *Revue de statistique Appliquee* 11 (2) : 89-97.

Zurell, Damaris, Uta Berger, Juliano S Cabral, Florian Jeltsch, Christine N Meynard, Tamara Münkemüller, Nana Nehrbass, Jörn Pagel, Björn Reineking, et Boris Schröder. 2010. « The Virtual Ecologist Approach : Simulating Data and Observers ». *Oikos* 119 (4) : 622-35.

Appendix A : Poster presented at “les journées R”

De nouveaux packages pour sélectionner des points d’échantillonnage spatialement équilibrés sous

Kermorvant Claire[(1)], D’Amico Frank[(1)], Bru Noëlle[(1)], Caill-Milly Nathalie[(2)]

(1) CNRS / UNIV PAU & PAYS ADOUR, UFR Sciences et Techniques de la Côte Basque (FED 4155 MIRA) – 1 Allée Parc Montaury, 64600 Anglet, France,
(2) IFREMER - Laboratoire Environnement Ressources Arcachon (FED 4155 MIRA) – 1 Allée Parc Montaury, 64600 Anglet, France.

Contexte: Les protocoles d’échantillonnage constituent un mode de sélection des unités d’échantillonnages de la population statistique, permettant la représentativité des résultats. Dans les suivis écologiques, la population statistique est une zone spatiale. Une étude récente montre que seulement 21% des articles scientifiques parus en 2012 utilisent explicitement un protocole d’échantillonnage probabiliste[1]. Si un protocole non-probabiliste est utilisé, les résultats obtenus ne répondent généralement pas aux règles d’application des statistiques inférentielles et doivent être manipulés prudemment. R est un outil qui permet d’appliquer facilement des protocoles probabilistes complexes facilement. Ici, nous présentons deux packages R (« Spsurvey » et « SDraw ») qui permettent, en plus de protocoles simple, la production de points d’échantillonnages par des **protocoles probabilistes spatialement équilibrés**.

Les protocoles probabilistes spatialement équilibrés :
→génèrent des échantillons bien **répartis spatialement** sur la zone d’étude,
→sont plus **performants** = ↘ nombre d’ échantillons ↗ précision (comparé à un protocole aléatoire simple (**SRS**)).
Deux sont utilisés ici en exemples :
→le **GRTS** (Generalized Random Tessellation Stratified) –> un des plus utilisé,
→le **BAS** (Balanced Acceptance Sampling) –> capable d’intégrer plusieurs couches d’informations.

Palourdes japonaises

Cincle plongeur

« Spsurvey »

Spsurvey: Survey Design and Analysis
Version: 3.3
Depends: R (≥ 2.10), sp
Imports: methods, deldir, foreign, graphics, grDevices, Hmisc, MASS, rgeos, stats
Published: 2016-08-19
Author: Tom Kincaid [aut, cre], Tony Olsen [aut], Don Stevens [ctb], Christian Platt [ctb], Denis White [ctb], Richard Remington [ctb]

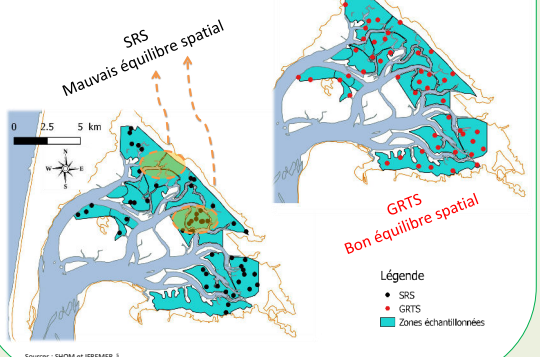
« SDraw »

SDraw: Spatially Balanced Sample Draws for Spatial Objects
Version: 2.1.3
Depends: R (≥ 2.10), sp
Imports: spsurvey, utils, rgeos, graphics, methods, deldir, stats
Published: 2016-06-11
Author: Trent McDonald

Application 1 : Campagne d’évaluation du stock de palourdes japonaises (*Venerupis philippinarum)* dans le Bassin d’Arcachon (France).

Exemple de tirage de points d’échantillonnage avec « Spsurvey », qui permet, entre autres, l’utilisation du **SRS** et du **GRTS.**

```
## Code GRTS avec Spsurvey
>library(Spsurvey)
>nrow<-50                          # Nombre d'échantillons voulus
>Equaldsgn list(None=list(panel=c(PanelOne=nrow),
        seltype="Equal")
)
>Equalsites <- grts(design=Equaldsgn,
        src.frame="shapefile",     # Couche géographique d'entrée
        in.shape="bassinarcachon",
        att.frame= read.dbf("bassinarcachon"),
        type.frame="area",
        DesignID="EQUAL",
        shapefile=TRUE,            # Sortie sous forme de couche géographique
        out.shape= "my shape"
)
>coord<-cbind(Equalsites$xcoord,Equalsites$ycoord)
>write.table(coord,file="GRTSbassinarcachon.txt")
```



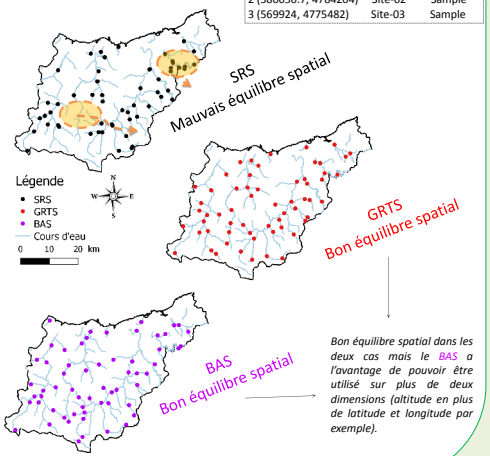
Sources : SHOM et IFREMER [5]

Application 2 : Suivi de la population de cincle plongeur (*Cinclus cinclus)* dans la province du Gipuzkoa (Espagne).

Exemple de tirage de points d’échantillonnage avec « SDraw». En plus du **SRS** et du **GRTS**, « SDraw » permet l’utilisation du **BAS.**

```
##Code SRS, GRTS et BAS avec SDraw
>library(SDraw)
>library(maptools)
>Shape<-readShapeSpatial("RiosGipuzkoa.shp")  # Couche géographique en entrée
>nb<- 60                          # Nombre d'échantillons voulus
>over.n<-0                        # Nombre de sur-échantillons voulus
>samples<-bas.line(Shape, nb, over.n) #ou grts.line() ou srs.line()
>write.table(samples,file="BASgipuzkoa.txt")
```

```
>head(samples[,1:2])
  coordinates              sampleID  pointType
1 (581033.7, 4771133)      Site-01   Sample
2 (580056.7, 4784204)      Site-02   Sample
3 (569924, 4775482)        Site-03   Sample
```



Bon équilibre spatial dans les deux cas mais le BAS a l’avantage de pouvoir être utilisé sur plus de deux dimensions (altitude en plus de latitude et longitude par exemple).

Sources : GeoEuskadi et ARANZADI [6]

Conclusions : ● Les packages présentés ici sont des outils de tirage de points d’échantillonnage aléatoires **gratuits** et **faciles** à utiliser. L’utilisateur doit seulement avoir un fichier shapefile (couche géographique) de sa zone d’étude et connaître le nombre de points d’échantillonnage qu’il souhaite sélectionner.

● Les protocoles spatialement équilibrés sont aussi faciles à mettre en œuvre que les protocoles non-spatialement équilibrés. Avec de tels outils, l’utilisation de protocoles d’échantillonnage probabilistes n’est plus aussi complexe et devient accessible pour les écologistes (et tout « enquêteur » au sens large).

Perspectives : Le domaine de l’échantillonnage assisté par R est en constante évolution. Par exemple, le nouveau package MBHdesign[4], récemment développé, permet au BAS d’intégrer des sites de suivi qui sont historiques.

