Optimisation du parcours intra-parcellaire pour l'échantillonnage en production végétale

Baptiste Oger

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Génie des procédés

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau
Portée par l'Université de Montpellier

Unité de recherche ITAP – Information, Technologies, Analyse environnementale, Procédés agricoles
Unité de recherche MISTEA – Mathématiques, Informatique et Statistique pour l'Environnement et l'Agronomie

Optimisation du parcours intra-parcellaire pour l'échantillonnage en production végétale

Présentée par Baptiste OGER
Le 27 Novembre 2020

Sous la direction de Bruno TISSEYRE
Et Philippe VISMARA

Devant le jury composé de

Jean-Pierre DA COSTA, Professeur, Bordeaux Sciences Agro, France	Rapporteur
Gonzaga SANTESTEBAN, Associate professor, Université publique de Navarre, Espagne	Rapporteur
Gilles TROMBETTONI, Professeur, Université de Montpellier, France	Président du jury
Aurélie METAY, Maitre de conférence, L'institut Agro | Montpellier Supagro, France	Examinatrice
Harold CLENET, Enseignant-chercheur, Ecole d'ingénieurs de Purpan, France	Examinateur
Bruno TISSEYRE, Professeur, L'institut Agro | Montpellier Supagro, France	Directeur de thèse
Philippe VISMARA, Professeur, L'institut Agro | Montpellier Supagro, France	Invité, Co-directeur de thèse
Gilles LE MOGUEDEC, Chargé de recherche, Inrae, France	Invité, Encadrant

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

Remerciements

Je souhaiterais adresser un remerciement tout particulier à mes directeurs de thèse. Merci à Bruno Tisseyre pour sa pédagogie et son analyse critique sur mes travaux. Merci à Philippe Vismara pour sa patience, sa disponibilité et pour avoir été à l'écoute. Merci de m'avoir fait confiance. Merci pour votre bienveillance, j'ai beaucoup appris à vos côtés tout au long de ces trois années.

Un grand merci à Gilles Le Moguédec qui a accepté de m'encadrer sur la deuxième partie de la thèse et avec qui j'ai pris plaisir à travailler. Merci pour l'intérêt que tu as porté à mes travaux et merci pour ta rigueur et ton aide sur les aspects statistiques.

Merci également à Sébastien Roux et Olivier Strauss pour leur aide et le regard extérieur qu'ils ont pu apporter à différents moments de la thèse.

Je remercie Christophe Abraham et Bénédicte Fontez de m'avoir fait confiance pour l'encadrement de TD et pour leur aide en statistiques.

Mes remerciements vont également aux autres doctorants et anciens doctorants de l'UMR ITAP : Cécile, Léo, Eva, Anice, Julien et Corentin avec qui j'ai apprécié échanger au cours de ces trois années, nos discussions m'ont toujours beaucoup apporté. Plus généralement, merci à tous les membres de l'équipe ITAP pour leur bonne humeur, et ce, même durant le confinement.

Je remercie tous les membres de l'UMR MISTEA qui m'ont accueilli pendant ces trois ans, en particulier toute l'équipe des non-permanents (et anciens non-permanents) du bâtiment 29. Il me serait bien difficile de nommer tout le monde mais je ne vous oublie pas. Merci à Girault, co-bureau et ami, avec qui j'ai partagé mes joies et galères.

Un grand merci à mes proches et à ma famille sans qui tout ça n'aurait pas été possible. Merci à mes parents. Merci à Johannie pour ses encouragements et son soutien. Enfin, un remerciement tout particulier à mes petits frères, Luc, Clément & Louis, qui ont toujours été là.

Table des matières :

Introduction générale :

L'objectif de cette thèse est de proposer de nouvelles approches pour raisonner les pratiques d'échantillonnage en production végétale. En agriculture, l'échantillonnage est généralement réalisé à l'échelle de la parcelle qui représente l'unité de gestion habituelle des agriculteurs. C'est à cette échelle que sont généralement appliquées les pratiques culturales et à laquelle les interventions (y compris la récolte) sont raisonnées. L'échantillonnage a alors pour but d'estimer différents paramètres qui caractérisent l'ensemble de la parcelle. Les estimations résultant de l'échantillonnage permettent donc de générer des connaissances locales sur le système de production que les agriculteurs utilisent pour ajuster leurs prises de décision.

Les estimations sont cependant toujours associées à une imprécision. Cette imprécision dépend de plusieurs facteurs, les principaux étant : (i) la variabilité de la parcelle considérée, (ii) le choix du nombre de sites de mesure et (iii) de leur position ainsi que (iv) la méthode qui sera utilisée pour inférer l'estimation à partir des observations effectuées au cours de l'échantillonnage. Elle se traduit par une erreur d'estimation, qui correspond à l'écart observé entre la valeur prédite et la valeur réelle. Diverses stratégies d'échantillonnage et d'inférence, souvent issues d'autres domaines scientifiques, proposent des solutions afin de réduire l'imprécision.

Dans le contexte agricole, il convient néanmoins de prendre en compte les contraintes opérationnelles intervenant dans la réalisation de l'échantillonnage. Les parcelles peuvent présenter des caractéristiques qui contraignent souvent fortement les déplacements. La réalisation d'un parcours d'échantillonnage reliant les différents sites de mesure représente donc un coût non-négligeable pour la personne effectuant les observations (agriculteur, conseiller, etc.). Si les aspects relatifs à l'imprécision de l'estimation sont généralement bien documentés dans la littérature scientifique, les contraintes opérationnelles relatives aux déplacements des opérateurs dans les parcelles, pourtant limitantes, n'y sont que très peu considérées.

Cette thèse propose donc de répondre à la question suivante : « Comment améliorer les pratiques d'échantillonnage en production végétale pour mieux répondre aux enjeux du coût et de l'imprécision de l'estimation ? »

Pour résoudre ce problème, les travaux présentés dans cette thèse mobilisent plusieurs approches appartenant à des domaines différents. Des méthodes stochastiques caractérisent les propriétés de l'estimation en prenant en compte le choix des sites de mesure et les méthodes d'inférence associées. Ces méthodes sont associées à des méthodes d'optimisation appartenant à la programmation par contraintes et à la recherche opérationnelle pour déterminer un échantillon avec des propriétés opérationnelles (coût d'échantillonnage) optimales.

Les approches sont testées sur le cas de l'estimation du rendement en viticulture. Réalisée à quelques jours des vendanges, cette application constitue un enjeu important et présente des propriétés complexes susceptibles de répondre à la plupart des problématiques d'échantillonnage en production végétale. Elle est en effet essentielle à la bonne gestion logistique de la récolte et présente des enjeux commerciaux importants pour les exploitations. La structure palissée des vignobles et le peu de temps disponible pour l'estimation mettent en exergue la nécessité de raisonner le coût de l'estimation.

Le premier chapitre de la thèse introduit les concepts et les enjeux relatifs à l'échantillonnage en production végétale. Il présente également le cas d'étude qui sera suivi tout au long du document.

Le deuxième chapitre présente la démarche générale proposée et une première approche de mise en œuvre pour le choix des sites de mesure. Cette première approche est décrite dans un article qui constitue le troisième chapitre de cette thèse. Celle-ci a été utilisée pour identifier les questions soulevées par la mise en œuvre de méthodes stochastiques en association avec des méthodes d'optimisation dans une seule et même approche. Cette association n'est pas triviale et soulève des questions scientifiques spécifiques. Ces questions spécifiques sont abordées dans les chapitres suivants.

Le chapitre quatre s'attache au choix des méthodes d'inférence en quantifiant l'apport d'un estimateur basé sur un modèle linéaire plutôt que sur la moyenne. Ces considérations sont basées sur un formalisme statistique comparant la variance et l'espérance des estimateurs. Le chapitre suivant (chapitre cinq) poursuit cette réflexion dans le cas du modèle linéaire et fait le lien entre la méthode d'inférence et le choix des sites de mesure.

Le chapitre six présente la mise en œuvre des méthodes. Il débute par une présentation rapide de la programmation par contraintes et décrit les mécanismes utilisés pour modéliser le problème. Dans une deuxième partie est présentée une autre modélisation basée sur les outils de la recherche opérationnelle.

Le chapitre sept s'intéresse aux types de parcours optimaux trouvés par les méthodes d'optimisation. Ces parcours sont mis au regard des pratiques actuelles couramment utilisées en viticulture pour donner quelques considérations faisant le lien entre échantillonnage et propriétés des parcelles.

Enfin le chapitre huit présente la mise en place d'une première approche autour de la détection des valeurs aberrantes dans le contexte du cas d'étude. Ce chapitre ouvre sur les perspectives plus générales de ces travaux de thèse.

Chapitre 1 : L'échantillonnage en production végétale

1.1 L'échantillonnage, un outil au service de l'agronomie

1.1.1 Principes généraux de l'échantillonnage

De manière générale, on appelle **échantillonnage** les méthodes de sélection d'un sous-ensemble d'individus statistiques à l'intérieur d'une population. Le résultat de cette procédure est un **échantillon**. A partir de l'échantillon obtenu et de ses propriétés, il est possible de caractériser l'ensemble de la population étudiée par un procédé appelé **inférence** (Thompson, 2012).

Deux grands avantages découlent de l'utilisation de ces méthodes. Etudier un sous-ensemble plutôt que l'intégralité de la population dont il est issu nécessite significativement moins de ressources, qu'il s'agisse de moyens financiers, d'outils, de temps ou de main-d'œuvre. Il n'est parfois tout simplement pas possible d'étudier une population dans son intégralité. A titre d'exemple, il serait impossible de connaitre le nombre exact de feuilles présentes dans une forêt à un instant donné sans une main d'œuvre pléthorique. Mais en comptant le nombre de feuilles sur quelques arbres ou quelques rameaux et le nombre de rameaux par arbre, il devient possible d'estimer le nombre total de feuilles de la forêt. Un autre avantage intervient lorsque la réalisation des mesures implique de détruire ou de modifier les propriétés de l'objet étudié, on parle alors de mesures destructives. C'est le cas par exemple lorsque l'on cherche à connaitre la quantité de sucre dans un fruit ou lorsque l'on fait appel à des procédés de datation au carbone 14. La pratique de l'échantillonnage est alors nécessaire sous peine de dégrader l'intégralité de l'objet d'étude.

Pour ces raisons, les méthodes d'échantillonnage sont mobilisées dans de nombreux domaines tels que la physique des matériaux (caractérisation des propriétés des matériaux Machado et al. 2020), l'étude des signaux (échantillonnage d'un signal variable au cours du temps, Akers et al. 1986), la sociologie (sondages et échantillonnage d'une population humaine, Lameck, 2013), la santé (étude des épidémies, Maganni et al., 2005, échantillonnage de virus, Rahmani et al., 2020, et recherche clinique, Elfil et al., 2017), ou encore l'étude de la biodiversité (recensement d'individus d'une espèce au sein d'un écosystème donné, Mann et al. 2006). Dans chacun de ces domaines, les méthodes d'échantillonnage sont adaptées aux spécificités des problèmes traités mais le principe général reste le même.

1.1.2 Echantillonnage en agronomie

En agronomie, les méthodes d'échantillonnage sont utilisées pour caractériser les propriétés d'espèces, de parcelles ou de territoires. Ces méthodes sont mises en place à des échelles très variables pouvant aller de l'intra plante (échantillonnage de fruits sur un rameau ou de grains à l'intérieur d'un épi) aux territoires (échantillonnage de la production par pays ou région) (Huddleston, 1978 ; FAO, 2010). La parcelle représente l'unité de gestion agricole. C'est à cette échelle que sont raisonnées la gestion des intrants et les pratiques culturales en général (Sheaffer & Moncada,2012). Elle constitue naturellement l'échelle spatiale privilégiée pour l'échantillonnage afin d'apporter l'information nécessaire à la prise de décision en agriculture. S'agissant d'un enjeu important pour améliorer les prises de décision en agriculture, c'est donc à cette échelle qu'est raisonné l'échantillonnage dans les travaux présentés.

L'échantillonnage d'une parcelle peut poursuivre différents objectifs. Il s'agit toujours de caractériser une grandeur d'intérêt mais la manière dont on cherche à la caractériser peut varier. Le plus souvent,

l'échantillonnage cherche simplement à approcher une valeur sur l'ensemble de la parcelle. Il peut s'agir d'une valeur moyenne comme une quantité de sucre moyen par fruit (Kasimatis & Vilas 1985) ou d'une valeur cumulée comme le rendement d'une culture (Sampford, 1962 ; Anderson et al. 2019, Uribeetxebarria et al. 2019). L'échantillonnage peut également avoir pour but la réalisation de cartes de la grandeur étudiée (Fleischer et al., 1999 ; Jordan et al. 2003). Ces cartes doivent permettre de représenter les variations de la grandeur mesurée sur la parcelle, on parle alors de variabilité intra-parcellaire. Dans le même sens, l'échantillonnage peut avoir pour but de subdiviser une parcelle hétérogène en sous-zones homogènes. Les pratiques culturales (implantation et choix dans la conduite de la culture) et intrants utilisés (énergie, engrais, produits phytosanitaires) peuvent ainsi être modulés et adaptés à chaque zone identifiée afin de mieux correspondre aux besoins réels (González-Fernández et al. 2013, Fortes et al. 2015). Les travaux de cette thèse se concentrent principalement autour de la première situation, c'est à dire l'estimation d'une valeur approchant un paramètre de la distribution pour la grandeur étudiée pour la parcelle agricole. Remarquons toutefois, que le terme de « parcelle » sera utilisé pour simplifier le discours mais les concepts et approches présentés s'appliquent à toute unité de gestion agricole, qu'il s'agisse d'une zone intra-parcellaire, de la parcelle dans sa totalité ou d'un ilot de parcelles résultant de l'agrégation de plusieurs unités de gestion agricoles.

1.1.3 Echantillonnage, estimations et erreurs

1.1.3.1 *Définitions*

L'échantillonnage est la première étape pour l'obtention d'une valeur approchée d'un paramètre d'une parcelle. C'est l'étape au cours de laquelle un échantillon de **sites de mesure** est sélectionné parmi l'ensemble des sites disponibles de la parcelle. Des **observations** (ou mesures) de la grandeur d'intérêt sont réalisées sur chacun des sites constituant l'échantillon. La deuxième étape est appelée **inférence** et consiste à déduire des observations la valeur prise par le paramètre d'intérêt pour la parcelle (Figure 1.1). Cette inférence est réalisée en utilisant un **estimateur**, une fonction faisant correspondre à chaque échantillon une valeur que l'on nomme **estimation**. En statistique, l'estimation d'un paramètre quelconque θ est notée $\hat{\theta}$.