1. SMITH, Adam NH, ANDERSON, Marti J et PAWLEY, Matthew DM. 2017. Could ecologists be more random? Straightforward alternatives to haphazard spatial sampling. *Ecography*.
2. KINCAID, Tom et OLSEN, Tony. *spsurvey: Spatial Survey Design and Analysis* [en ligne]. 2016. Disponible à l’adresse : https://CRAN.R-project.org/package=spsurvey
3. MCDONALD, Trent. *SDraw: Spatially Balanced Sample Draws for Spatial Objects* [en ligne]. 2016. Disponible à l’adresse : https://github.com/tmcd82070/SDraw/wiki/SDraw
4. FOSTER, Scott D, *MBHdesign: Spatial Designs for Ecological and Environental Surveys* [en ligne]. 2017. Disponible à l’adresse: https://CRAN.R-project.org/package=MBHdesign
5. SANCHEZ F, CAILL-MILLY N, DE CASAMAJOR MARIE-NOELLE LM. 2012. *Campagne d’évaluation du stock de palourdes du bassin d’Arcachon*. Ifremer, France
6. ARIZAGA J. , SÁNCHEZ J.M. & D’AMICO F. 2016. *Desarrollo, puesta a punto y método de censo del mirlo acuático (Cinclus cinclus) en la CAPV*. Departamento de Medio Ambiente, Planificación Territorial y Vivienda del Gobierno Vasco. http://www.euskadi.eus/gobierno-vasco/departamento-medio-ambiente-politica-territorial/inicio/ . 12 pp.

Appendix B : Optimal release of mosquitoes to control dengue transmission

ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2019, 1-10
Editors: Will be set by the publisher

OPTIMAL RELEASE OF MOSQUITOES TO CONTROL DENGUE TRANSMISSION *

Luis Almeida[1], Antoine Haddon[2,3], Claire Kermorvant[4], Alexis Léculier[5], Yannick Privat[6], Martin Strugarek[7], Nicolas Vauchelet[8] and Jorge P. Zubelli[9]

Abstract. In order to prevent the propagation of human diseases transmitted by mosquitoes (as dengue or zika), a solution is to release mosquitoes infected by *Wolbachia*. In this study, we model the release and the propagation over time and space of such infected mosquitoes in a population of uninfected ones. The aim of this study is to investigate the best location in space of the release to ensure invasion by the infected mosquitoes.

Résumé. Afin de prévenir la propagation de maladies transmises à l'homme par les moustiques (comme la dengue ou le zika), une solution consiste à relâcher des moustiques infectés par la bactérie *Wolbachia*. Dans cette étude, nous modélisons le relâcher et la propagation dans le temps et l'espace de ces moustiques infectés dans une population de moustiques hôtes non infectés. Le but de cette étude est d'étudier le meilleur emplacement dans l'espace des relâchers afin d'assurer l'invasion par les moustiques infectés.

Introduction

Aedes aegypti is the main vector transmitting dengue viruses. This mosquito can also transmit chikungunya, yellow fever and Zika infection. According to the World Health Organization, 390 million people are infected by

* *The authors acknowledge supports by the Project "Analysis and simulation of optimal shapes - application to life-sciences" of the Paris City Hall. N.V. and J.P.Z. acknowledge support from the France-Brasil network for mathematics.*

[1] Sorbonne Université, CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France. luis@ann.jussieu.fr

[2] MISTEA, Université Montpellier, INRA, Montpellier SupAgro, 2 pl. Viala, 34060 Montpellier, France. antoine.haddon@etu.umontpellier.fr

[3] Mathematical Engineering Department and Center for Mathematical Modelling (CNRS UMI 2807), Universidad de Chile, Beauchef 851, Santiago, Chile.

[4] CNRS, Univ Pau & Pays Adour, E2S UPPA, Laboratoire de Mathématiques et de leurs Applications de Pau, MIRA, UMR 5142, 64600, Anglet, France. claire.kermorvant@univ-pau.fr

[5] Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex 9. alexis.leculier@math.univ-toulouse.fr

[6] IRMA, Université de Strasbourg, CNRS UMR 7501, Équipe TONUS, 7 rue René Descartes, 67084 Strasbourg, France (yannick.privat@unistra.fr).

[7] Sorbonne Université, CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France. martin.strugarek@gmail.com

[8] LAGA, UMR 7539, CNRS, Université Paris 13 - Sorbonne Paris Cité, Université Paris 8, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse - France. vauchelet@math.univ-paris13.fr

[9] IMPA, Estrada Dona Castorina, 110 Jardim Botˆanico 22460-320 Rio de Janeiro, RJ - Brazil. zubelli@gmail.com

dengue every year and 3.9 billion people, in 128 countries, are at risk of infection by dengue viruses. As there is no treatment for dengue fever, the current method of preventing dengue virus transmission and epidemics is to target the vector, i.e. the mosquito. Beyond preventing mosquitoes from accessing egg-laying habitats by environmental management and modification, one of the most promising control techniques is to transform mosquito population with a virus-suppressing *Wolbachia* bacteria. The idea of using *Wolbachia* for disease control was first proposed in the 1960s [8] but applying it to *Aedes aegypti* population is very recent. *Wolbachia* bacterium strains were isolated from *Drosophila melanogaster* in laboratory just before 2000 [7, 10] but were introduced into *Aedes aegypti* embryos only on 2009 [11]. The capability of this bacteria to suppress dengue virus and other pathogens transmission by *Aedes aegypti* was shown in laboratory around 2010 [2,12,21]. It was also shown that this bacteria shortens life span [22] and most of the infected adults do not reach the infectious stage. But the most important modification induced by the bacteria is cytoplasmic incompatibility (CI) [11]. Cytoplasmic incompatibility is used by the bacteria to spread rapidly into natural population [18] by producing non-viable eggs when uninfected females mate with infected males. Reproduction between infected males and females lead to infected eggs. As this bacteria is vertically transmitted (from mother to off-springs), uninfected males mating with infected females give rise only to infected eggs.

We are interested on optimizing the release of *Wolbachia*-infected mosquitoes into a wild host population of mosquitoes. Thus, the aim of the study is to model the propagation across time and space of the density of infected mosquitos, denoted n_2, starting from a controlled release u into an existing population of uninfected. In what follows, we will denote by n_1 the density of uninfected mosquitos.

Formally, a proportion $1 - s_h$ of uninfected female's eggs fertilized by infected males actually hatch. Cytoplasmic incompatibility is perfect when $s_h = 1$. We denote by b_1, respectively b_2, the net fecundity rate of uninfected females, respectively infected females. Death rate for uninfected mosquitoes is denoted d_1. As *Wolbachia* decreases lifespan, death rate of infected mosquitoes d_2 verifies $d_2 > d_1$. Is is also observed that *Wolbachia* infected mosquitoes tend to have reduced fertility, then $b_2 \leq b_1$. Finally, we denote κ the carrying capacity. Cytoplasmic incompatibility and vertical transmission drive the spatial spread of the infected population producing a bistable dynamic of *Wolbachia* [19]. If the infected population is installed above a sufficient threshold frequency Θ compared to the uninfected population, it will spread and tend to increase to 1, otherwise it will tend to decline to zero.

For fixed maximal time $T > 0$ and domain Ω, the system of equation that we consider is the following:

$$\begin{cases} \partial_t n_1 - D\Delta n_1 = b_1 n_1 (1 - s_h \dfrac{n_2}{n_1+n_2})(1 - \dfrac{n_1+n_2}{\kappa}) - d_1 n_1 & \text{in } \Omega, \\ \partial_t n_2 - D\Delta n_2 = b_2 n_2 (1 - \dfrac{n_1+n_2}{\kappa}) - d_2 n_2 + u & \text{in } \Omega, \\ \partial_\nu n_1 = \partial_\nu n_2 = 0 & \text{on } \partial\Omega, \\ n_1(0,x) = n_1^0(x), \quad n_2(0,x) = n_2^0(x) & \text{in } \Omega. \end{cases} \tag{1}$$

The equations driving the dynamics of n_1 and n_2 are bistable and monostable reaction-diffusion equations, respectively. Note that in the reaction term of the first equation the term $-\frac{n_2}{n_1+n_2}$ stands for the vertical transition of the disease whereas the coefficient s_h models that this vertical transmission may or not be perfect because of the cytoplasmic incompatibility. More precisely, assuming homogeneous repartition of individuals, the probability to mate with an infected mosquito is $\frac{n_2}{n_1+n_2}$. Then, uninfected mosquitoes are generated from mating of uninfected mosquitoes with uninfected mosquito (probability $n_1 \frac{n_1}{n_1+n_2}$) or uninfected mosquitoes with infected mosquitoes but with a probability $(1 - s_h) n_1 \frac{n_2}{n_1+n_2}$. The first term in the right hand side is the sum of this latter quantities. The diffusion coefficient is denoted D; it is assumed to be the same for both populations since both populations belongs to the same genus of mosquitoes. The last term of the second equation $+u$ stands here to model the releases of infected mosquitoes developed in laboratory: it is on this control that we will act upon. More precisely, a question we want to address in this work is to know what should be the shape of the release function u to be as close as possible to the total invasion of the infected population into the domain.