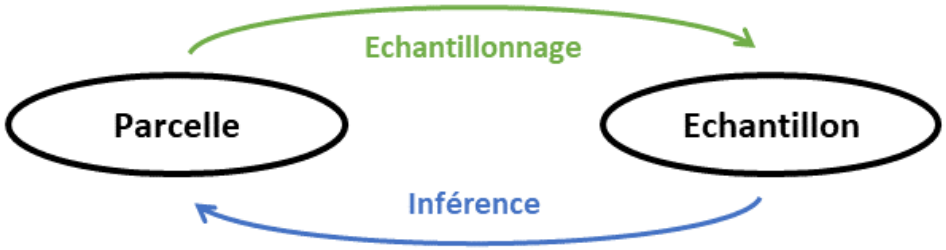

Figure 1.1 : Echantillonnage et inférence

Naturellement, il est souhaité que cette estimation se rapproche autant que possible de la valeur réelle du paramètre estimé. On définit comme **erreur d'estimation** l'écart relatif absolu entre la valeur estimée et la valeur réelle, cette grandeur est généralement exprimée en pourcentage. En reprenant les notations du paragraphe précédent, si la valeur réelle est différente de 0, elle est définie par l'équation :

$$Erreur\ (\%) = \left|\frac{\hat{\theta} - \theta}{\theta}\right| \times 100 \qquad Eq.1.1$$

1.1.3.2 *Erreur d'estimation, biais et variabilité*

Il est possible de décomposer l'écart $\hat{\theta} - \theta$ entre la valeur estimée et la valeur réelle en introduisant l'espérance de l'estimation, notée $E(\hat{\theta})$. Cette espérance représente la valeur moyenne pouvant être attendue pour l'estimateur. L'écart peut alors être écrit :

$$\hat{\theta} - \theta = \left(\hat{\theta} - E(\hat{\theta})\right) + \left(E(\hat{\theta}) - \theta\right) \qquad Eq.1.2$$

Cette décomposition fait apparaitre deux modalités**, le biais** $\left(E(\hat{\theta}) - \theta\right)$ et une mesure de l'écart de l'estimation à son espérance, associé au concept de **variabilité** $\left(\hat{\theta} - E(\hat{\theta})\right)$. Ces deux modalités sont illustrées dans la Figure 1.2.

Le biais contribue à sous-estimer ou surestimer la valeur d'intérêt. En reproduisant une estimation par échantillonnage un grand nombre de fois, le biais correspond à l'écart de la moyenne des estimations à la valeur réelle. Il s'agit d'une erreur « systématique ». Ce biais peut avoir plusieurs origines, par exemple le mauvais étalonnage d'un appareil de mesure, un échantillon délaissant systématiquement certains types d'individus ou un opérateur n'échantillonnant pas de manière objective.

La variabilité joue sur l'erreur d'estimation en réduisant la dispersion des estimations. Plusieurs estimations d'une même grandeur, réalisées selon le même protocole, tendent à donner des résultats différents lorsque la variabilité augmente. En reproduisant l'estimation un grand nombre de fois, la précision correspond à la dispersion des valeurs estimées autour de leur propre moyenne. En statistique, elle est généralement représentée par **la variance** ou **l'écart-type**. La variabilité de l'estimation est généralement dépendante de facteurs tels que le nombre de mesures constituant l'échantillon et la variabilité de la grandeur échantillonnée.

En pratique l'erreur d'estimation observée est le produit du biais et de la variabilité et l'absence de répétition ne permet pas de discriminer les deux types d'erreurs. Savoir quelle part de l'erreur provient du biais et quelle part provient de la variabilité nécessite soit des répétitions d'une même stratégie d'échantillonnage, soit un travail statistique théorique modélisant les différents phénomènes à l'origine de l'erreur d'estimation. Ces concepts apparaissent également dans d'autres domaines sous la dénomination de justesse et fidélité.

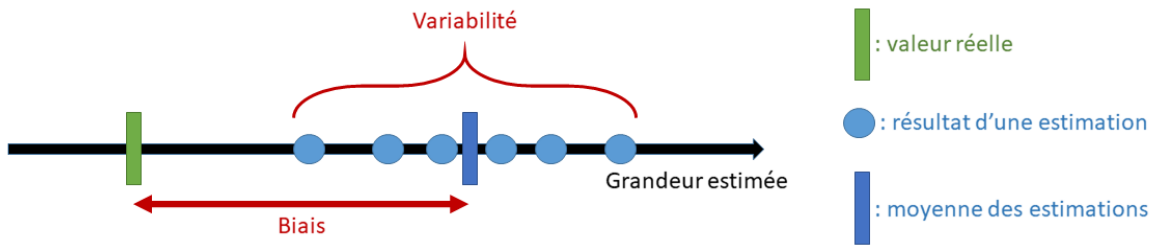

Figure 1.2 : Schéma explicatif du biais et de la variance de l'estimation. Le biais est représenté par l'écart de la moyenne des estimations à la valeur réelle que l'on souhaite atteindre. La variabilité est représentée par la dispersion des estimations.

1.1.3.3 *Inférence et estimateurs*

Dans la majorité des cas, l'échantillonnage est réalisé de manière à caractériser un ou plusieurs paramètres inconnus relatif à la loi de probabilité des observations. Sur une parcelle, il s'agit très généralement de l'espérance, notée μ. Ce paramètre prend en effet tout son sens lorsqu'il s'agit d'estimer un rendement ou de raisonner une quantité d'intrant à l'échelle d'une parcelle. Plus rarement, l'estimation est utilisée pour caractériser une mesure de la variabilité (Yemefak et al., 2005 ; Chang et al., 2003) ou une proportion (Hodgson et al., 2004 ; Halle et al., 2007).

En pratique en agriculture, l'inférence de l'espérance de la loi dont sont issues les observations, est presque systématiquement basée sur une moyenne arithmétique. Cette méthode est choisie pour ses propriétés statistiques en plus de sa simplicité d'utilisation. Le théorème central limite (Laplace, 1809) énonce que la moyenne, notée $\hat{\mu}$, des observations $X_N = (X_1, X_2, \dots, X_n)$ d'un échantillon converge vers une loi normale tant que ces observations sont indépendantes et suivent une même loi de variance finie. Cette loi normale est centrée sur l'espérance des observations et sa variance dépend de la variance de la loi des observations σ^2 et de n, la taille de l'échantillon :

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Eq.1.3$$

Il s'agit donc d'un estimateur non-biaisé de l'espérance puisque $E[\hat{\mu}] = \mu$. Ce théorème est valable pour la grande majorité des lois que peuvent suivre les observations, sous réserve que celles-ci soient indépendantes les unes des autres.

L'inégalité de Cramér-Rao (Kagan 2001) indique que la variance d'un estimateur convergent ne peut être aussi petite que voulue, il existe un minimum qu'il est impossible de dépasser. Pour un estimateur sans biais de μ, cette inégalité s'exprime sous la forme suivante :

$$Var[\hat{\mu}] > \frac{1}{I_{X_N}(\hat{\mu})} \qquad Eq.1.4$$

La variance minimum de l'estimateur $\hat{\theta}$ y apparaît exprimée en fonction de $I_{X_N}(\hat{\mu})$, la quantité d'information de Fisher, une grandeur qui quantifie l'information relative à μ contenue dans l'échantillon d'observation $X_N = (X_1, X_2, \dots, X_n)$. Un estimateur non-biaisé et de variance minimum selon l'inégalité de Cramer-Rao est qualifié d'estimateur efficace. Pour les distributions les plus usuelles (loi normales, lois binomiales), la moyenne est justement un estimateur efficace de l'espérance, il s'agit alors du meilleur estimateur possible pour inférer ce paramètre (Wasserman, 2004).

1.1.3.4 *Erreur d'estimation et imprécision*

Pour un échantillon, l'erreur d'estimation n'est jamais exactement nulle. Cet écart entre estimation et grandeur estimée représente une réalisation de l'**imprécision** associée à l'estimation. L'inégalité de Cramér-Rao (Eq 1.4) montre qu'il existe une imprécision pour l'estimation quelle que soit la taille de l'échantillon et l'estimateur utilisé.

En statistique cette imprécision est généralement représentée par l'intervalle de confiance. Cet intervalle est construit de telle manière qu'il est possible d'affirmer que la valeur réelle se trouve à l'intérieur d'un intervalle défini avec une certaine probabilité. Cette probabilité est exprimée sous la forme $1 - \alpha$, où α représente le risque que la valeur réelle soit située en dehors de l'intervalle. La façon dont est construit l'intervalle de confiance dépend de l'estimateur. En se plaçant dans le cas le plus courant de l'estimation d'une espérance par la moyenne d'un échantillon de taille n, $X_N = (X_1, X_2, \dots, X_n)$ suivant une loi normale, celui-ci est de la forme :

$$\left[\overline{X_N} \pm t_{n-1,\alpha/2} \times \frac{S_{X_N}}{\sqrt{n}}\right] \qquad Eq.1.5$$

Où $\overline{X_N}$ représente la moyenne de l'échantillon, S_{X_N} son écart-type et $t_{n-1,\alpha/2}$ le quantile d'ordre $\alpha/2$ pour une loi de student à $n-1$ degrés de liberté. On a alors :

$$P\left(\mu \in \left[\overline{X_N} - t_{n-1,\alpha/2} \times \frac{S_{X_N}}{\sqrt{n}}, \overline{X_N} + t_{n-1,\alpha/2} \times \frac{S_{X_N}}{\sqrt{n}}\right]\right) = 1 - \alpha \qquad Eq.1.6$$

Comme cela apparait dans la formule de l'intervalle de confiance de la moyenne (Eq. 1.5), l'imprécision d'une estimation dépend principalement du nombre d'observations, représenté par n, et de la variabilité de celles-ci (représenté par S_{X_N}). Le choix du nombre d'observations qui constituent l'échantillon doit donc être réalisé au regard de la variabilité de la grandeur étudiée sur la parcelle et du niveau d'imprécision de l'estimation considéré comme étant acceptable. Pour un même intervalle de confiance, une parcelle présentant une forte variabilité nécessitera ainsi davantage de mesures échantillonnées qu'une parcelle plus homogène.

1.1.3.5 *Choix des sites de mesure et indépendance des observations en agriculture de précision*

Les méthodes d'inférence usuelles se basent sur l'hypothèse que les observations qui constituent l'échantillon sont indépendantes les unes des autres, conditionnellement au fait qu'elles sont issues du même phénomène. En d'autres termes, cela signifie qu'une observation de l'échantillon ne doit en aucun cas être influencée par une autre observation (Wasserman, 2004). Jusqu'à peu, ce concept d'indépendance des observations était couramment admis en agriculture. Toutefois, les développements récents de l'agriculture de précision ont apporté de nouvelles connaissances sur la variabilité intra-parcellaire de certaines variables liées à l'environnement (caractéristiques du sol) mais aussi liées aux cultures (rendement, expression végétative, teneur en chlorophylle, etc.) (Pierse & Nowak, 1999 ; Whelan & Mcbratney, 2000).

L'existence des structures spatiales intra-parcellaires et d'auto-corrélations spatiales mises en évidence par l'agriculture de précision révèle de nouveaux enjeux pour l'échantillonnage (Kerry et al., 2010 ; Brus & Gruijter, 1997). La variabilité entre deux observations devient fonction de leurs positions respectives : deux observations proches auront tendance à présenter davantage de similarité que deux observations situées aux extrémités d'une même parcelle. Ce principe remet en cause l'hypothèse d'indépendance des observations. Dans ce contexte, il apparait nécessaire de pouvoir raisonner le choix des sites de mesure et la distance qui les sépare afin de s'assurer de la qualité de l'estimation finale. Toutefois, la connaissance, avant échantillonnage de la structure spatiale de la variable d'intérêt n'est pas triviale. En effet, elle supposerait que la variabilité spatiale de la variable d'intérêt soit connue à l'avance. Pratiquement, il en résulte que le définition minimale d'une distance d'échantillonnage garantissant l'indépendance des observations pendant l'échantillonnage n'est pas possible en considérant la variable d'intérêt dont les caractéristiques sont inconnues *a priori*. Une approche proposée dans la littérature (Adamchuk et al., 2011, Kerry et al., 2010) consiste à raisonner sur des variables dites auxiliaires faciles à mesurer ou disponibles avec une haute résolution spatiale. La variable auxiliaire est choisie pour sa pertinence à expliquer la variabilité spatiale de la variable d'intérêt : par exemple le rendement avec un indice de végétation obtenu par télédétection (Carillo, 2016). En faisant l'hypothèse que la variable d'intérêt est plus ou moins liée à la variable auxiliaire (connue), il devient donc possible de mieux raisonner l'indépendance spatiale des sites d'observation. En agriculture de précision, l'étude de la variabilité spatiale est presque systématiquement associée au semi-variogramme ou variogramme (Bachmaier et Backes, 2008 ; Kerry & Oliver, 2003), un modèle décrivant la variance entre deux observations en fonction de la distance qui les sépare. Une brève présentation de cet outil est disponible en annexe (Annexe : A propos du semi-variogramme).

1.1.4 Effort de l'estimation

Le recours à l'échantillonnage permet d'estimer une variable avec un effort (coût) bien moindre que des mesures exhaustives sur l'ensemble des sites de la parcelle. Il existe néanmoins un effort d'échantillonnage associé aux temps de main d'œuvre nécessaire. Ce temps de main d'œuvre est défini par l'observation à effectuer sur chaque site de mesure (comptage, prise d'échantillon, etc.) et aux déplacements de l'opérateur sur la parcelle entre les différents sites d'observation et entre les

parcelles. Pour une variable d'intérêt donnée, le temps nécessaire à l'observation est indépendant de la stratégie d'échantillonnage car il dépend d'un protocole et/ou de choix de méthodes propres aux acteurs et aux organisations impliqués (agriculteurs, techniciens, coopératives, etc.). De la même manière, l'effort lié aux déplacements pour aller d'une parcelle à une autre est jugé indépendant de la stratégie d'échantillonnage puisque dépendant de l'organisation du parcellaire. Il en résulte que pour une stratégie d'échantillonnage donnée, l'effort est associé au temps nécessaire au déplacement entre les sites à échantillonner. Cet effort est donc représenté par la durée de l'échantillonnage. Il peut être décomposé en deux composantes, la durée nécessaire pour réaliser les mesures et la durée des déplacements intra parcellaires :

$$Durée_{échantillonnage} = Durée_{mesure} + Durée_{déplacement} \qquad Eq.1.7$$

En décomposant la durée des déplacements intra parcellaires on obtient :

$$Durée_{échantillonnage} = Durée_{mesure} + Distance_{déplacement} \times Vitesse_{déplacement} \qquad Eq.1.8$$

En supposant la durée nécessaire à la réalisation des mesures spécifiques à la grandeur mesurée et la vitesse de déplacement de l'observateur constantes, réduire l'effort d'échantillonnage revient à minimiser la distance qu'il est nécessaire de parcourir dans la parcelle. Pour cela, les sites de mesure doivent être choisis au regard de leurs positions respectives dans la parcelle et de celle du point de départ de l'échantillonneur. L'ordre dans lequel les sites de mesures sont parcourus influe également sur cette distance. En pratique, le coût de l'estimation et la distance qu'il est nécessaire de parcourir sont généralement pris en compte de manière empirique mais représente des contraintes souvent négligées pour la mise en place de nouvelles méthodes d'échantillonnage.