The outline of this paper is the following. In the next section, we introduce the optimal control problem and prove the existence of an optimum for this problem. In Section 2, we consider a toy problem, which is a very simplified version of the full problem, for which we can solve explicitly the optimal problem and find the optimum. In Section 3, we investigate numerically the optimization of the spatial releases of mosquitoes. Finally, we end this paper with a conclusion and perspective for future works. An appendix is devoted to recalling the reduction of system (1).

1. Optimal Control Problem

We are going to simplify the problem. Instead of studying the coupled equations (1), we are going to follow the proportion of mosquitoes $p(t,x) = \frac{n_2(t,x)}{n_1(t,x)+n_2(t,x)}$ as in [15]. This reduction is clearly justified in the limit of large population in [15] (see also [1, Section 2.3]). Details on the formal computation are provided in Appendix A. In order to simplify the reading, we perform the scaling $x = \frac{\tilde{x}}{\sqrt{D}}$ not to keep the diffusion coefficient along the computations. Obviously, for the numerical simulations performed in Section 3, we have to keep in mind this scaling.

Denoting by p the proportion of infected mosquitoes, and u the release function, the dynamics is governed by the reaction-diffusion equation

$$\begin{cases} \dfrac{\partial p}{\partial t} - \Delta p = f(p) + ug(p), & t \in (0,T), \quad x \in \Omega, \\ \partial_\nu p(t,x) = 0, & x \in \partial\Omega, \\ p(0,x) = 0, & x \in \Omega, \end{cases} \tag{2}$$

where

$$f(p) = p(1-p)\frac{d_1 b_2 - d_2 b_1(1-s_h p)}{b_1(1-p)(1-s_h p) + b_2 p} \quad \text{and} \quad g(p) = \frac{1}{\kappa} \cdot \frac{b_1(1-p)(1-s_h p)}{b_1(1-p)(1-s_h p) + b_2 p}. \tag{3}$$

The general optimal control problem we want to investigate involves the least-squares functional J defined by

$$\hat{J}(u) = \frac{1}{2}\int_\Omega (1 - p(T,x))^2 \, dx, \tag{4}$$

which models that one aims at steering the system as close as possible to the target state. In some sense, it stands for the research of a control strategy ensuring the persistence of infected mosquitoes at the time horizon T.

Of course, it is relevant from the biological point of view to impose several constraints on the control function u. Indeed, the production of *Wolbachia*-infected mosquitoes is limited, which imposes that the total number of mosquitoes released is bounded. Hence, the control function u is assumed to belong to the set

$$\mathcal{U}_{T,C,M} = \left\{ u \in L^\infty([0,T]\times\Omega), \quad 0 \le u \le M \text{ a.e. }, \int_0^T \int_\Omega u(t,x)\, dxdt \le C \right\}. \tag{5}$$

modeling an upper limit on the instantaneous number of *Wolbachia*-infected individuals released at time t, as well as on the total number of released mosquitoes.

We then deal with the following optimal control problem:

$$\boxed{\inf_{u \in \mathcal{U}_{T,C,M}} \hat{J}(u).} \tag{$\mathcal{P}_{\text{full}}$}$$

Since this problem involves the minimization over function depending on time and space variables, it is difficult to study. Then, we will reduce it to a simpler one by assuming that the time distribution of the control function is given.

1.1. **Modeling of the optimal control problem**

In order to weaken the difficulty of Problem ($\mathcal{P}_{\text{full}}$), we introduce a simpler, although still relevant, problem by assuming that:

- releases are done periodically in time (for instance every week) and are impulses in time[1];
- at each release, the largest allowed amount of mosquitoes is released, corresponding to the maximal production capacity per week (which is relevant, according to the comparison principle).

As a consequence, we will be interested in determining the optimal way of releasing spatially the infected mosquitoes. Let us denote by $t_0 = 0 < t_1 < \ldots < t_N = T$, $t_i = i\Delta T$, the release times. Rewriting the L^1 constraint on the control as $\langle u, 1\rangle_{\mathcal{D}',\mathcal{D}((0,T)\times\Omega)} \leq C$, the control function reads

$$u(t,x) = \sum_{i=0}^{N-1} u_i(x)\delta_{\{t=t_i\}}, \quad \text{with } \sum_{i=0}^{N-1} \int_\Omega u_i(x)\,dx \leq C,$$

where the pointwise constraint is modified into $0 \leq u_i(\cdot) \leq M$.

The new optimal design problem reads

$$\inf_{\mathbf{u}\in\mathcal{V}_{T,C,M}} \tilde{J}(\mathbf{u}), \quad \text{where } \mathbf{u} = (u_i)_{0\leq i\leq N-1}, \quad \tilde{J}(\mathbf{u}) = J\left(\sum_{i=0}^{N-1} u_i(\cdot)\delta_{\{t=t_i\}}\right) \tag{$\mathcal{P}'_{\text{full}}$}$$

and

$$\mathcal{V}_{T,C,M} = \left\{ \mathbf{u} = (u_i(\cdot))_{0\leq i\leq N-1}, \quad 0 \leq u_i \leq M \text{ a.e. in } \Omega,\ i \in \{0,\ldots,N-1\},\ \sum_{i=0}^{N-1} \int_\Omega u_i(x)\,dx \leq C \right\}.$$

It is possible to recast System (2) without source measure terms, coming from the specific form of the control functions. For the sake of simplicity, we provide here a naive formal analysis, but claim that this can be proven rigorously by using a standard variational analysis.

Let us approximate the Dirac measure at $t = t_i$ by the function $\frac{1}{\varepsilon}\mathbf{1}_{[t_i,t_i+\varepsilon]}$. Making the change of variable $t = t_i + \tau\epsilon$, and introducing $\tilde{p}$ given by $\tilde{p}(\tau,x) = p(t,x)$, one gets from system (2) that $\tilde{p}$ solves

$$\frac{\partial\tilde{p}}{\partial\tau} - \varepsilon\Delta\tilde{p} = \varepsilon f(\tilde{p}) + u_i g(\tilde{p}). \qquad \tau \in [0,1],\ x \in \Omega.$$

Letting formally ε go to 0 and denoting, with a slight abuse of notation, still by $\tilde{p}$ the formal limit of the system above yields

$$\frac{\partial\tilde{p}}{\partial\tau}(\tau,x) = u_i(x)g(\tilde{p}(\tau,x)), \qquad \tau \in [0,1],\ x \in \Omega. \tag{6}$$

Let us denote G the anti-derivative of $\frac{1}{g}$ vanishing at 0, namely

$$G(p) = \int_0^p \frac{dq}{g(q)}.$$

Then, by a direct integration of (6) on $[0,1]$, we obtain

$$G(\tilde{p}(1,x)) = G(\tilde{p}(0,x)) + u_i(x), \qquad x \in \Omega.$$

[1]We consider Dirac measures since at the time-level of the study (namely, some generations), the release can be considered as instantaneous.

Coming back on the function p yields

$$p(t_i^+, x) = G^{-1}(G(p(t_i^-, x)) + u_i(x)), \qquad x \in \Omega.$$

Hence we arrive at the system

$$\begin{cases} \dfrac{\partial p}{\partial t} - \Delta p = f(p), & t \in (0,T) \setminus \{t_i\}_{i\in\{1,\dots,N-1\}}, \quad x \in \Omega, \\ \partial_\nu p(t,x) = 0, & x \in \partial\Omega, \\ p(0^+,\cdot) = G^{-1}(u_0(\cdot)), & \\ p(t_i^+,\cdot) = G^{-1}\left(G(p(t_i^-,\cdot)) + u_i(\cdot)\right), & i \in \{1,\dots,N-1\} \end{cases} \tag{7}$$

and the optimization problem reads

$$\boxed{\inf_{\mathbf{u}\in\mathcal{V}_{T,C,M}} J(\mathbf{u}) \quad \text{with } J(\mathbf{u}) = \frac{1}{2}\int_\Omega (1-p(T,x))^2\,dx}, \tag{$\mathcal{P}_{\text{reduced}}$}$$

where p is the solution of (7). In the next Section, we investigate the existence of solutions for this problem.

1.2. **Existence of minimizers**

Theorem 1.1. *Problem ($\mathcal{P}_{\text{reduced}}$) has a solution.*

Proof. For the sake of readability, we only provide the proof in the case $N = 2$. Indeed, there is no additional difficulty to deal with the general case whose proof follows exactly the same lines. The proof is divided into several steps.

Let $u^n = \{u_i^n\}_{i\in\{1,\dots,N\}} \in (\mathcal{V}_{T,C,M})^{\mathbb{N}}$ be a minimizing sequence for Problem ($\mathcal{P}_{\text{reduced}}$).

Notice that, since u belongs to $\mathcal{V}_{T,C,M}$ and G^{-1} takes its value in $[0,1[$, we infer from the maximum principle that $0 \le p(t,\cdot) < 1$ for a.e. $t \in [0,T]$ so that one has for all $u \in \mathcal{V}_{T,C,M}$

$$0 \le J(u) \le \frac{|\Omega|}{2}.$$

It follows that $\inf_{u\in\mathcal{V}_{T,C,M}} J(u)$ belongs to $(0, \frac{|\Omega|}{2})$ and, in particular, is finite.

Step 1: *Convergence of the minimizing sequence.*
Let p^n be the solution to (7) associated to the control function u^n and let us introduce

$$\begin{aligned} v_0^n(.) &= u_0^n(.) \\ v_1^n(.) &= G^{-1}(G(p^n(t_1^-,.)) + u_1^n(.)). \end{aligned}$$

By induction, one easily shows that v^n is uniformly bounded in L^∞. Since the class $\mathcal{V}_{T,C,M}$ is closed for the L^∞ weak-star topology, there exists $v^\infty \in \mathcal{V}_{T,C,M}$ such that, up to a subsequence, v^n converges weakly-star to v^∞ in L^∞. Here and in the sequel, we will denote similarly with a slight abuse of notation a given sequence and any subsequence.

Multiplying the main equation of (7) by p^n and integrating by parts, we infer from the above estimates the existence of a positive constant C such that

$$\frac{1}{2}\int_0^T\int_\Omega \partial_t(p^n(t,x)^2)dxdt + \int_0^T\int_\Omega |\nabla p^n(t,x)|^2 dxdt \le C$$

for every $n \in \mathbb{N}$, which also reads

$$\frac{1}{2}\int_\Omega \left(\left[(p^n(t,x))^2 \right)\right]_{t=0}^{t=t_1} + \left[(p^n(t,x))^2 \right)\right]_{t=t_1}^{t=T} \Big)\, dx + \int_0^T \int_\Omega |\nabla p^n(t,x)|^2\, dxdt \leq C$$

for every $n \in \mathbb{N}$.

By using the pointwise bounds on p^n, it follows that p^n is uniformly bounded in $L^2([0,T], H^1(\Omega))$. Furthermore, by using (7), one gets that the sequence $\partial_t p^n$ is uniformly bounded in $L^2([0,T], W^{-1,1}(\Omega))$. According to the Aubin-Lions theorem (see [14]) we infer that p^n converges (up to a subsequence) to $p^\infty \in L^2([0,T], H^1(\Omega))$, strongly in $L^2([0,T], L^2(\Omega))$ and weakly in $L^2([0,T], H^1(\Omega))$. Passing to the limit in (7) yields that p^∞ is a weak solution to

$$\begin{cases} \partial_t p^\infty - \Delta p^\infty = f(p^\infty), & t \in (0,T), \quad x \in \Omega, \\ \partial_\nu p^\infty(t,x) = 0, & t \in (0,T), \quad x \in \partial\Omega, \end{cases} \tag{8}$$

It is standard that any solution to this bistable reaction-diffusion equation is continuous in time.

Introducing $u_0^\infty := G(p^\infty(0^+,.))$ and $u_1^\infty := G(v_1^\infty) - G(p^\infty(t_1^-,.))$, one shows that $p^\infty(t_1^-,.) = v_1^\infty(.)$ by passing to the limit as $n \to +\infty$ in the variational formulation on p^n, using adapted test-functions belonging to

$$V_1 = \left\{ q \in C^\infty([0,T[, C^\infty(\Omega) \cap C^0(\bar{\Omega})) \text{ whose support is contained in } [t_i, t_{i+1}[\right\}.$$

This is a consequence of the weak convergence of p^n in $H^1(\Omega)$ to p^∞. Notice, in particular, that $G(v_1^n)$ converges weakly star in L^∞ to $G(v_1^\infty)$. Indeed, this is a consequence of the continuity and convexity since one has $G''(p) = \kappa b^2 \dfrac{(1 - s_h p^2)}{((p-1)(s_h p - 1))^2}$ which is positive whenever p belongs to $[0,1]$.

Step 2: *Conclusion.*
Let us first show that u^∞ belongs to $\mathcal{V}_{T,C,M}$. Since the derivative of G is $1/g$ which is positive, G is increasing and therefore, one has $0 \leq u^\infty \leq m$ a.e. in Ω.
For the integral condition (namely, $\int_\Omega u \leq C$), let us distinguish between two cases:

Case 1: if $m|\Omega| \leq C$, the conclusion follows immediately.

Case 2: if $m|\Omega| > C$, let us use that G is, as aforementioned, lower semi-continuous for the weak-star topology of L^∞. Thus, we deduce that

$$\int_\Omega u^\infty = \int_\Omega G(v^\infty) \leq \liminf_{n\to+\infty} \int_\Omega G(v^n) = \liminf_{n\to+\infty} \int_\Omega u^n \leq C.$$

It follows that u^∞ belongs to $\mathcal{V}_{T,C,M}$ and one concludes by using the Fatou Lemma:

$$\begin{aligned} J(u^\infty) = \frac{1}{2}\int_\Omega (1 - p^\infty(T,x))^2 dx &= \frac{1}{2}\int_\Omega \liminf_{n\to+\infty} (1 - p^n(T,x))^2 dx \\ &\leq \liminf_{n\to+\infty} \frac{1}{2}\int_\Omega (1 - p^n(T,x))^2 dx = \liminf_{n\to+\infty} J(u_n) = \inf_{u \in \mathcal{V}_{T,C,M}} J(u). \end{aligned}$$

We finally infer that u^∞ solves Problem $(\mathcal{P}_{\text{reduced}})$. □

Remark 1.2. The uniqueness issue remains open, even for simple domain. It is likely that symmetries of the release domain play an important role.

It is interesting to notice that, in a very particular case, we have an explicit expression of the minimizer for this problem.

Proposition 1.3. *Let $N \in \mathbb{N}^*$ and $M \leq \frac{C}{|\Omega|}$. Then $u = M$ is the unique solution of Problem $(\mathcal{P}_{\text{reduced}})$.*

Proof. It is a direct application of the comparison principle. Let u^* be a solution of Problem $(\mathcal{P}_{\text{reduced}})$. By contradiction, let us assume that u^* is not identically equal to M a.e. in Ω. Then, let t_i be a release time for which the associated control function u_i^* is not identically equal to M in Ω. Recall that $u_i^* \leq M$. Let us denote by $p*$ the solution of the problem (2) associated to the control function u^*. Let u^M be the control function defined by

$$u_i^M = M \quad \text{and} \quad u_j^M = u_j^* \quad \text{for all } j \in \{0, \dots, N-1\} \backslash \{i\}.$$

Let p^M be the solution of (2) associated to u^M identically. Since G^{-1} is an increasing function by the comparison principle we have for all time $t \in [0, T]$ and a.e. $x \in \Omega$,

$$0 < p^*(t, x) \leq p^M(t, x) < 1.$$

Evaluating this expression at time $t = T$, the expected conclusion follows by noting that the constant function equal to M on $(0, T) \times \Omega$ belongs to $\mathcal{U}_{T,C,M}$. □