1.1.5 Stratégies d'échantillonnage en agronomie

1.1.5.1 *Echantillonnage, représentativité et données auxiliaires*

Au-delà du problème d'indépendance évoqué dans la sous-partie 1.1.3.5, le choix des sites de mesure est soumis à un second enjeu lié à la qualité de l'estimation, celui de la représentativité. Idéalement, un site de mesure doit représenter une partie de la population. C'est à dire présenter des caractéristiques similaires à un ensemble de sites de mesure qui n'ont pas été sélectionnés dans l'échantillon final, cet ensemble de sites étant généralement situé dans une même zone géographique de la parcelle (Binns et al., 2000). Cette particularité définit la **représentativité** d'un échantillon (Kruskal & Mosteller, 1979 ; Ramsey, 2005). Il est généralement préférable d'éviter les sites de mesure associés à des évènements rares tels que ceux associés à des zones très localisées de mortalité ou affectées par des événements extraordinaires (dégâts de gibier). Dans l'ensemble, l'échantillon doit pouvoir permettre de dresser un portrait général de la parcelle en question. A titre d'exemple, la Figure 1.3 présente une parcelle constituée de deux zones, une zone A plus productive pour un tiers de sa surface et une zone B moins productive pour les deux tiers restant. Un échantillonnage représentatif devrait en théorie présenter deux fois plus de sites de mesures dans la zone B que dans la zone A.

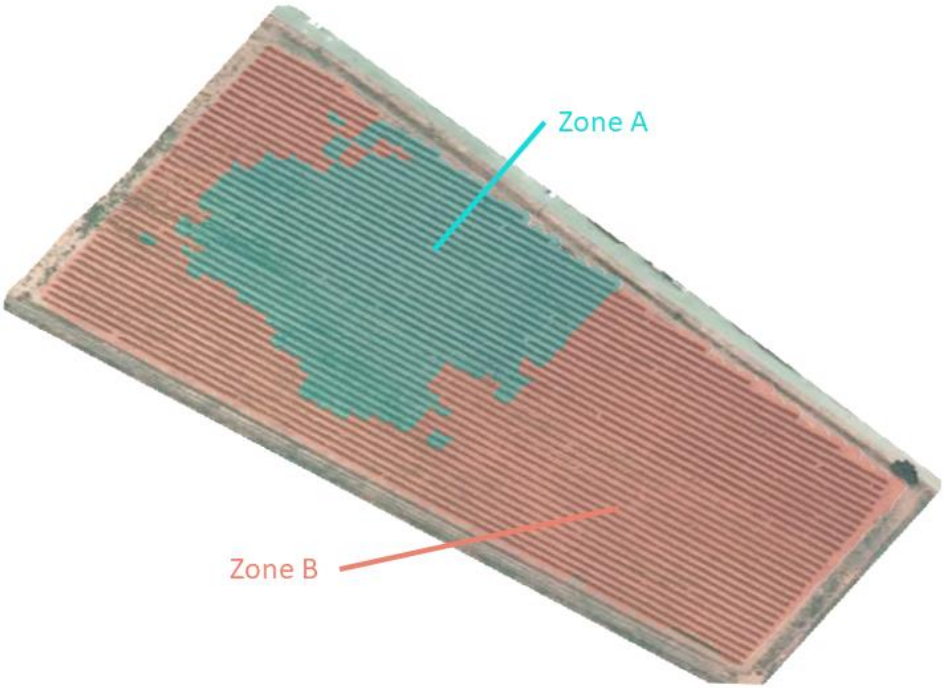

Figure 1.3 : Illustration d'une parcelle présentant deux zones distinctes (A & B). Un des enjeux de la représentativité de l'échantillonnage est de répartir correctement les sites de mesure entre les deux zones.

En pratique il est souvent difficile de s'assurer de la représentativité d'un plan d'échantillonnage. La variable d'intérêt n'étant pas connue *a priori*, comme évoqué dans la section 1.1.3.5. Pour contourner ce problème certaines approches se basent sur l'utilisation de données dites auxiliaires. Les **données auxiliaires** correspondent à des variables déjà connues ou accessibles à moindres coûts avec une haute résolution spatiale et qui ont pour propriété d'être corrélées à la variable d'intérêt. Ces variables doivent être disponibles avec une résolution suffisante afin de pouvoir discriminer les différents sites de la parcelle. Il peut par exemple s'agir de données historiques provenant des années précédentes (Araya-alman et al., 2019), de données acquises par imagerie aérienne ou satellitaire tel que des indices de biomasse ou d'information sur les propriétés pédologiques de la parcelle (Carrillo et al., 2016 ; Meyers et al., 2020). L'utilisation de telles données doit être conduite au regard de leur résolution, corrélation à la variable d'intérêt, origine et transformations éventuelles (interpolation, lissage, fusion …).

Le choix des sites de mesures qui constitueront l'échantillon ainsi que leur nombre apparaît donc comme une des questions centrales pour l'échantillonnage des parcelles en agriculture et ce, quelle que soit la manière dont sont réalisées les mesures. Les sous-parties suivantes proposent une liste non exhaustive des approches d'échantillonnage.

1.1.5.2 *Echantillonner sans information a priori*

Lorsqu'aucune donnée auxiliaire ou information n'est disponible sur la variable d'intérêt, un échantillonnage aléatoire est généralement le plus indiqué. Cette stratégie donne à chaque site une même probabilité d'être sélectionné dans l'échantillon final. L'absence d'information ne permettant pas de sélectionner préférentiellement un individu par rapport à un autre, le choix des sites d'échantillonnage se fait donc avec équiprobabilité. Cette approche peut sembler difficile à mettre en œuvre sur le terrain, car le caractère aléatoire peut souvent être biaisé par des contraintes pratiques, telles que les distances à parcourir ou le point d'entrée dans la parcelle.

Les alternatives à l'échantillonnage aléatoire reposent sur la réalisation de mesures sur une maille régulière (grid sampling). Cela peut se faire en localisant des sites de mesure sur les nœuds d'une grille régulière sur la parcelle (grid sampling) ou en parcourant l'ensemble de la parcelle et en effectuant une mesure sur des bases régulières (systematic sampling et systematic uniformly random sampling, Wulfsohn et al., 2010). Le Grid Sampling requiert cependant un nombre suffisant de sites de mesure pour être pertinent et reste principalement utilisé pour la construction de carte (Spezia et al., 2012) plutôt que pour l'estimation d'un paramètre moyen.

1.1.5.3 *Echantillonnage avec données auxiliaires*

Lorsqu'elles sont disponibles, l'intégration de données auxiliaires dans le plan d'échantillonnage peut améliorer considérablement la qualité de l'estimation. Des approches telles que le targeted sampling (ou échantillonnage orienté) exploitent le lien entre les variables pour améliorer la représentativité. Le targeted sampling propose simplement de sélectionner les sites à mesurer en fonction de la valeur de leurs données auxiliaires. Le processus statistique de sélection des sites cibles peut varier (quantile, k-means, rank set sampling, etc.) mais l'objectif est toujours de définir un ensemble de sites d'échantillonnage qui assure une certaine représentativité au regard de la distribution des données auxiliaires. Ces approches sont largement utilisées, en particulier dans les études de sol (Adamchuck et al., 2011) et ont été proposées pour le rendement en viticulture (Bramley, 2001 ; Carrillo et al., 2016). Certaines variantes, telles que le Ranked Set sampling (Dell & Cluter, 1972 ; Chen et al., 2004), ont également été spécifiquement déployées dans les cultures fruitières horticoles pérennes (Uribeetxebarria et al., 2019).

Le model sampling (ou échantillonnage par modèle) suit les principes du targeted sampling en allant plus loin dans l'exploitation des données auxiliaires disponibles (Basso et al., 2001 ; Särndal et al., 1992). Cette stratégie d'échantillonnage utilise les observations faites sur les sites de mesure de l'échantillon pour étalonner les paramètres d'un modèle reliant la variable de rendement aux données auxiliaires. Dans un deuxième temps, le modèle nouvellement étalonné est utilisé pour prédire les valeurs de la variable de rendement à partir de l'ensemble des données auxiliaires disponibles. L'estimation finale est alors effectuée en utilisant la moyenne de toutes les valeurs prédites. Cette approche a montré de très bons résultats notamment pour l'estimation du rendement en viticulture (Carrillo et al., 2016).

D'autres méthodes sont utilisées pour échantillonner des populations complexes qui peuvent être subdivisées en sous-populations. Les critères utilisés pour former ces sous-populations doivent être adaptés aux objectifs de l'échantillonnage. Ce type de méthode peut par exemple être utilisé pour échantillonner une parcelle subdivisée en plusieurs îlots ou sur laquelle des éléments connus et parfaitement délimités affectent la distribution de la variable d'intérêt, comme les unités pédologiques différentes par exemple. Il existe différentes façons d'échantillonner ce type de population, comme le cluster sampling. Cette stratégie propose de choisir des sites de mesure en tirant au hasard une sous-population puis un individu de la sous-population, en laissant la liberté d'attribuer différentes probabilités à la sous-population et aux individus. Le poids attribué à chaque observation dans la moyenne finale peut également varier en fonction de la sous-population d'origine. Certaines variantes de ces approches d'échantillonnage stratifié ont déjà été appliquées en agronomie (Wulfsohn et al., 2010).

1.2 Cas d'étude : L'échantillonnage au service de l'estimation du rendement en viticulture

Les travaux de thèse présentés dans ce document se concentrent autour du cas d'étude de l'estimation du rendement en viticulture. Cette section présente les spécificités associées à cette estimation. Une grande partie des informations de cette section proviennent d'un état de l'art récent (Laurent et al. 2020) en cours de soumission sur l'estimation du rendement auquel j'ai participé.

1.2.1 Le rendement viticole et son estimation

1.2.1.1 *Qu'est-ce que le rendement viticole ?*

Le rendement viticole correspond à un poids ou un volume de raisin récolté à la vendange par unité de surface ou par parcelle. L'enjeu de l'estimation est d'obtenir une valeur approchée de ce rendement avant même la récolte. Pour faire cette estimation, l'approche la plus commune consiste à décomposer le rendement en composante (Dry, 2000 ; Clingeleffer et al., 2001). Ces approches partent généralement du nombre de ceps par parcelle qui, multiplié au nombre de grappes par cep, au nombre de baies par grappe et au poids moyen par grappe, permet d'obtenir une masse pour la parcelle (Eq. 2).

$$Rendement = Nbr_{ceps} \times \frac{Nbr_{grappes}}{cep} \times \frac{Nbr_{baies}}{grappe} \times Poids_{baie} \qquad Eq\ 1.9$$

En raisonnant sur des unités :

$$\left(\frac{kg}{parcelle}\right) = \left(\frac{cep}{parcelle}\right) \times \left(\frac{grappe}{cep}\right) \times \left(\frac{baie}{grappe}\right) \times \left(\frac{kg}{baie}\right) \qquad Eq\ 1.10$$

Toutes ces composantes n'interviennent pas à égalité dans l'élaboration du rendement. Une étude menée sur trois années par Guilpart et al. (2014) montre que le nombre de grappes par cep, le nombre de baies par grappe et le poids moyen d'une baie expliquent respectivement 55%, 14% et 26% de la variabilité temporelle du rendement. Le choix des composantes mesurées est donc adapté aux objectifs de prédiction mais aussi à l'effort d'échantillonnage que requiert chacune d'elle et qui vont définir la durée d'une prise de mesure sur un site d''échantillonnage. Il en résulte que toutes les composantes ne sont pas systématiquement mesurées.

1.2.1.2 *Les enjeux de l'estimation du rendement en viticulture*

L'estimation du rendement est un enjeu à plusieurs niveaux pour les viticulteurs et les coopératives. Celle-ci est nécessaire pour soutenir les prises de décision allant des pratiques culturales (interventions sur la parcelle avant la récolte) à la commercialisation et la comptabilité, en passant par la logistique de la récolte (besoin en main d'œuvre et en équipement) et de la vinification (assignement des parcelles à différentes cuves).

Table 1.1 : Résumé des principaux enjeux de l'estimation du rendement par Laurent et al. 2020.

Expected date	Expected spatial scale	Expected unit	Associated operational decisions	Expected benefits
before budbreak of season n	*field or within-field zones*	*mass per unit area*	*vineyard operations: reasoning pruning intensity and soil fertilization*	*optimized management of marketable yield*
	wine blends	*volume of wine blends after press*	*winery logistics: purchase of the barrels*	*costs saving, possibility to order any required material*
during season n	*block or within blocks zones*	*mass per unit area*	*vineyard operations: reasoning bunch thinning intensity, eventual fertilisation and irrigation level*	*optimized management of marketable yield*
	one or several blocks	*final volume of wine blends*	*accounting: managing stocks, planning revenue*	*good accounting, investment reasoning*
	production area	*final volume of wine blends*	*territorial agency: planning marketing & commercialisation* *wine traders: purchase contracting*	*profitable sales and purchases*
just before harvest n	*field or within-field zones in regards to all the fields to be harvested*	*mass per field*	*vineyards operations : organizing harvest, planning work force and allocating transport equipment*	*optimized harvest decision*
	one or several blocks	*mass per wine blend*	*winery logistics : making and allocating space in tanks, purchasing wine-making consumables, planning tasks and work force, scheduling harvest intakes and treatment*	*optimal harvest blending and gain in wine quality*
	production area	*wine final volume*	*territorial agency : announcement of an eventual regulation for harvest volumes*	*optimized commercialisation*
longer term	*production area*	*wine final volume*	*whole industry : anticipation of the effects of future contexts on wine production, market price etc. for research orientation and strategic development*	*business sustainability*

Les conséquences d'une mauvaise estimation impactent significativement la production (traitement inadapté, mauvaise gestion des volumes, etc.). À l'échelle d'une zone d'approvisionnement ou d'un territoire, l'estimation du rendement est également une aide importante à la prise de décision à des fins commerciales comme dans le cas d'appellations contrôlées par exemple. La Table 1.1 reprend les différents enjeux liés à l'estimation du rendement (Laurent et al. 2020).