1.3. **Computation of derivatives**

As a preliminary remark, we claim that for any element $\mathbf{u}$ of the set $\mathcal{V}_{T,C,M}$ and any admissible perturbation $\mathbf{h}$, the mapping $\mathcal{V}_{T,C,M} \ni \mathbf{u} \mapsto p \in L^2(0, T, H^1(\Omega))$, where p denotes the unique weak solution of (7), is differentiable in the sense of Gâteaux at $\mathbf{u}$ in the direction $\mathbf{h}$. Indeed, proving such a property is standard in calculus of variations and rests upon an application of the implicit function theorem. In the sequel, and with no confusion possible, we will denote by $\dot{p}$ the Gâteaux-differential of p at $\mathbf{u}$ in direction $\mathbf{h}$ and by $\langle dJ(\mathbf{u}), \mathbf{h} \rangle$ the Gâteaux-differential of J at $\mathbf{u}$ in direction $\mathbf{h}$, namely

$$\langle dJ(\mathbf{u}), \mathbf{h} \rangle = \lim_{\varepsilon \searrow 0} \frac{J(\mathbf{u} + \varepsilon \mathbf{h}) - J(\mathbf{u})}{\varepsilon}.$$

Let us make the cone of admissible perturbations precise. We call "admissible perturbation" any element of the tangent cone $\mathcal{T}_{\mathbf{u}, \mathcal{V}_{T,C,M}}$ to the set $\mathcal{V}_{T,C,M}$ at $\mathbf{u}$.

Definition 1.4. The cone $\mathcal{T}_{\mathbf{u}, \mathcal{V}_{T,C,M}}$ is the set of N-tuples $\mathbf{h} = (h_0, \dots, h_{N-1}) \in (L^\infty(\Omega))^N$ such that, for any $i \in \{0, \dots N-1\}$ and for any sequence of positive real numbers ε_n decreasing to 0, there exists a sequence of functions $h_i^n \in L^\infty(0, T)$ converging to h_i as $n \to +\infty$, and $u_i + \varepsilon_n h_i^n \in \mathcal{V}_{T,C,M}$ for every $n \in \mathbb{N}$ (see e.g. [5]).

Proposition 1.5. *Assume that $N = 1$. Let $\mathbf{u} = (u_0) \in \mathcal{V}_{T,C,M}$ and $\mathbf{h} = (h_0) \in \mathcal{T}_{\mathbf{u}, \mathcal{V}_{T,C,M}}$. One has*

$$\langle dJ(\mathbf{u}), \mathbf{h} \rangle = \int_\Omega h(x) (G^{-1})'(u_0(x)) q(0, x) \, dx,$$

where q is the unique solution of the backward problem

$$\begin{cases} -\partial_t q(t, x) - \Delta q(t, x) - f'(p(t, x)) q(t, x) = 0, & (t, x) \in (0, T) \times \Omega, \\ \partial_n q(t, x) = 0, & (t, x) \in (0, T) \times \partial\Omega, \\ q(T, x) = p(T, x) - 1, & x \in \Omega. \end{cases}$$

Proof. By using the preliminary discussion, one has

$$\langle dJ(\mathbf{u}), \mathbf{h} \rangle = \int_\Omega \dot{p}(T, x)(p(T, x) - 1) \, dx, \tag{9}$$

where $\dot{p}$ denotes the unique solution of the system

$$\begin{cases} \dfrac{\partial \dot{p}}{\partial t} - \Delta \dot{p} = f'(p)\dot{p}, & t \in (0,T), \quad x \in \Omega, \\ \partial_\nu \dot{p}(t,x) = 0, & x \in \partial\Omega, \\ \dot{p}(0^+,\cdot) = (G^{-1})'(u_0(\cdot))h. \end{cases} \tag{10}$$

Let us multiply the main equation of this system by q and then integrate by parts with respect to the variables t and x. By using in particular the Green formula, we get successively that

$$\begin{aligned} \int_0^T \int_\Omega q \frac{\partial \dot{p}}{\partial t}\, dxdt &= -\int_0^T \int_\Omega \dot{p} \frac{\partial q}{\partial t}\, dxdt + \int_\Omega q(T,x)\dot{p}(T,x)\, dx - \int_\Omega q(0,x)\dot{p}(0^+,x)\, dx, \\ -\int_0^T \int_\Omega q\Delta \dot{p}\, dxdt &= -\int_0^T \int_\Omega \dot{p}\Delta q\, dxdt, \end{aligned}$$

and therefore,

$$\langle dJ(\mathbf{u}), \mathbf{h}\rangle = \int_\Omega q(T,x)\dot{p}(T,x)\, dx = \int_\Omega q(0,x)\dot{p}(0^+,x)\, dx,$$

yielding the desired conclusion. □

Remark 1.6. For practical purposes, it may be useful to notice that

$$q(t,x) = \widetilde{q}(T-t,x), \qquad t \in [0,T],\ x \in \Omega,$$

where $\widetilde{q}$ denotes the solution of the initial boundary value problem

$$\begin{cases} \partial_t \widetilde{q}(t,x) - \Delta \widetilde{q}(t,x) - f'(p(T-t,x))\widetilde{q}(t,x) = 0, & (t,x) \in (0,T)\times\Omega, \\ \partial_n \widetilde{q}(t,x) = 0, & (t,x) \in (0,T)\times\partial\Omega, \\ \widetilde{q}(0,x) = p(T,x) - 1, & x \in \Omega. \end{cases}$$

2. A toy Problem

This section is devoted to investigating a simpler version of $(\mathcal{P}_{\text{full}})$ corresponding to the case $N = 1$ with $f = 0$. More precisely, let p be the solution of

$$\begin{cases} \dfrac{\partial p}{\partial t} - \Delta p = 0, & t \in (0,T), \quad x \in \Omega, \\ \partial_\nu p(t,x) = 0, & x \in \partial\Omega, \\ p(0^+,\cdot) = u_0(\cdot). \end{cases} \tag{11}$$

Then, the optimization toy problem reads

$$\boxed{\inf_{u_0 \in \mathcal{V}_{T,C,M}} \hat{J}(u_0) \quad \text{with } \hat{J}(u_0) = \frac{1}{2}\int_\Omega (1 - p(T,x))^2\, dx}, \tag{$\mathcal{P}_{\text{toy}}$}$$

where $p \in C^0([0,T], L^2(\Omega))$ is the unique solution of Equation (11). Note that the equation (11) has to be understood in a weak sense, since $u_0 \in L^\infty(\Omega) \subset L^2(\Omega)$ (see for example [17, Section 10.7]).

For this simple problem, we are able to solve explicitely the optimization problem :

Theorem 2.1. *Problem* (11) *has a unique solution* u_0*, which is constant and equal to* $\min\left(1, M, \frac{C}{|\Omega|}\right)$.

Proof. First, note that Problem $(\mathcal{P}_{\text{toy}})$ has a solution. Indeed, it is standard that the mapping $L^2(\Omega) \ni u_0 \mapsto p \in C^0([0,T], L^2(\Omega))$ is continuous. Therefore, so is $\hat{J}$ by composition of continuous mappings. The conclusion follows by observing that $\mathcal{V}_{T,C,M}$ is a compact subset of $L^2(\Omega)$.