1.2.1.3 *Variabilité du rendement en viticulture*

Plusieurs travaux scientifiques ont mis en évidence une variabilité spatiale intra-parcellaire importante du rendement en viticulture (Taylor et al., 2005). Ces travaux ont proposé de décomposer cette variabilité intra-parcellaire en fonction de deux composantes. La première composante correspond à la variabilité structurée spatialement, responsable du fait que deux sites éloignés auront tendance à présenter une plus grande différence que deux sites proches. Cette première variabilité résulte de phénomènes variant à l'échelle de la parcelle comme l'accès aux ressources par les plantes (lumière, eau, minéraux), la profondeur du sol, l'exposition aux maladies etc. La deuxième composante est non structurée spatialement et est qualifiée d'erratique (Figure 1.4). Celle-ci est indépendante de la position du site de mesure et résulte de facteurs très localisés comme la compétition entre les ceps par exemple ou de tout autre facteur incontrôlé comme la variance associée à l'observation elle-même (capteur, observateur, etc.). Pour le rendement viticole, la part de variabilité erratique représente entre 20 et 50% de la variabilité totale (Taylor et al., 2005). Cette forte variabilité spatiale justifie le

recours aux outils de l'agriculture de précision et s'étend plus généralement à d'autres variables utilisées en viticulture (Taylor et al., 2013).

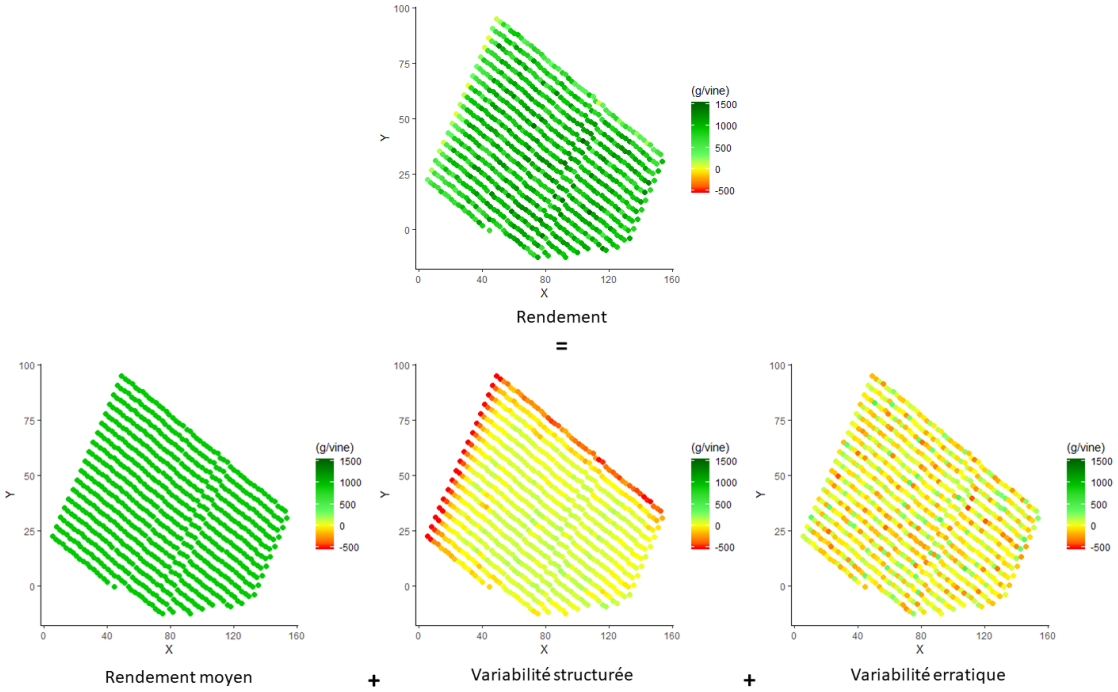

Figure 1.4 : Décomposition du rendement et de sa variabilité. En chaque site de la parcelle, le rendement est exprimé comme la somme du rendement moyen, de sa variabilité spatialement structurée et de sa variabilité erratique.

Il existe également une variabilité temporelle du rendement. Elle correspond aux différences observables d'une saison à l'autre. (Chloupek et al., 2004 ; Sommer et al. 2008). Cette variabilité temporelle est le résultat de l'exposition à des facteurs externes (climat, ravageurs) variant au cours du temps. Le cycle de reproduction de la vigne se déroule sur deux saisons (Howell, 2001 ; Clingeleffer et al., 2001 ; Carmona et al., 2008 ; Vasconcelos et al., 2009 ; Guilpart et al., 2014). Le rendement de l'année n est donc le résultat des évènements ayant eu lieu au cours des deux saisons précédentes (année n et n-1). En dehors du processus d'élaboration du rendement, celui-ci peut varier de manière importante dans les derniers jours avant la vendange lorsque les baies sont exposées à une déshydratation pouvant affecter significativement leur poids.

1.2.2 L'échantillonnage du rendement en viticulture

1.2.2.1 *La mesure de l'estimation du rendement en viticulture*

L'estimation du rendement en viticulture peut prendre plusieurs formes selon les composantes du rendement considérées. Cette estimation s'appuie presque toujours sur un échantillonnage puisque les composantes du rendement, à l'exception du nombre de ceps, requièrent des mesures directes sur la parcelle. Ces mesures correspondent à des comptages (nombre de grappes ou nombre de baies) et à des mesures de poids ou de volume (poids / volume des baies ou des grappes), ces dernières mesures étant généralement destructives (Clingeleffer et al., 2001). Ces comptages sont effectués sur un site de mesure. Un échantillon est donc caractérisé pour une valeur issue des résultats de comptage et une localisation correspondant au site sur lequel les observations ont été effectuées. Pour composer avec la forte variabilité erratique du rendement en viticulture, la mesure retenue pour chaque site de mesure correspond généralement à la valeur moyenne mesurée sur un ensemble de 4 à 5 ceps consécutifs *de part et d'autre/centrés autour* du site de mesure. Le nombre de baies et de grappes

étant importants sur chaque site de mesure, leur comptage systématique correspond à une tâche fastidieuse et particulièrement coûteuse en temps. De nombreux travaux récents ont proposé de nouvelles méthodes afin d'améliorer et de faciliter la réalisation de ces mesures (Aquino et al., 2018 ; Liu et al., 2018 ; Pothen and Nuske, 2016 ; Abdelghafour et al., 2017). Ces méthodes se basent pour la plupart sur des outils d'analyse d'image destinés à automatiser la quantification des composantes du rendement en proposant un comptage automatique du nombre de baies et du nombre de grappes. Cependant l'utilisation généralisée de ces approches sur le terrain reste difficile de par leur sensibilité à la variabilité des conditions de mesure (Nuske et al., 2014 ; Grimm et al. 2019). Il existe donc plusieurs manières de réaliser ces mesures. Le point commun à toutes ces approches réside dans la nécessité de raisonner la position de ces observations sur la parcelle.

1.2.2.2 *Contraintes de l'échantillonnages du rendement en viticulture*

En viticulture, l''échantillonnage visant à estimer le rendement d'une parcelle est souvent systématiquement réalisé en fin de saison pour répondre à des problématiques organisationnelles et logistiques de vendange et de réception à la cave. L'estimation du rendement avant vendange constitue donc une problématique importante pour la profession. Il est généralement réalisé quelques jours avant la récolte afin de ne pas être affecté par la variation du volume des baies et de correspondre au mieux au rendement qui sera récolté. Il intervient donc au cours d'une période déjà chargée pour la profession et doit généralement être réalisé sur un grand nombre de parcelles dans un laps de temps très limité. Pour l'échantillonneur, cela se traduit par un temps disponible par parcelle de l'ordre de quelques dizaines de minutes. Le nombre de mesures réalisables dans ce temps imparti est généralement faible (inférieur à 10), ce qui tend à augmenter les erreurs d'estimation. Plusieurs auteurs se sont intéressés à ce problème de la temporalité de l'estimation du rendement en proposant des approches visant à produire des estimations plus précoces de certaines composantes du rendement (Serrano et al., 2005 ; Roscher et al. 2014 ; Nuske et al., 2014). Le recours à ces méthodes est cependant associé à une imprécision supplémentaire sur la valeur de la composante au moment de la récolte variant entre 10% et 25%.

Une autre contrainte de l'échantillonnage en viticulture est associée à la structure palissée de la vigne. Notons que d'autres modes de conduite non palissés existent en viticulture, c'est le cas du mode de conduite en gobelet qui était traditionnel sur le vignoble méditerranéen. Toutefois, avec la mécanisation systématique des opérations, ces modes de conduits sont aujourd'hui minoritaires. Il en résulte que les vignes sont aujourd'hui majoritairement palissées avec des rangs de plantes soutenues verticalement par des fils de palissages (à un, deux ou trois niveaux de hauteur) soutenus par des piquets. Cette organisation rend la traversée d'un rang impossible. Cette organisation permet d'optimiser l'action des machines et la mécanisation le long des rangs mais elle constitue une contrainte pour les déplacements du personnel sur la parcelle. L'échantillonneur est donc forcé à se déplacer uniquement dans le long d'inter-rangs, et impose de se rendre à une extrémité de la parcelle pour changer d'inter-rang. A noter que des passages peuvent être aménagés à l'intérieur des rangs sur des parcelles de grandes dimensions. Le changement d'inter-rang nécessite alors tout de même d'atteindre le passage le plus proche. En se déplaçant dans un inter-rang donné, l'échantillonneur a accès à deux rangs distincts (un à droite et un à gauche), ce qui signifie que chaque cep est accessible depuis deux inter-rangs. Ces propriétés spécifiques complexifient la gestion des déplacement en multipliant les chemins permettant de relier sites de mesure donnés comme le montre la Figure 1.5.

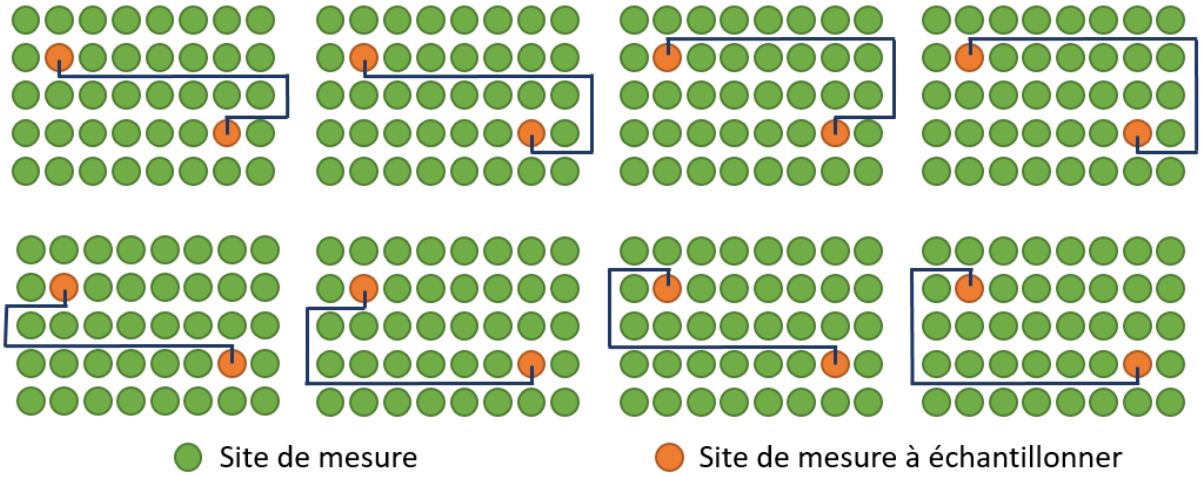

Figure 1.5 : Illustration de la diversité des chemins existants pour relier deux pieds de vignes situés sur des rangs différents lorsque la vigne est palissée. Il existe jusqu'à 8 manière de relier sites de mesures suivant les inter-rangs considérés et les extrémités de parcelles choisis pour changer d'inter-rang.

Un site de mesure est par la suite défini par le support spatial d'échantillonnage pratiqué par les opérateurs. Un site de mesure sera donc défini par les habitudes ou le protocole d'échantillonnage utilisé localement par les professionnels. Un site de mesure pourra donc correspondre à un cep de vigne (dans le cas où le protocole utilisé se base sur l'observation des composantes sur un pied isolé) comme à un ensemble de 4 ou 5 pieds de vigne contigües dans le cas où le protocole défini par les professionnels est basé sur ce support spatial (Figure 1.6).

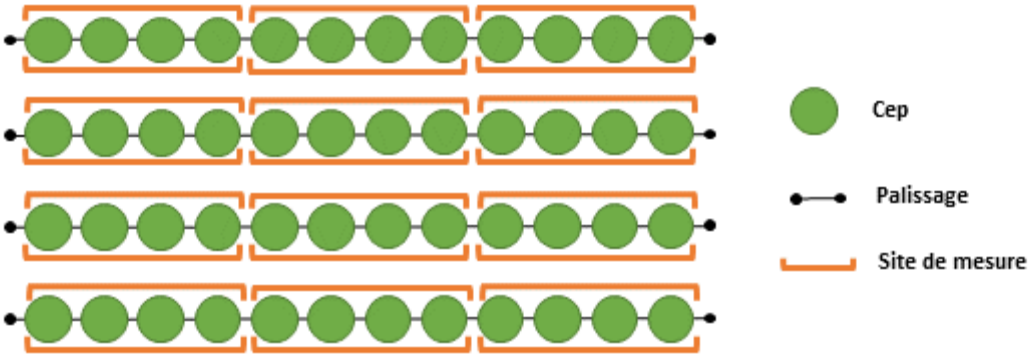

Figure 1.6 : Un site de mesure est un ensemble de plusieurs ceps.

De par la structure des parcelles, il existe une question importante relative aux sites de mesures selon s'ils sont associés ou non à un inter-rang. Pour un site de mesure qui n'est pas associé à un inter-rang, l'observation pourra être réalisée d'un côté ou de l'autre du rang selon l'inter-rang d'accès (Figure 1.5). Dans ce cas, il est possible de définir trois sites A, B et C tels que la « distance » séparant le site A du site B est inférieur à la somme des « distances » entre les sites A et C et entre les sites C et B (Figure 1.7). Il ne s'agit donc pas ici d'une distance à proprement parler puisque celle-ci ne respecte pas l'inégalité triangulaire.