The proof relies on a well-adapted rewriting of the criterion $\hat{J}$. For that purpose, let us introduce the Neumann operator $-\Delta_N$ on Ω defined on

$$\mathcal{D}(-\Delta_N) = \{y \in H^2(\Omega) \mid \frac{\partial y}{\partial n}_{|\partial\Omega} = 0 \text{ and } \int_\Omega y(x)\,dx = 0\}.$$

According to the spectral theorem, there exists an orthonormal family $(\phi_j)_{j\geq 1}$ consisting of (real-valued) eigenfunctions of $-\Delta_N$, associated with the non-decreasing sequence positive eigenvalues $(\lambda_j)_{j\geq 1}$. Moreover, by setting $\lambda_0 = 0$ and $\phi_0 = \frac{1}{\sqrt{|\Omega|}}$, the sequence $(\phi_j)_{j\geq 0}$ is a Hilbert basis of $L^2(\Omega)$ and any solution p of (11) can be expanded in a unique way in $L^2(\Omega)$ as

$$p(t,x) = \sum_{j=0}^{+\infty} \langle p(0,\cdot), \phi_j\rangle_{L^2(\Omega)} e^{-\lambda_j t}\phi_j(x) = \sum_{j=0}^{+\infty} u_{0j} e^{-\lambda_j t}\phi_j(x), \tag{12}$$

with $u_{0j} = \langle u_0, \phi_j\rangle_{L^2(\Omega)}$. By expanding the square in the definition of $\hat{J}$, we then infer that

$$\begin{aligned}
\hat{J}(u_0) &= \frac{|\Omega|}{2} - \int_\Omega p(T,x)\,dx + \frac{1}{2}\int_\Omega p(T,x)^2\,dx \\
&= \frac{|\Omega|}{2} - \sqrt{|\Omega|}u_{00} + \frac{1}{2}\sum_{j=0}^{+\infty} e^{-2\lambda_j T} u_{0j}^2 \\
&= \frac{|\Omega|}{2} - \int_\Omega u_0(x)\,dx + \frac{1}{2}\sum_{j=0}^{+\infty} e^{-2\lambda_j T}\left(\int_\Omega u_0(x)\phi_j(x)\,dx\right)^2.
\end{aligned}$$

Let u be a solution of Problem (11) and $h \in \mathcal{T}_{u_0,\mathcal{V}_{T,C,M}}$. Then, one has

$$\begin{aligned}
\langle d\hat{J}(u_0), h\rangle &= -\int_\Omega h(x)\,dx + \sum_{j=0}^{+\infty} e^{-2\lambda_j T}\left(\int_\Omega u_0(x)\phi_j(x)\,dx\right)\left(\int_\Omega h(x)\phi_j(x)\,dx\right) \\
&= \int_\Omega h(x)\psi(x)\,dx,
\end{aligned}$$

where $\psi(x) = -1 + \sum_{j=0}^{+\infty} e^{-2\lambda_j T} u_{0j}\phi_j(x)$.

The first order optimality conditions reads

$$\langle d\hat{J}(u_0), h\rangle \geq 0, \qquad \forall h \in \mathcal{T}_{u_0,\mathcal{V}_{T,C,M}}. \tag{13}$$

The analysis of such optimality condition is standard in optimal control theory (see for example [9]) and yields the existence of a Lagrange multiplier $\xi \leq 0$ such that

- on $\{u_0 = M\}$, $\psi(x) \leq \xi$,
- on $\{u_0 = 0\}$, $\psi(x) \geq \xi$,
- on $\{0 < u_0 < M\}$, $\psi(x) = \xi$,
- $\xi\left(\int_\Omega u_0(x)\,dx - C\right) = 0$ (complementarity condition).

Let us investigate the optimality of constant functions. To this aim, notice that the functional $\hat{J}$ is strictly convex[2]. It follows that the optimality conditions (13) are at the same time necessary and sufficient and that Problem (11) has a unique solution.

Let u_0 be an admissible constant function for Problem $(\mathcal{P}_{\text{toy}})$. Then, $u_0 \in [0, M]$ whenever $M \leq C/|\Omega|$ and $u_0 \in [0, C/|\Omega|]$ elsewhere.

Furthermore, if u_0 is constant, then,

$$\psi(x) = -1 + u_0.$$

Let us now investigate each case separately. If $u_0 = 0$, then, from the complementarity condition, $\xi = 0$ and $\psi(x) = -1$ which is in contradiction with the optimality conditions above. Let us assume that $u_0 \neq 0$.

- If $u_0 = \frac{C}{|\Omega|}$, then this is admissible only if $C \leq M|\Omega|$. In this case we find $\psi(x) = \frac{-|\Omega|+C}{|\Omega|}$, and thus the optimality conditions are satisfied if and only if $|\Omega| \geq C$. All in all, $u_0 = \frac{C}{|\Omega|}$ is indeed a solution if, and only if, $\min(1, M)|\Omega| \geq C$.
- If $u_0 \neq \frac{C}{|\Omega|}$: then $\xi = 0$. Either $u_0 = M$, in which case the optimality conditions are satisfied only if $M \leq 1$, and this solution is admissible only if $M|\Omega| \leq C$; or $0 < u_0 < M$, in which case the optimality conditions hold only if $u_0 = 1$ (since $\psi \equiv \xi = 0$ in this case), which is admissible only if $M \geq 1$ and $|\Omega| \leq C$. All in all, $u_0 = \min(1, M)$ is a solution if and only if $\min(1, M)|\Omega| \leq C$.

The conclusion follows.

□

3. Gaussian Releases

From a practical point of view, not all controls $u \in \mathcal{V}_{T,C,M}$ correspond to a release that could actually be conducted, as for example the constant solution of the toy problem of the previous section. To guarantee a solution that could be implemented, we restrict here the admissible controls to more accurately model the way mosquitoes are released in practice.

We thus consider that there are $K \in \mathbb{N}$ simultaneous releases and that each one results in a Gaussian distribution of mosquitoes centered around the position of the release $x_k \in \Omega$ for $k = 1, ..., K$. Then, the feasible controls are of the form

$$u_K(x, x_1, ..., x_K) = \sum_{k=1}^{K} m \exp\left(-\frac{\|x - x_k\|^2}{\sigma^2}\right), \tag{14}$$

where the constants m and σ are chosen such that $u_K(\cdot, x_1, ..., x_K) \in \mathcal{V}_{T,C,M}$. In particular, we choose to saturate the constraint on the total number of mosquitoes released, i.e. we take $\int_\Omega u_K(x)dx = C$.

The goal is then to find the best position of the releases and the optimization problem becomes

$$\boxed{\inf_{(x_1,...,x_K)\in\Omega^K} J_K(x_1, ..., x_K) \quad \text{with } J_K(x_1, ..., x_K) = \frac{1}{2}\int_\Omega (1 - p(T, x))^2\, dx}, \tag{$\mathcal{P}_K$}$$

where $p \in C^0([0, T], L^2(\Omega))$ is the unique solution of (7) with control $u_K(\cdot, x_1, ..., x_K)$.

Remark 3.1. Since Ω is a bounded domain in $\mathbb{R}^2$, the question of the existence of a minimizer is trivial. But, the uniqueness is still a challenging problem.

[2]The convexity results from the convexity of the square function combined with the linearity of $u_0 \mapsto p(T, \cdot)$. Furthermore,

$$\langle d^2\hat{J}(u_0), h, h\rangle = \sum_{j=0}^{+\infty} e^{-2\lambda_j T}\left(\int_\Omega h(x)\phi_j(x)\, dx\right)^2 \geq 0$$

and vanishes if, and only if, $\int_\Omega h(x)\phi_j(x)\, dx = 0$ for all j, meaning that $h = 0$ since $(\phi_j)_{j\geq 1}$ is a Hilbert basis of $L^2(\Omega)$. The strict convexity of $\hat{J}$ follows.

Proposition 3.2. *Let $(x_1, ..., x_K) \in \Omega^K$. For $k \in \{1, ..., K\}$, one has*

$$\frac{\partial J_K}{\partial x_k}(x_1, ..., x_K) = \int_\Omega (G^{-1})'(u_K(x))q(0,x)\frac{\partial u_K}{\partial x_k}(x, x_1, ..., x_K)\, dx,$$

where q is the unique solution of the backward problem

$$\begin{cases} -\partial_t q(t,x) - \Delta q(t,x) - f'(p(t,x))q(t,x) = 0, & (t,x) \in (0,T) \times \Omega, \\ \partial_n q(t,x) = 0, & (t,x) \in (0,T) \times \partial\Omega, \\ q(T,x) = p(T,x) - 1, & x \in \Omega. \end{cases}$$

Proof. It is an easy application of the chain rule. First, we notice that

$$J_K(x_1, ..., x_K) = J(u_K(x, x_1, ..., x_K)).$$

Next, using Proposition 1.5, we find thanks to the chain rule that for all $k \in \{1, ..., K\}$

$$\begin{aligned} \nabla J_K(x_1, ..., x_K) &= < dJ(u(x, x_1, ..., x_K)), \nabla u(x, x_1, ...x_K) > \\ &= \int_\Omega (G^{-1})'(u_K(x, x_1, ..., x_K)q(0,x)\nabla u_K(x, x, x_1, ..., x_K)dx. \end{aligned}$$

We deduce the result from the last equality. □

3.1. **Numerical Resolution**

We now present the computation of the numerical solution of $(\mathcal{P}_K)$. For this we use a direct method which consists in carrying out a discretization of Equation (7) and of the control in order to obtain a finite dimensional optimization problem with constraints. We can then compute an approximation of a local minimizer of $(\mathcal{P}_K)$ with a numerical optimization solver. Our results were obtained with the finite element toolbox *FreeFem++* [6] which contains an implementation of the optimization routine *Ipopt* [20].