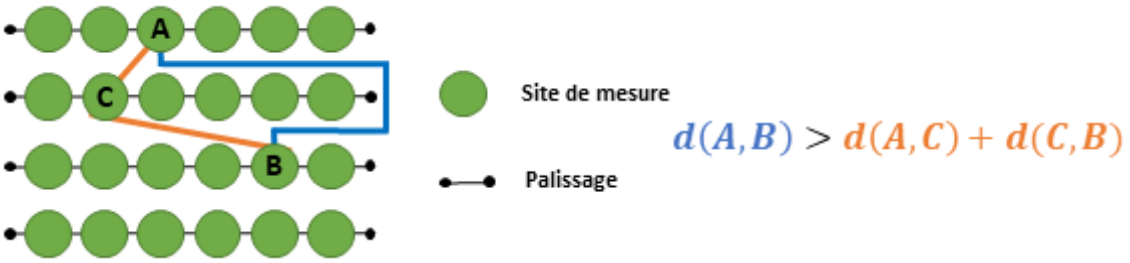

Figure 1.7 : La longueur des chemins séparant deux sites de mesure ne respecte pas l'inégalité triangulaire. Sans assigner un site de mesure à un inter-rang d'accès, la définition de distance ne s'applique pas.

Pour raisonner sur des distances au sens mathématique du terme, il est nécessaire que les sites de mesure soient définis par un inter-rang d'accès en plus du support spatial de la mesure. Chaque cep ou ensemble de ceps correspond alors à deux sites de mesure possibles. Dans ce cas, on définit la distance entre deux sites :

- Comme une distance euclidienne s'ils partagent le même inter-rang ;
- Comme la somme de trois distances euclidiennes lorsqu'ils ne sont pas dans le même inter-rang. On somme ainsi les distances entre les sites de mesure et l'extrémité de leur rang respectif et la distance qui sépare les deux rangs. A noter que chaque rang présente deux extrémités, on choisit alors l'extrémité permettant d'aboutir à la distance la plus courte.

Par ailleurs, l'échantillonnage devra nécessairement être adapté à chaque parcelle et à ses spécificités. En effet, chaque parcelle est unique en terme de de taille (surface), de forme et d'orientation des rangs qu'il est nécessaire de prendre en compte pour le positionnement des sites de mesures. Enfin, d'un point de vue opérationnel, l'échantillonnage efficace et pertinent doit également être conditionné par le point d'entrée dans la parcelle puisque celui-ci influe sur les distances qu'il sera nécessaire de parcourir pour accéder aux premiers sites de mesure.

1.2.2.3 *Echantillonnage opérationnel en viticulture*

Il n'existe pas de convention ou de protocole largement répandu définissant un ensemble de règles pour l'échantillonnage en viticulture. Les pratiques varient donc en fonction des pays, des régions, des domaines viticoles ou des coopératives et parfois en fonction des personnes en charge de l'échantillonnage. La pratique la plus commune consiste à réaliser un ou plusieurs allers-retours dans deux inter-rangs distincts pour optimiser le temps de parcours sur la parcelle. Le choix des inter-rangs et des sites de mesure à l'intérieur des inter-rangs se base alors sur la perception de la parcelle par l'échantillonneur et sur ses observations faites en temps réel sur la parcelle. En fonction des situations, il existe toutefois des règles empiriques visant à prendre en compte des éléments visibles et remarquables susceptibles d'expliquer la variabilité du rendement de la parcelle. C'est par exemple le cas de la topométrie susceptible d'expliquer des conditions pédologiques différentes. L'échantillonneur est alors invité à choisir les inter-rangs à suivre et/ou à répartir les sites d'observation du rendement le long de la plus grande pente. Pour toutes ces raisons, l'échantillonnage du rendement viticole est souvent confronté à des problèmes de reproductibilité et l'information apportée par la position des sites de mesure dans la parcelle n'est pas mobilisée ou perdue. Ces facteurs, associés à un nombre limité de mesures (variant communément entre 5 et 10 par parcelles) et à la forte variabilité des rendements entraînent souvent des erreurs d'estimation importantes de l'ordre de 30%. Cette incapacité à évaluer le rendement avec suffisamment de précision est aujourd'hui un verrou important pour la profession viticole en particulier pour optimiser la gestion logistique des vendanges (Laurent et al. 2020).

1.2.2.4 *Etat de l'art de l'échantillonnage du rendement en viticulture*

Peu de travaux traitent spécifiquement de la stratégie d'échantillonnage pour l'estimation du rendement en viticulture. Les premiers font état du nombre de sites de mesure nécessaires pour obtenir une précision d'estimation suffisante. Wolpert & Vilas (1992) expriment le nombre de sites de mesure nécessaires en fonction de la variance stochastique, de la taille de la zone échantillonnée et de la précision souhaitée. En moyenne, le nombre de sites de mesure nécessaires pour atteindre une précision de l'ordre de 5% est compris entre 20 et 30 (Clingeleffer et al., 2001). Cela suppose également que toutes les composantes sont échantillonnées. Ces valeurs sont à mettre en perspective avec le nombre effectif de mesures réalisées en pratique (inférieur à 10). L'écart entre ces valeurs justifie les importantes erreurs d'estimation existantes aujourd'hui qui se situent souvent aux alentours de 30% et le besoin d'optimiser le temps associé aux parcours d'estimation.

Des travaux plus récents tentent de résoudre le problème du choix des sites de mesure par l'utilisation de données auxiliaires. Cette question a déjà été partiellement abordée au au paragraphe 1.1.5. Carillo et al. (2016) se base sur le NDVI (indice de végétation par différence normalisée), un indice de biomasse caractérisant la proportion de végétation photo-synthétiquement active qui peut être acquis par imagerie aérienne, satellitaire ou par drone (l'Annexe : A propos du NDVI propose un encart sur le NDVI et les indices de végétation). Ce travail met en lumière la corrélation qui existe entre cet indice de végétation et les différentes composantes du rendement. Sur une base de donnée constituée de 9 parcelles situées dans le sud de la France, ce travail montre que la corrélation avec le rendement total varie entre 0.25 et 0.8 avec de fortes variabilités d'une parcelle à l'autre. Dans un second temps, l'article montre le gain d'un échantillonnage orienté (*target sampling*) et d'un échantillonnage basé sur un modèle (*model sampling*). Sur ces exemples, l'erreur diminue de 5 à 7 pour cent lorsque ces méthodes d'échantillonnage basée sur le NDVI sont comparées à une méthode d'échantillonnage aléatoire. Plus récemment, Araya-Alman et al. (2017 & 2019) ont appliqué une méthodologie similaire en se basant sur une autre donnée auxiliaire : les données historiques de rendement. Les gains apportés par cette approche se révèlent être du même ordre de grandeur qu'avec des données de biomasse NDVI.

Ces travaux montrent qu'il est possible d'améliorer significativement les erreurs d'estimation issues d'un processus d'échantillonnage grâce aux données auxiliaires. Cependant, la qualité d'un plan d'échantillonnage ne se définit pas uniquement par l'erreur associée. Le coût de l'échantillonnage, résumé par la durée de l'échantillonnage doit être pris en compte. Cet aspect est particulièrement important dans le cas d'une culture palissée. En effet, échantillonner des sites de mesure dispersés sur une parcelle entraîne nécessairement une augmentation drastique des distances du fait de la structure palissée de la vigne. Ces contraintes opérationnelles tendent à rendre inapplicables ces nouvelles approches dans un contexte opérationnel. Ce double enjeu de l'échantillonnage autour de la qualité et de l'applicabilité se retrouve dans des publications récentes. Meyers et al. (2020) notamment tentent de contourner ce problème en ne proposant de sélectionner que des sites de mesure adjacent pour prendre en compte le caractère opérationnel de l'échantillonnage proposé sur la base de données auxiliaires. Il s'agit néanmoins d'une approche simplifiée et limitée, autant pour la prise en compte de la distance que pour la représentativité des sites choisis.

1.3 Concevoir un approche d'échantillonnage opérationnelle

Dans ce premier chapitre, nous avons présenté le contexte et les enjeux d'une estimation par échantillonnage en production végétale. Y sont présentées les questions de la qualité de l'estimation à travers l'influence de la représentativité et de l'indépendance des sites de mesure ainsi que les

réponses apportées par les principales stratégies d'échantillonnage existantes. Un second enjeu autour du coût de l'estimation et des contraintes opérationnelles s'appliquant au choix des sites de mesure est également présenté.

L'objectif principal des chapitres suivants est de montrer comment construire une nouvelle approche pour le choix des sites d'échantillonnage en production végétale. Cette approche devant résoudre le double problème de la qualité et du coût de l'estimation. Pour répondre à ces enjeux, plusieurs contraintes et critères statistiques et agronomiques sont identifiés. Les parcours d'échantillonnage sont ensuite optimisés et sélectionnés au regard des critères retenus. Toute l'originalité du sujet se trouve dans la résolution d'un problème agronomique d'échantillonnage en combinant les méthodologies issues de deux domaines scientifiques très différents : des approches stochastiques visant à considérer et caractériser un grand nombre de candidats (les sites potentiels d'échantillonnage) ainsi que des approches d'optimisation informatique pour identifier un échantillon solution parmi un grand nombre de possibilités. Tout cela afin de prendre en compte les contraintes opérationnelles et *in fine* proposer un échantillonnage optimal sur deux plans : minimiser l'erreur d'estimation et l'effort d'échantillonnage.

Les résultats sont obtenus en se basant sur un cas d'étude, celui du rendement en viticulture. Ce cas est intéressant car il cumule des propriétés permettant de rendre l'approche proposée généralisable à un grand nombre de productions végétales : une grande variabilité de la variable d'intérêt au niveau intra-parcellaire, une autocorrélation spatiale, l'existence de données auxiliaires susceptibles d'aider au positionnement des sites de mesures et un environnement très structuré contraignant le déplacement d'un site de mesure à un autre. L'échantillonnage du rendement en viticulture présente des propriétés susceptibles d'être généralisables à d'autres variables d'intérêt, que ce soit en viticulture ou à d'autres productions végétales, moyennant l'hypothèse de l'existence de données auxiliaires pertinentes. Les méthodes développées dans le cadre de ce travail sont donc directement mobilisables pour d'autres problématiques d'échantillonnage en agriculture. Une application directe concerne les autres cultures palissées telles que certains vergers, que ce soit pour estimer le rendement, un statut hydrique, une quantité de sucre ou les propriétés du sol. Les contraintes opérationnelles liées au palissage, telles que prises en compte dans ce travail, peuvent avec des modalités différentes s'appliquer à des cultures non palissées dès lors qu'une contrainte de déplacement est à prendre en compte. En grande culture, cela pourrait par exemple être le cas lorsque l'estimation de la variable d'intérêt est réalisée par un capteur positionné sur une plateforme motorisée et que le déplacement de cette plateforme est contraint par les lignes de semis afin de minimiser l'impact de la prise de mesure sur la culture.

Chapitre 2 : De la littérature à une nouvelle approche prenant en compte les contraintes opérationnelles

2.1 Parcours d'échantillonnage pour les stratégies d'échantillonnage existantes dans la littérature scientifique

Comme évoqué dans le premier chapitre, de nombreuses stratégies d'échantillonnage sont présentes dans la littérature, chacune pouvant être adaptée en fonction du problème spécifique rencontré. Dans le cas du rendement en viticulture, des travaux ont montré que des approches prenant en compte des données auxiliaires à haute résolution spatiale permettait de réduire l'erreur d'estimation du rendement. Les approches proposées correspondent au *model sampling* et au *target sampling* (Carillo et al. 2017 ; Araya-Alman et al. 2017 et 2019). Telles qu'elles sont implémentées dans ces travaux, ces approches se basent sur une liste des ceps (ou ensembles de ceps) qu'il est possible d'observer par échantillonnage sur la parcelle. Dans le reste du document, on parlera de site potentiel d'échantillonnage pour définir les ceps ou les groupes de ceps de la parcelle sur lesquels un échantillon peut être réalisé. Les approches proposées dans la littérature classifient ces sites potentiels d'échantillonnage en sous-groupes selon leur valeur attributaire pour la donnée auxiliaire. Le nombre de groupes est choisi égal à la taille de l'échantillon final désirée. La méthode de classification retenue dans ces articles est une approche basée sur les quantiles. Si k représente le nombre total de sites potentiels présents sur la parcelle et n le nombre de mesures désiré dans l'échantillon final, on utilise les quantiles pour définir n groupes contenant k/n points de mesure potentiels chacun. Un site potentiel est ensuite tiré aléatoirement dans chaque groupe pour constituer l'échantillon final. L'image du bas de la Figure 2.1 montre un échantillonnage basé sur le choix de cinq sites de mesure choisis dans 5 groupes définis à partir de quantiles définis sur une donnée auxiliaire.

Ces méthodes d'échantillonnage basées sur les données auxiliaires permettent d'obtenir des échantillons plus représentatifs des parcelles et tendent à fournir de meilleures estimations. Cependant pour ces approches, il n'existe aucune contrainte liée à la position des sites de mesures sur la parcelle en particulier la question du cheminement imposé par le palissage et des inter-rangs qui en résultent n'est jamais considérée. De la même manière, le point d'accès à la parcelle et la distance qui sépare le point de départ de l'observateur du premier site de mesure ne sont jamais considérés. Les parcours d'échantillonnage associés aux stratégies « optimales » d'échantillonnage telles que proposées dans la littérature peuvent donc nécessiter de parcourir des distances importantes, qui peuvent aller jusqu'à rendre l'échantillonnage irréaliste pour les parcelles les plus grandes ou présentant des rangs de longueur importante.

La Figure 2.1 représente ce phénomène en illustrant les parcours d'échantillonnage résultant d'un *random sampling* (Figure 2.1 haut) et d'un *target sampling (*Figure 2.1 bas*)*. Dans le cas du *target sampling*, les sites d'échantillonnage ont été définis en fonction d'une donnée auxiliaire : indice de végétation obtenu par télédétection). Les trajets associés aux deux méthodes d'échantillonnage sont présentés sur les figures. Ces trajets ont été générés à l'aide d'un algorithme de recherche opérationnelle (cf sous-section 6.3.2.1) prenant en compte les contraintes de déplacement d'un piéton au sein de la parcelle palissée. Les parcours représentés correspondent donc en théorie aux parcours les plus courts permettant de relier l'ensemble des sites de mesure en partant d'un point d'accès de la parcelle et en revenant à ce même point d'accès (en faisant l'hypothèse que ce point est aussi le point de sortie de la parcelle).

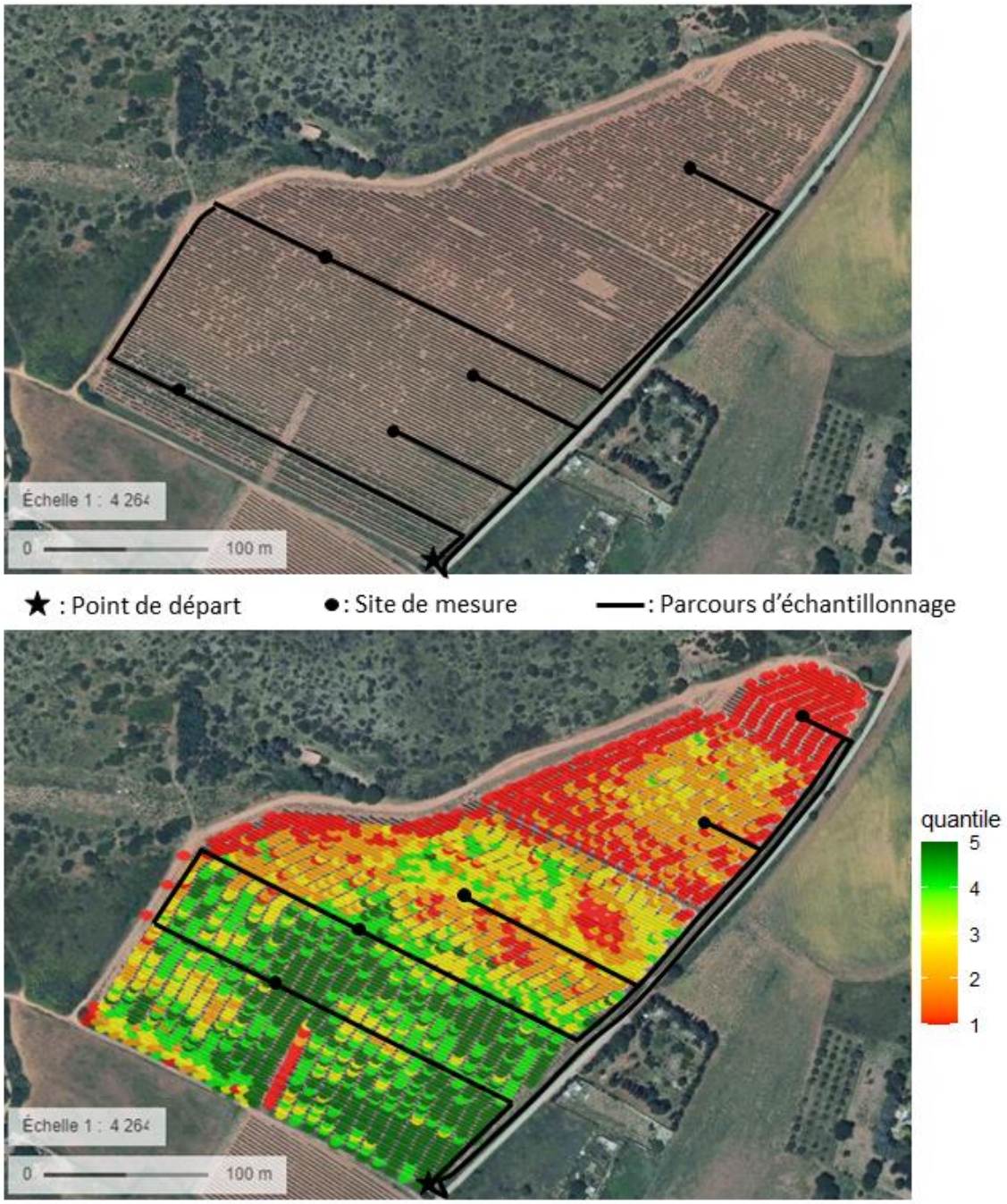

Figure 2.1 : Illustration de deux parcours d'échantillonnage avec 5 sites de mesure résultant de deux méthodes différentes sur une même parcelle, un random sampling (haut) et un target sampling (bas). Les couleurs sur l'image du bas correspondent aux quantiles de la donnée auxiliaire (indice de végétation obtenu par télédétection).

La Figure 2.1 montre clairement les limites d'une stratégie d'échantillonnage qui ne prend pas en compte les contraintes opérationnelles de déplacement au sein de la parcelle. En se focalisant sur le cas de l'échantillonnage stratifié, il apparaît clairement que le choix judicieux d'un ou deux rangs permettrait simultanément de garantir : i) une répartition acceptable des sites de mesures en fonction de la donnée auxiliaire, ii) de minimiser la distance à parcourir entre les sites d'échantillonnages et de iii) proposer une répartition des sites de mesure mieux adaptée au point d'entrée et de sortie de la parcelle.

2.2 Echantillonnage et sélection aléatoire de sites de mesure

Il existe donc un problème pour la mise en place de ces approches d'échantillonnage sur le terrain lié à la longueur des parcours générés par les méthodes d'échantillonnage qui ne prennent pas en compte les contraintes de déplacement au sein des parcelles. Généralement, quelle que soit l'approche d'échantillonnage retenue, le choix des sites est conditionné en totalité ou en partie par une part liée à un tirage aléatoire. Si cela est évident dans le cas du *random sampling*, la part aléatoire apparaît également partiellement avec les approches stratifiées basées sur les données auxiliaires puisque le choix des sites de mesures à l'intérieur des classes ou des groupes est réalisé par tirage aléatoire parmi l'ensemble des sites de mesure potentiels appartenant à chaque classe ou groupe. Cette part d'aléatoire dans le choix des sites de mesure est l'une des causes principales du manque d'opérabilité puisqu'elle tend à augmenter la dispersion des sites de mesure sur la parcelle. Seules quelques approches telles que le *grid sampling* limitent la part d'aléatoire dans le choix de l'échantillon, mais requièrent des échantillons plus complets car elles sont souvent destinées à la cartographie de la variable d'intérêt et non à l'estimation.

Ce recours systématique au hasard peut paraitre surprenant de prime abord. Il est facile d'imaginer qu'il est possible de faire mieux qu'un simple tirage aléatoire. Les tirages aléatoires uniformes restent néanmoins des méthodes simples, faciles à implémenter et dont les propriétés sont très bien connues des statisticiens. Une des principales caractéristiques de ces tirages réside dans leur absence de biais, chaque site ayant la même probabilité d'être tiré. Ils favorisent également l'indépendance des observations. La loi des grands nombres assure que les propriétés de l'échantillon convergent vers celles de l'ensemble de la parcelle lorsque la taille de l'échantillon augmente (Borel, 1909). Le recours aux tirages aléatoires apporte également une importante variabilité, pour une parcelle donnée, les probabilités d'obtenir deux fois le même échantillon tendent rapidement vers 0 lorsque le nombre de sites de mesure potentiels augmente. Cette variabilité tend à augmenter les erreurs d'estimation.

2.3 Vers une intégration du coût de l'échantillonnage

2.3.1 Echantillonnage opérationnel et random sampling

L'échantillonnage opérationnel généralement utilisé pour l'estimation du rendement viticole correspond à une version « altérée » du traditionnel *random sampling*. Notons que la littérature scientifique est inexistante sur le recensement des pratiques réelles des professionnels. Les considérations évoquées ci-après sont donc le résultat d'échanges effectués avec des professionnels confrontés à la mise en œuvre de l'estimation du rendement en viticulture. Ces échanges mettent en évidence des approches communes visant à intégrer contraintes de déplacement en choisissant un nombre très limité d'inter-rangs, généralement au nombre de deux avec un inter-rang montant et un inter-rang descendant. Cette organisation permettant de limiter les déplacements dans la parcelle en positionnant les sites d'échantillonnage sur un aller-retour.

La Figure 2.2 illustre la mise en place telles approches sur trois parcelles de l'INRAE Pech Rouge (Narbonne, France). Pour chaque parcelle, on dispose de la liste des ceps pour lesquelles une donnée auxiliaire de NDVI est disponible. Ces données sont extraites d'image aérienne de résolution $1\ pixel = 0.25m^2$. Une donnée de rendement est générée pour chaque cep en utilisant un modèle linéaire selon l'approche présentée dans la section 3.3.5 du chapitre 3. On s'assure d'une corrélation de 0.5 entre NDVI et donnée de rendement.

Deux approches d'échantillonnage sont comparées, un échantillonnage aléatoire simple et un deuxième échantillonnage aléatoire pour lequel les sites de mesure se concentrent sur deux inter-rangs non-consécutifs choisis au hasard permettant de simuler les pratiques usuelles des professionnels et correspondant à un aller-retour sur la parcelle. La taille des échantillons varie entre $n = 5$ et $n = 10$. Pour chaque échantillon sont calculées : la longueur du parcours et l'erreur d'estimation. Sur les courbes, chaque point représente la moyenne de 3000 répétitions (1000 par parcelle).

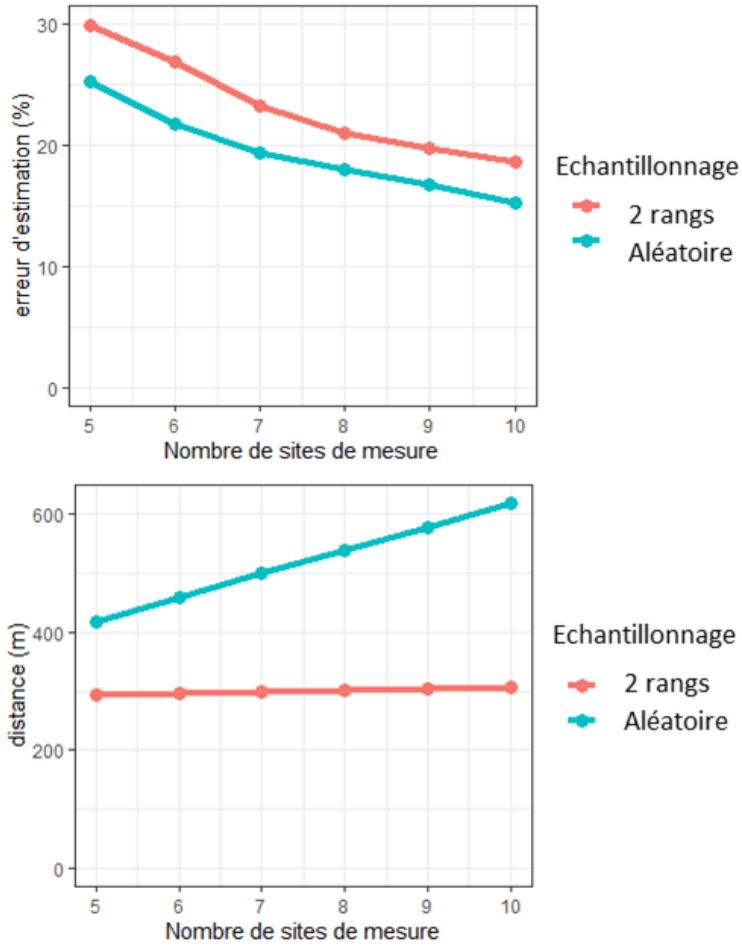

Figure 2.2: (haut) Eerreurs d'estimation et (bas) distances de parcours pour un échantillonnage aléatoire simple et un échantillonnage aléatoire limité à deux inter-rangs (aller-retour sur la parcelle).

La Figure 2.2 (haut) met clairement en évidence (même s'il s'agit de cas de rendement simulé) qu'un échantillonnage basé sur deux inter-rangs entraîne des erreurs d'estimation plus importantes de 4 à 5 % par rapport à un *random sampling* simple. En contrepartie la longueur des parcours d'échantillonnage (en ne prenant en compte que le temps de déplacement) est évidemment moindre lorsque les échantillons sont répartis aléatoirement sur deux rangs. Logiquement, le temps de parcours entre site de mesure n'augmente pas avec le nombre de sites de mesure (avec une longueur d'environ 300m) lorsque ces derniers sont répartis sur deux rangs (pour rappel, le temps de mesure sur chaque site n'est pas pris en compte puisque pour un nombre d'échantillons donné, il est indépendant de la méthode d'échantillonnage). L'écart de distance entre les approches varie entre 100 m pour $n = 5$ et 300 m pour $n = 10$. Ces résultats mettent en évidence le nécessaire compromis qu'il est nécessaire de réaliser entre erreur d'estimation et distance de parcours.

2.3.2 Vers une intégration du coût de l'échantillonnage aux approches de model sampling

Dans le cas où un indice de végétation est disponible avec une haute résolution spatiale, les travaux de Carillo (2016) ont montré que les approches de *model sampling* permettaient de réduire l'erreur d'estimation du rendement. La première piste étudiée au travers de cette thèse s'est basée sur ces approche pour la mise en place d'une nouvelle stratégie d'échantillonnage. Tout comme certaines pratiques opérationnelles intègrent le coût de l'échantillonnage aux approches classiques, l'objectif a été de proposer une version améliorée du *model sampling* proposant un meilleur compromis entre longueur du parcours et erreur d'estimation.

Plutôt que de limiter directement le nombre d'inter-rangs à parcourir ou de privilégier certains sites à d'autres dans un objectif de réduction de la distance de parcours, il a semblé plus pertinent d'intégrer directement la longueur du parcours comme une grandeur à réduire ou minimiser au moment de la sélection des sites de mesure. Concrètement, la partie aléatoire du *model sampling* est modifiée afin d'optimiser la longueur du parcours.

2.4 Première mise en œuvre de l'approche

Mettre en place une telle approche requiert, en plus des méthodes stochastiques traditionnelles, le recours à des outils d'optimisation. Deux types d'outils ont été mobilisés dans ce but, la mise en place de ces méthodes est décrite dans le chapitre 6. La version présentée dans le chapitre suivant (Chapitre 3 :) met en œuvre la programmation par contraintes afin de rechercher la meilleure solution au problème posé. Dans une première instance d'implémentation, la programmation par contrainte a été choisie pour au regard de plusieurs critères décrits ci-après. Tout d'abord, sa souplesse a permis de faire évoluer la méthode au fur et à mesure de l'avancée de la thèse, d'intégrer de nouvelles contraintes en l'adaptant à plusieurs types de données. Enfin son efficacité rend possible l'obtention d'un résultat dans un temps limité et même de trouver la valeur optimale sur des instances de taille raisonnable.

Le résultat de cette implémentation est présenté dans le chapitre suivant (Chapitre 3 :). Il s'agit d'un article proposé et accepté dans la revue *Precision Agriculture* (Oger et al. 2020). Cet article correspond à une version étendue d'un article présenté lors de la conférence ECPA 2019 (Oger et al. 2019). Ce dernier compare différentes variantes du *model sampling* reposant sur des méthodes de classification différentes. L'algorithme des K-means (Steinhaus 1956 ; MacQueen 1967) et une variante de la méthode de Kennard & Stones (1969) y sont ainsi comparés à l'approche plus traditionnelle des quantiles. L'approche de Kennard & Stone a pour objectif de trouver les meilleures observations pour la calibration d'un modèle linéaire. Les valeurs sélectionnées par cette méthode sont ici utilisées comme valeurs centrales d'un intervalle dont la largeur dépend d'un paramètre noté β. Les détails de cette approche sont donnés dans la note de conférence présente en annexe (Annexe, Article de conférence : ECPA 2019) de ce document. Deux méthodes de classification ressortent des résultats obtenus : celle basée sur la méthode de Kennard & Stone et la méthode des quantiles.