We therefore consider a finite element basis of functions $(\varphi_i)_i$ that allows us to discretize the control as $u_h(x, x_1, ..., x_K) = \sum_i u_i\varphi_i(x)$ and the proportion of infected mosquitoes as $p_h(t,x) = \sum_i p_i(t)\varphi_i(x)$, the finite element approximation of the solution of the PDE (7) with initial condition $G^{-1}(u_h(x, x_1, ..., x_K))$. The cost function can be computed with numerical integration as $J_h(x_1, ..., x_K) = \int_\Omega (1 - p_h(T,x))^2\, dx$. In addition, *Ipopt* requires the gradient of the cost function and thanks to Proposition 3.2 we have

$$\frac{\partial J_h}{\partial x_k} = \int_\Omega (G^{-1})'(u_h(x))q_h(0,x)\frac{\partial u_h}{\partial x_k}(x, x_1, ..., x_K)\, dx$$

where $q_h(0,x)$ is the finite element approximation of the solution of the backwards PDE.

Remark 3.3. Because of Proposition 1.3, we were interested in the case $M > \frac{C}{|\Omega|}$. In addition, we have fixed C such that the constant solution $u = \frac{C}{|\Omega|}$ leads to extinction (as T tends to $+\infty$) but there exists $R \in]0, \sqrt{\frac{C}{\pi M}}[$ such that the function $u(x) = M \times 1_{B(0,R)}(x)$ belongs to $\mathcal{V}_{T,C,M}$ and leads to invasion (as T tends to $+\infty$).

We now present numerical simulations for the parameters given in Table 1. The birth and death rates are given per day, whereas the unit of the carrying capacity is per m^2 and the diffusion coefficient is given per m^2 per day. The numerical values are taken from [1] and references therein. We consider a square domain of 1 hectare, a final time of 200 days and we set the total amount of mosquitoes released such that $C < G(\theta)|\Omega|$. In Figure 1 we show the control $u_K(\cdot, x_1, ..., x_K)$ for $K = 3, 4, 5, 6$ releases and for each case the same total amount of mosquitoes is released. For the case of 6 releases we display in Figure 2 the time dynamics of the proportion of infected mosquitoes $p(t, \cdot)$. As expected, it leads to the total invasion of the domain.

Parameter	b_1	b_2	d_1	d_2	κ	D
Value	1.12	1.12	0.27	0.36	$6 \cdot 10^{-2}$	2.5

TABLE 1. Model parameters

Our simulations seem to be very sensitive on the initial data given to *Ipopt*. Indeed, for most choices of initial datum in the optimization algorithm, the best solution provided by *Ipopt* has the "same shape" as the initial datum (more precisely, by assimilating the Gaussian releases to domains, the optimal solution seems to have the same number of connected components as the initialization). Heuristically, this suggest that the function J_K we aim at minimizing mainly penalizes a lot the final time (here $T = 200$) and does not take into account what occurs at intermediate times. Since most of the initial data lead to invasion with the set of parameter we considered (in other words, the global minimum of J_K is almost reached), the considered interior points algorithm (via the software *Ipopt*) tries some new configurations relatively close to the initial data to find out that it was already an "almost" global minimum of J_K. We have tried unsuccessfully to make tests with a lower final time, and the results are similar. In a future work, to avoid such bad boundary effects, we foresee to consider another functional J_K taking into account not only the final time but also several intermediate times.

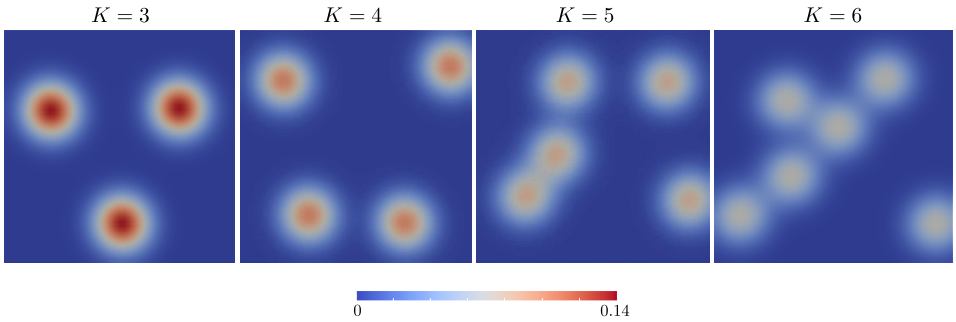


FIGURE 1. $u_K(\cdot, x_1, ..., x_K)$ for $K = 3, 4, 5, 6$ releases and $C = 0.017$.

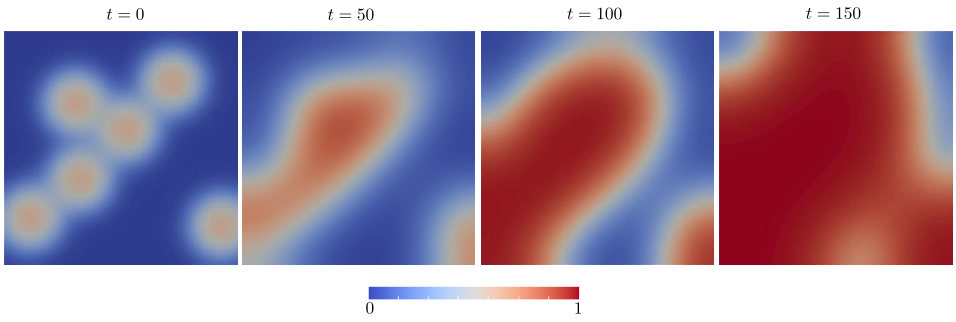


FIGURE 2. $p(t, \cdot)$ for $K = 6$ releases, $t = 0, 50, 100, 150$ days and $C = 0.017$.

4. Conclusion

We investigate in this work the optimization of the release of *Wolbachia*-infected mosquitoes into a host population in the aim to replace the wild population by a *Wolbachia*-infected population unable to transmit several diseases to human. To conduct this study, we first reduce the optimal problem under investigation by assuming that the time distribution is given. Then we obtain existence of a minimum for this latter problem. Finally, reducing again the control problem by considering that the releases are modeled by Gaussian distributions, some numerical computations are performed.

Optimization strategies for release protocols of mosquitoes have been investigated by several authors [3,4,16]. However, in these papers, only the time optimization of the releases is investigated. Up to our knowledge, this work is the first attempt in optimizing spatially the releases, which is of great interest for experiments in the field. The preliminary results obtained in this paper should be continued. In particular, the optimality conditions for the system $(\mathcal{P}_K)$ should be studied in a future work in the aim to find properties of the optimal solution. The numerical simulations should also be continued to have a better representation of what is observed in the field.