Le Chapitre 3 part de ces résultats en se basant sur l'approche des quantiles (privilégiée pour sa simplicité et sa faible sensibilité aux données aberrantes). Le chapitre 3 intègre une étude complémentaire sur des données simulées. Ces simulations permettent de tester les approches sur des jeux de données en contrôlant certaines caractéristiques (variance, structure spatiale, etc.) ce qui permet de mieux étudier la sensibilité de la méthode à ces caractéristiques.

Chapitre 3 : Combining target sampling with within field route-optimization to optimise on field yield estimation in viticulture

B. Oger[1-2], P. Vismara[2-3] and B. Tisseyre[1]

[1] *ITAP, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[2] *MISTEA, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[3] *LIRMM, Univ. Montpellier, CNRS, France*

3.1 Abstract

This paper describes a new approach for yield sampling in viticulture. It combines approaches based on auxiliary information and path optimization to offer more consistent sampling strategies, integrating statistical approaches with computer methods. To achieve this, groups of potential sampling points, comparable according to their auxiliary data values are created. Then, an optimal path is constituted that passes through one point of each group of potential sampling points and minimizes the route distance. This part is performed using constraint programming, a programming paradigm offering tools to deal efficiently with combinatorial problems. The paper presents the formalization of the problem, as well as the tests performed on 9 real fields were high resolution NDVI data and medium resolution yield data were available. In addition, tests on simulated data were performed to examine the sensitivity of the approach to field data characteristics such as the correlation between auxiliary data and yield, the spatial auto-correlation of the data among others. The approach does not alter much the results when compared to conventional approaches but greatly reduces sampling time. Results show that, for a given amount of time, combining model sampling and path optimization can give estimation error up to 30% lower for a given amount of time compared to previous methods.

Keyword: Yield estimation · Sampling · NDVI · Constraint programming · Simulation · Spatial data · Viticulture

3.2 Introduction

In order to optimize harvest organization and quality management, the wine industry needs to know the yield of each vine field. Ideally, yield has to be estimated a few days before harvest with a relative error of less than 10 % (Carrillo et al., 2016). Although models have been developed to forecast the yield at the regional level (Cristofolini and Gottardini 2000), their results were not precise enough to manage logistic issues in relation to harvest operations at the farm or at the winery level. Therefore, precise estimation of vine field yield always requires fruit sampling and counting (Clingeleffer et al. 2001). This estimation must be carried out quickly (few minutes per field) at a time when the workload at harvest or for the preparation of the harvest is critical. Practical constraints, like the time available to visit all the fields before harvest, limit the number of sampled sites per field. Therefore, yield estimation is based on a low number of sites sampled (4 to 5 sites per field) where yield components (number of clusters, number of berries per cluster, mean berry weight) are manually measured by a practitioner. Due to these practical constraints and the high within-field variability of grape yield usually observed (Taylor et al., 2005), the small number of observations results in high errors in yield estimation (generally around 20 to 30%).

Recent works (Carrillo et al., 2016, Uribeetxebarria et al., 2019, Arnó et al., 2017) have shown the interest of integrating auxiliary data to improve sampling strategies and yield estimation for perennial crops. Among possible auxiliary information, vegetation index derived from multispectral airborne

images is of great interest since they permit to characterise the spatial variability of several fields; in one acquisition, with a high spatial resolution (< 1 m.) and at an optimal date. In viticulture, Carrillo et al., (2016) shows the potential of NDVI to drive target sampling of the main grape yield components (bunch number, berry weight, etc.) to improve yield estimation. Although spatial patterns of yield and vegetative expression, estimated by vegetation indices (i.e. NDVI) may not match systematically in all the situations (Bramley et al., 2019). Carrillo et al., (2016) demonstrated, in a dry vineyard of southern France, the value of using NDVI information to determine relevant within field sampling sites selection based on the distribution of NDVI values.

Although interesting, the methodology proposed by Carrillo et al. (2016) presents a significant drawback. Indeed, it does not take into consideration the relative position of the sites to be sampled and the fact that vine fields are structured in rows. This peculiarity implies that rows cannot be crossed, leading to sampling plans optimized in terms of prediction but potentially unrealistic in terms sampling routes and resulting travelled distance (and time) for the operator. This paper proposes a new approach to optimally design within-field sampling routes which takes into account the spatial organisation of the crop (rows) and spatial location of sampling sites. The originality of this approach, called constrained sampling, is to combine statistical and computer methods. It can be decomposed into two steps. In the first step, potential sampling sites are sorted into different groups according to their auxiliary data value in a similar way to traditional targeted sampling. The second step finds an optimal route that passes through one sampling site from each group. A constraint programming solver is used to build an optimal route in terms of travelled distance. This kind of solver has already been used in precision viticulture to solve the differential harvest problem (Briot et al. 2016).

The objective of this work is therefore to propose the resolution of a sampling problem by combining two methodologies from two very different scientific fields: a stochastic approach aimed at considering a large number of representative candidates (the sampling sites) with a constraint programming approach whose objective is to search for the optimal solution among a number of identified candidates. This combination is an interesting scientific question since it is necessary to simplify the exploration of possible candidates in order to be able to calculate an optimal route, while preserving the abundance of information provided by the variability provided high spatial resolution data. The scientific hypothesis of this work is therefore to test and validate the contribution of a combination of methods for spatial sampling in agriculture. In particular, the aim is to verify to what extent the introduction of an optimisation approach decreases the error of the estimates by taking conventional sampling approaches as a reference. It will also examine whether gains in sampling time compensate for decreases in the quality of the estimates. Tests are performed either on experimental data from France or on simulated plots to better characterize its performance under different spatial structures.

3.3 Materials and methods

3.3.1 Sampling sites and selection principles

3.3.1.1 *Overview*

The purpose of constrained sampling (CS) is to select N sampling sites constituting a sampling route in the vine field. Accounting for classical sampling practices in viticulture, N will vary between 5 and 10. It is assumed that there is a finite number of sites on the plot where sampling can be carried out, these sites are called potential sampling sites. For instance, considering the plant as a potential sampling site, a one-hectare vineyard plot planted at 4000 vines/ha consists in 4000 potential sampling sites. For each potential sampling site (PSS), the method assumes that: i) the coordinate of the PSS, ii) the row that the PSS belongs to and iii) the corresponding auxiliary data value are known.

The method requires computation of a distance matrix. This matrix gives the distance between each couple of PSSs. Distance must correspond to the shortest walking distance between the two PSSs. It must take into account the structure of the vineyard. Each PSS, located on one vine rows, can be accessed through two inter-rows (Figure 3.1). As a result, each site appears twice in the distance matrix. If the sites are in the same inter-row, it corresponds to the classical Euclidian distance. If they belong to different inter-rows, the distance is computed considering that the practitioner has to get out the row, to reach the desired rows passing by all ends of intermediate rows and finally to reach the targeted sampling site (Figure 3.1). As rows have two extremities, two different distances can be computed and the shortest one is kept.

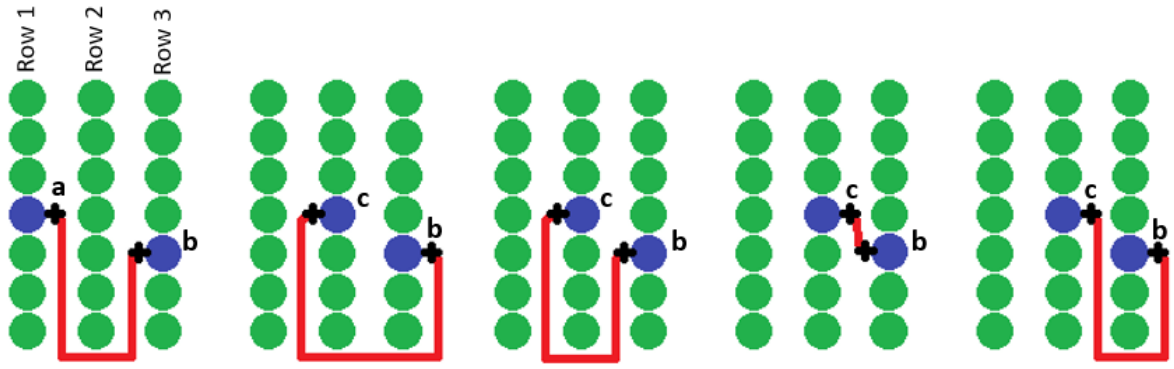

Figure 3.1: Distances across vineyards. The left illustrates how the vineyard structure affect the moving from point a to point b. The others illustrate the four different ways of going from point c to point b.

3.3.1.2 *Representative sampling based on auxiliary information*

The same approach as Carrillo et al. (2016) was used to consider auxiliary data. It is summarised hereafter. The sampling approach aims at calibrating a linear regression which relates the yield (sampled) to auxiliary data. Carrillo et al. (2016) showed that yield components, especially berry weight, are linearly related to auxiliary data (NDVI) in non-irrigated vineyard of south of France. The sampling approach aims at selecting sampling sites (SSs) representatively to build this linear model. This model is then used to estimate yield using all available high-resolution auxiliary data.

The approach proposed in this paper relies on the following principle. PSSs are split into n homogeneous groups (n corresponding to the number of samples) according to their auxiliary data values. Once groups are formed, one PSS from each group is selected in order to optimize the length of the route connecting all selected PSSs. This ensures the PSSs are representative of the field as selected points are spread across the auxiliary data distribution. Quantiles were used to make groups. Related work concluded that this approach was adapted to yield estimation (Oger et al. 2019). Quantiles method guarantees that the groups have the same number of PSS, k being the number of PSSs and n the number of samples required, then each group has $\frac{\mathrm{k}}{\mathrm{n}}$ elements (Figure 3.2).

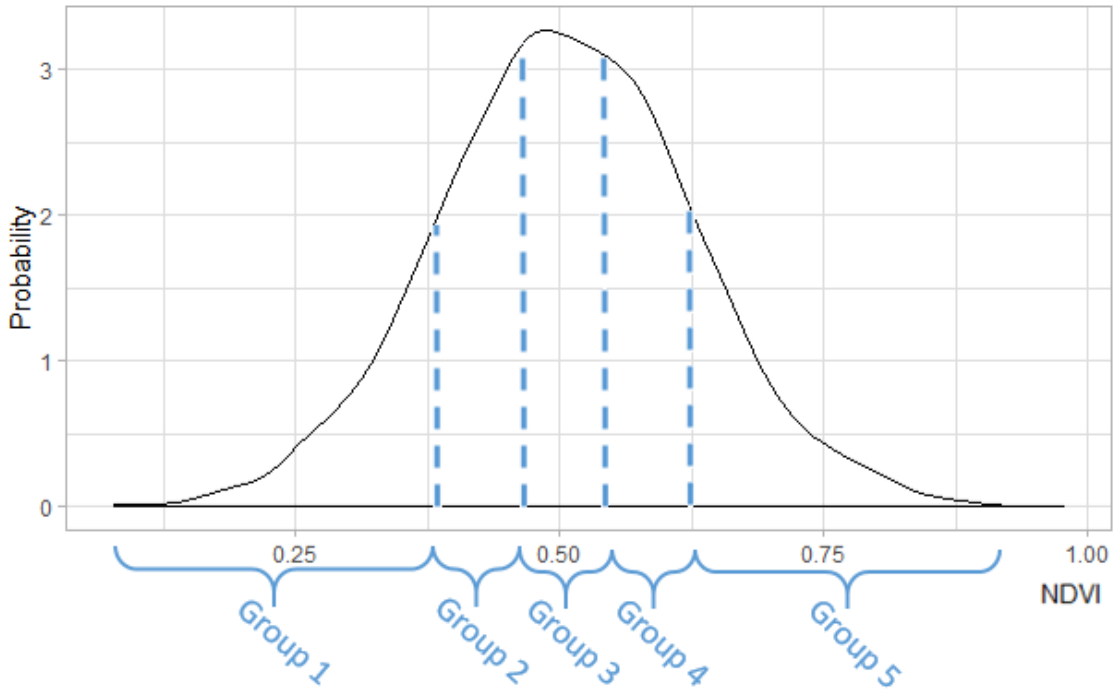

Figure 3.2 : NDVI values for groups with quantile approach and n = 5. Each group contains 20% (100/n) of PSS according to their NDVI values.

3.3.1.1 *Route optimisation*

The second step of the approach consists in selecting n sampling sites (one sampling site per group). These N sampling sites (SSs) must be all different and have to form the shortest possible sampling circuit. There are many possible choices to select these SSs and many ways to order them to form a circuit. It is therefore a highly combinatorial optimization problem. Constraint Programming is one of the programming paradigms able to deal with such problems. It aims at solving a problem expressed as a set of variables and a set of constraints on these variables. Such a problem is called a Constraint Satisfaction Problem (CSP). A Constraint Solver is used to find a solution to the problem that satisfies all the constraints. The efficiency of these solvers relies on the implementation of many methods such as filtering, which allows a quick detection of combinations of values that do not lead to an optimal solution. The interest of constraint solvers lies in their ability to address many types of constraints.

Without going into detail, let $\mathcal{S} = \{1, \dots, K\}$ be the set of PSSs and $\{G_i\}_{i\in(1,\dots,N)}$ the set of quantile groups covering S, formed in the previous step from the auxiliary data. Since all these groups are disjoint subsets, $\{G_i\}_{i\in(1,\dots,N)}$is a partition of $\mathcal{S}$. P_i is defined as the selected site for group G_i. Hence, set $\mathcal{S}_{Selected}$ (the set sampling sites) will be equal to $\{P_i\}_{i\in(1,\dots,N)}$.

The first constraint imposes that all P_i must be different; because all quantile groups are disjoint subsets, it is immediately satisfied in the case of the presented approach. P_0 represents the point of departure and arrival, it is a fixed parameter representing the initial position of the practitioner. The length of the optimum route passing through all the $P_{i\in(0,\dots,N)}$ must be minimum. This is a particular case of vehicle routing problem (VRP) where the goal is not to find a Hamiltonian tour (visiting once every site) but a tour covering only a subset of sites. Recent work about the WeightedSubCircuit constraint (Vismara et al. 2018) has proposed a filtering algorithm that is well adapted to address this type of situation. All these constraints and variables constitute the constraint satisfaction problem. An instance of this problem is built from each dataset and solved with the solver in order to get an ordered

set of sites that form a sampling circuit. The program returns the list of sampling sites, the order in which they are visited and the associated walking distance.