A. Appendix – Reduction of system (1)

For the sake of completeness and for reader facility, we explain briefly in this appendix how to reduce system (1) to system (2). We will not provide all the details of this reduction but only the main steps. We refer to [15] and [1, Section 2.3] for the interested reader. The starting point is to introduce a small parameter $0 < \varepsilon \ll 1$ modeling the ratio of the fertility on the death rate. Indeed for mosquitoes population, the fertility is large compared to death rates. System (1) reads then

$$\begin{cases} \partial_t n_1^\varepsilon - D\Delta n_1^\varepsilon = \dfrac{b_1}{\varepsilon} n_1^\varepsilon (1 - s_h \dfrac{n_2^\varepsilon}{n_1^\varepsilon + n_2^\varepsilon})(1 - \dfrac{n_1^\varepsilon + n_2^\varepsilon}{\kappa}) - d_1 n_1^\varepsilon, \\ \partial_t n_2^\varepsilon - D\Delta n_2^\varepsilon = \dfrac{b_2}{\varepsilon} n_2^\varepsilon (1 - \dfrac{n_1^\varepsilon + n_2^\varepsilon}{\kappa}) - d_2 n_2^\varepsilon + u. \end{cases}$$

As $\varepsilon \to 0$, we expect from this system that $n_1^\varepsilon + n_2^\varepsilon \to \kappa$. Hence we introduce the quantity $n^\varepsilon = \frac{1}{\varepsilon}(1 - \frac{n_1^\varepsilon + n_2^\varepsilon}{\kappa})$ and denote $p^\varepsilon = \frac{n_2^\varepsilon}{n_1^\varepsilon + n_2^\varepsilon}$ the proportion of infected mosquitoes. From straightforward computations, we deduce the system satisfied by $(n^\varepsilon, p^\varepsilon)$:

$$\begin{cases} \partial_t n^\varepsilon - D\Delta n^\varepsilon = \dfrac{1}{\varepsilon}\left((1 - \varepsilon n^\varepsilon) a(p^\varepsilon)(Z(p^\varepsilon) - n^\varepsilon) - \dfrac{u}{\kappa}\right), \\ \partial_t p^\varepsilon - D\Delta p^\varepsilon + \dfrac{2\varepsilon}{1 - \varepsilon n^\varepsilon} \nabla p^\varepsilon \cdot \nabla n^\varepsilon = p^\varepsilon (1 - p^\varepsilon)(n^\varepsilon (b_2 - b_1(1 - s_h p^\varepsilon)) + d_1 - d_2) + \dfrac{u(1 - p^\varepsilon)}{\kappa(1 - \varepsilon n^\varepsilon)}, \end{cases}$$

where we use the notations $a(p) = b_1(1-p)(1 - s_h p) + b_2 p > 0$ and $Z(p) = \frac{d_1(1-p) + d_2 p}{a(p)} > 0$. Assuming that the sequences $(n^\varepsilon)_\varepsilon$ and $(p^\varepsilon)_\varepsilon$ admit limits denoted n and p respectively, we deduce from the first equation that, formally,

$$n = Z(p) - \frac{u}{\kappa} a(p). \tag{15}$$

Passing into the limit into the equation satisfied by p, we get

$$\partial_t p - D\Delta p = p(1-p)(n(b_2 - b_1(1 - s_h p)) + d_1 - d_2) + \frac{u}{\kappa}(1 - p),$$

Injecting the expression of n (15) into this latter equation, we recover the equation

$$\partial_t p - D\Delta p = f(p) + u g(p),$$

with f and g defined in (3)

References

[1] L. Almeida, Y. Privat, M. Strugarek, N. Vauchelet, *Optimal releases for population replacement strategies, application to Wolbachia*, Preprint HAL (2018).

[2] G. Bian, Y. Xu, P. Lu, Y. Xie, Z. Xi, *The endosymbiotic bacterium Wolbachia induces resistance to dengue virus in Aedes aegypti*, PLoS pathogens, **4** (2010).

[3] P.-A. Bliman, *Feedback Control Principles for Biological Control of Dengue Vectors*, preprint.

[4] D. E. Campo-Duarte, O. Vasilieva, D. Cardona-Salgado, M. Svinin, *Optimal control approach for establishing wMelPop Wolbachia infection among wild Aedes aegypti populations*, Journal of mathematical biology, **76** (2018), 1907–1950.

[5] R. Cominetti and J.-P. Penot. Tangent sets to unilateral convex sets. *C. R. Acad. Sci. Paris Sér. I Math.*, 321(12):1631–1636, 1995.

[6] F. Hecht, *New development in FreeFem++*, J. Numer. Math. **20** (2012), no. 3-4, 251–265.

[7] M. Kyung-Tai and B. Seymour, *Wolbachia, normally a symbiont of Drosophila, can be virulent, causing degeneration and early death*, Proceedings of the National Academy of Sciences, **20** (1997).

[8] H. Laven, *Eradication of Culex pipiens fatigans through cytoplasmic incompatibility*, Nature, **5113** (1967).

[9] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim. **29** (1991), no. 4, 895–908.

[10] C.J. McMeniman, A.M. Lane, A.W.C Fong, D.A. Voronin, I. Iturbe-Ormaetxe, R. Yamada, E.A. McGraw, S.L. O'Neill, *Host adaptation of a Wolbachia strain after long-term serial passage in mosquito cell lines*, Applied and environmental microbiology, **22** (2008).

[11] C.J. McMeniman, R.V. Lane, B.N. Cass, A.W.C. Fong, M. Sidhu, Y.F. Wang, S.L. O'neill, *Stable introduction of a life-shortening Wolbachia infection into the mosquito Aedes aegypti* Science, **5910** (2009).

[12] L.A. Moreira, I. Iturbe-Ormaetxe, J.A. Jeffery, G. Lu, A.T. Pyke, L.M. Hedges, B.C. Rocha, S. Hall-Mendelin, A. Day, M. Riegler, *A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya, and Plasmodium*, Cell, **7** (2009).

[13] E. Nelson, *Analytic vectors*, Ann. Math. **70** (1959), no. 3, 572–615.

[14] J. Simon, *Compact sets in the space $L^p(0,T;B)$* Ann. Mat. Pura Appl. (4) (1987).

[15] M. Strugarek, N. Vauchelet, *Reduction to a single closed equation for 2 by 2 reaction-diffusion systems of Lotka-Volterra type*, SIAM J. Appl. Math. **76** (2016) no 5, 2068–2080.

[16] R.C.A. Thome, H.M. Yang, and L. Esteva, *Optimal control of Aedes aegypti mosquitoes by the sterile insect technique and insecticide*, Math. Biosci. **223** (2010), pp. 12–23.

[17] M. Tucsnak, G. Weiss, *Observation and control for operator semi-groups*, Birkhäuser Verlag, Basel (2009).

[18] M. Turelli, *Cytoplasmic incompatibility in populations with overlapping generations*, Evolution: International Journal of Organic Evolution (2010).

[19] M. Turelli, N.H. Barton, *Deploying dengue-suppressing Wolbachia: robust models predict slow but effective spatial spread in Aedes aegypti*, Theoretical population biology (2017).

[20] A. Wächter, L. T. Biegler, *On the implementation of Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming*, Mathematical Programming **106** (2006), no. 1, 25–57.

[21] T.J.P.H. Walker, P.H. Johnson, L.A. Moreira, I. Iturbe-Ormaetxe, F.D. Frentiu, C.J. McMeniman, Y.S. Leong, Y. Dong, J. Axford, P. Kriesner, *The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations*, Nature, **7361** (2011).

[22] H.L. Yeap, P. Mee, T. Walker, A.R. Weeks, S.L. O'Neill, P. Johnson, S.A. Ritchie, K.M. Richardson, C. Doig, N.M. Endersby, *Dynamics of the'popcorn'Wolbachia infection in outbred Aedes aegypti informs prospects for mosquito vector control*, Genetics (2010).

Appendix C : Optimal sampling design to survey riparian bird populations with low detection probability

Frank J.N. D'Amico, Claire Kermorvant, José M.Sánchez, Juan Arizaga. "Optimal sampling design to survey riparian bird populations with low detection probability". Submitted on Bird Study/Ringing & Migration

Abstract :

Linear censusing and occupancy models based on fixed sampling points are alternative widely used technique to determine bird densities in riparian ecosystems, although it cannot be always properly executed. The aim of the present article is to assess the census efficiency for river birds using occupancy models in contexts of impaired visibility owing to dense vegetation along the banks. We tested whether increasing sampling periods within each survey unit (point) at occupancy models would result in increasing detection probability values. We used two approaches in order to identify the "best" design for dippers along forested river stretches : minimizing survey effort of standard single-season site occupancy modeling and exploratory power analysis. With a detection probability of 0.26 (i.e. much lower than in previous studies), a design with 60 sites surveyed 10 minutes 6 times a year would be the option to survey Dippers in forested habitats if an acceptable power is required. Simulations further revealed the consistency of the results. A strength of our study was to first select the survey sites using a recent advanced probabilistic approach (here GRTS) providing, among other major interests, a spatially balanced geographic coverage. We provide guidelines to establish a cost-effective survey design for a long-term monitoring citizen-based program of a white-throated dipper (*Cinclus cinclus*) population when detection probabilities are low.

Keywords

GRTS ; power analysis ; site-occupancy models ; spatial balance ; white-throated dipper (*Cinclus cinclus*).