3.3.2 Yield estimation

The aim is to estimate Y, the average yield of the field. For each selected sampling site s ($s \in \mathcal{S}_{Selected}$), $GW(s)$ is the grape weight per vine value. A linear model linking the auxiliary data (AD) to GW is built from these sites (Eq. 3.1):

$$\widehat{GW}(s) = a \times AD(s) + b \qquad Eq.3.1$$

For a given site s, $\widehat{GW(s)}$ represents an estimate of $GW(s)$. The parameters a and b are obtained from a linear regression on the N sites selected by the sampling method.

With $\mathcal{S} = \{1, \dots, K\}$ being the full set of PSSs available, $\hat{Y}_{CS}$, the yield estimate with Constrained Sampling (CS) approach, can be computed from the model using all these PSSs (Eq. 3.2):

$$\hat{Y}_{CS} = mean_{s \in \mathcal{S}}(\widehat{GW}(s)) \qquad Eq.3.2$$

3.3.3 Estimation error

Regardless of the method used, the estimation error is a deviation from the actual yield value (Y), different from 0, and its estimation ($\hat{Y}$), expressed as a percentage (Eq. 3.3)**.**

$$Error\ (\%) = \left|\frac{Y - \hat{Y}}{Y}\right| \qquad Eq.3.3$$

3.3.4 Reference methods

The method is compared to two references:

A conventional random sampling (RS) approach where the N sampling sites are randomly selected among all the PSSs. $\hat{Y}_{RS}$, the yield estimated with RS, is therefore directly calculated from the mean of observed GW values. RS represents what is generally done in practice in terms of yield estimation.

$$\hat{Y}_{RS} = mean_{s \in \mathcal{S}_{Selected}}(GW(s)) \qquad Eq.3.4$$

The second approach is model sampling (MS) whose method principles have been described by (Carrillo et al. 2016). SSs are chosen according to NDVI values. One site is randomly selected for each of the N NDVI quantiles. MS uses a model based on the NDVI/yield relationship, as described in Eq. 3.1. Unlike constrained sampling method, the selection of SS does not consider their position on the plot.

To compare the length of the routes between the different methods, the optimal route between the selected sites was computed for both reference methods. This was done with the R TSP package (Hahsler et al. 2007).

3.3.5 Theoretical fields

3.3.5.1 *Methodology*

Simulated data were used to study the properties and limitations of the approach. Theoretical fields are intended to compare CS to reference methods in a wide range of known situations. Each simulation

aims at providing two variables for each theoretical field: auxiliary observation (i.e. NDVI) and variable of interest (i.e. the yield), both spatially auto-correlated.

The simulation assumed that a main underlying phenomenon (i.e. environmental factors like soil, climate, elevation, etc.) drives the within field variability of the plant response. The simulation process therefore starts by generating a theoretical auto-correlated variable (noted G), representing the spatial variability of the underlying factor. G is simulated as a spatialized Gaussian field with no nugget effect (Figure 3.3.A). Two new variables, respectively $V1$ (Figure 3.3.B) and $V2$ (Figure 3.3.E), were derived from G by adding a non-auto-correlated noise following a normal centred distribution of respective variances σ_{V_1} and σ_{V_2} (Eq. 3.3.5 and Eq. 3.3.6). V1 and V2 are therefore intrinsically correlated with each other, the level of correlation depends on σ_{V1} and σ_{V2}. In the followings, V1 will be used for the variable of interest while V2 will be used for the auxiliary variable.

$$V1_i = G_i + \mathcal{E}_{V1_i} \; with \; \mathcal{E}_{V1_i} \sim N(0, \sigma_{V1}^2) \qquad Eq.\,3.5$$

$$V2_i = G_i + \mathcal{E}_{V2_i} \; with \; \mathcal{E}_{V2_i} \sim N(0, \sigma_{V2}^2) \qquad Eq.\,3.6$$

$$With \; Cor(V1, V2) \; \epsilon \; [0,1]$$

Four parameters were considered to vary across the theoretical dataset: i) the distance of auto-correlation of G, V1 and V2, defined by the range of the semi-variogram of each variable; ii) the ratio of nugget effect/sill of the V1 variable; iii) the degree of correlation between V1 and V2 and iv) the number of outliers on V2. These parameters fit the diversity of theoretical data by controlling the link between the two variables and their semi-variograms. Hereafter is described the process used to set up these parameters during the simulation process:

σ_{V1}^2, the variance of the noise added to $V1$ was chosen to obtain the expected nugget effect/sill ratio on V1. G having no nugget effect, σ_{V1}^2 is therefore directly equal to the nugget effect of $V1$. It can be directly deduced from the ratio and σ_s^2 (Eq. 3.7):

$$Ratio = \frac{nugget(Y)}{sill(Y)} = \frac{\sigma_{V1}^2}{\sigma_G^2} \qquad Eq.\,3.7$$

$$\Rightarrow \quad \sigma_{V1}^2 = Ratio \times \sigma_G^2$$

σ_{V2}^2, the variance of the noise added to $V2$ was used to calibrate the degree of correlation between $V1$ and $V2$. This was deduced from the covariance and Pearson correlation formulas (Eq. 3.8 and Eq. 3.9).

$$Cov(V1, V2) = Cov(G + \mathcal{E}_{V1}, G + \mathcal{E}_{V2}) \qquad Eq.\,3.8$$

Hence:

$$Cov(V1, V2) = Cov(G, G) + Cov(G, \mathcal{E}_{V1}) + Cov(G, \mathcal{E}_{V2}) + Cov(\mathcal{E}_{V1}, \mathcal{E}_{V2})$$

G,$\mathcal{E}_{V1}$ and $\mathcal{E}_{V2}$ being independent random variables, their covariance are equal to zero. Finally:

$$Cov(V1, V2) = Cov(G, G) = Var(G) = \sigma_G^2$$

With the Pearson correlation formula:

$$Cor(V1, V2) = \frac{Cov(V1, V2)}{\sqrt{Var(V1)} \times \sqrt{Var(V2)}}$$

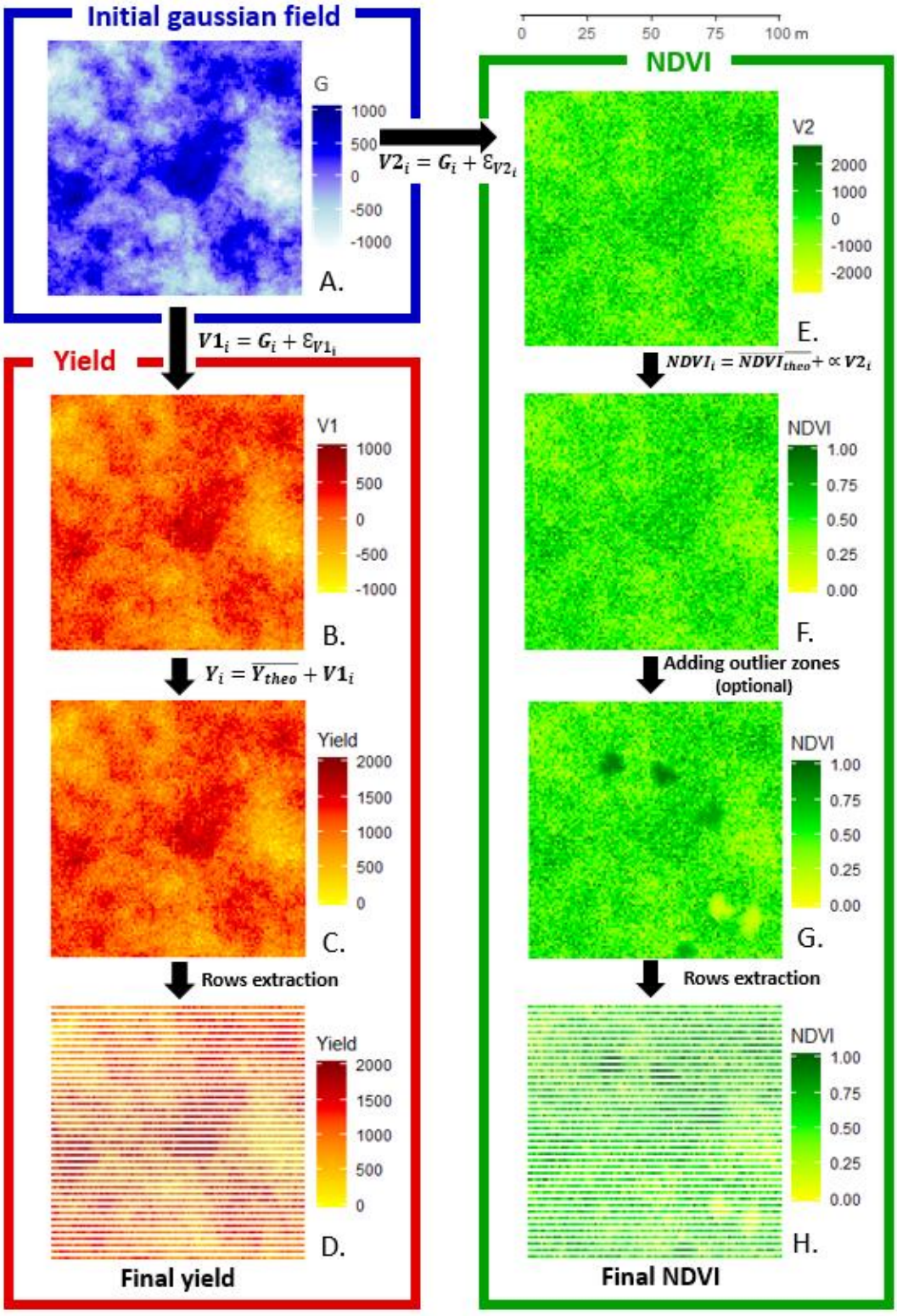

Figure 3.3: Workflow of the theoretical fields simulation process
3.A. Variable G is generated as a fully spatialized Gaussian field
3.B. Variable V1 derived from G by adding a random noise
3.C. Yield variable derived from V1(linear transformation)
3.D. Final theoretical yield data after row extraction
3.E. . Variable V2 derived from G by adding a random noise
3.F. Variable NDVI derived from V2 (linear transformation)
3.G. NDVI variable complemented with outlier zones
3.H. Final NDVI data after row extraction

It results in:

$$Cor(V1,V2) = \frac{\sigma_G^2}{\sqrt{\sigma_{sG}^2 + \sigma_{V1}^2} \times \sqrt{\sigma_G^2 + \sigma_{V2}^2}} \qquad Eq.3.9$$

And finally:

$$\sigma_{V2}^2 = \left(\frac{\sigma_G^2}{\sqrt{\sigma_G^2 + \sigma_{V1}^2} \times Cor(V1,V2)} \right)^2 - \sigma_G^2$$

The variable of interest will be the yield and the auxiliary data, the NDVI. They were respectively derived from $V1$ and $V2$ by a linear change of scale (α) to be centred around the desired average yield ($\overline{Y_{theo}}$) and the desired average NDVI ($\overline{\mathrm{NDVI_{theo}}}$) with the appropriate dispersion (Eq. 3.10.a. and Eq. 3.10.b.) This transformation does not affect the correlation or any of the three other parameters. (Figure 3.3.C, 3.3.F and Eq. 3.10):

$$Y_i = \overline{Y_{theo}} + V1_i \qquad Eq.3.10.a$$

$$NDVI_i = \overline{\mathrm{NDVI_{theo}}} + V2_i \times \alpha \qquad Eq.3.10.b$$

An optional step consists of adding outlier zones on the NDVI simulated maps. Outlier zones intended to represent abnormal phenomenon like weed patches (abnormally strong NDVI) or local diseased vines (abnormally low NDVI) who may locally alter the correlation between yield and NDVI.

The number of outlier zones will vary from 0 to 6 (Table 3.1). The Location of each outlier zones was randomly chosen. Their size varies from 10 to 30 pixels and all these pixels have the same NDVI value taken from $[0.1,0.25] \cup [0.75,0.9]$. All these parameters are drawn randomly for each desired outlier zones. Pixels around the outlier zone were smoothed in order to simulate a short gradient with "normal" surrounding NDVI values (Figure 3.3.G). Introduction of outlier zones may lead to a slight decrease for the correlation parameter.

The final step consisted in extracting the values of both information (Yield and NDVI) corresponding to the rows of the vine. The rows take the values of the nearest pixel. (Figure 3.3.D and figure 3.3.H).

3.3.5.2 *Implementation*

Theoretical fields were designed with an area of one hectare (100m×100m) with a 2.5 m distance between rows which corresponds more or less to typical fields area and plantation density of 4000 vines/ha found in south of France. The resolution was 1 pixel/m². For theoretical yield data, parameters of the simulation were defined so that average yield corresponds to common average yield of the region ($\overline{Y_{theo}} = 1000 g/vine$) and Coefficient of Variation (CV) to previous works: CV = 30%; $\sigma_G^2 = (1000 \times 0.3)^2$ (Krstic et all 1998, Dunn et al. 2000). For theoretical NDVI, the average value was set at $\overline{\mathrm{NDVI_{theo}}} = 1/2$, and the α factor in Eq. 10.b at $\frac{1}{18\,\sigma_G}$ to ensure that all the NDVI values lay within the range of $[0,1]$.

The choice of the different possible values for the four parameters (Table 3.1) was based on observations from literature in precision viticulture (Bramley et al. 2019, Bramley et al. 2004, Hall et al. 2010, Li et al. 2017, Taylor et al. 2005, Tisseyre et al. 2008). For each parameter, three possible levels were defined (Table 3.1), encompassing a range of variability allowing to account for the existing diversity in vineyard fields. Each parameter will vary individually by setting the others to their default

values, indicated in bold in Table 3.1. This procedure will ensure to test the effect of each parameter on the sampling results.

Table 3.1: Values for theoretical field parameters (default values in bold font)

Parameter	Low	Medium	High
Range (m)	10	**20**	40
Ratio Nugget effect / Sill	1/10	**1/3**	1/2
Pearson correlation coefficient	0.1	**0.4**	0.7
Number of Outlier zones	**0**	3	6

3.3.6 Real data

Real fields used to test the method come from INRA Pech-Rouge (Narbonne, France). The experiment and the data base was detailed by Carrillo et al., (2016). It is briefly summarised hereafter. NDVI values from 9 different vine fields were considered. All of them are non-irrigated and exposed to Mediterranean climate with precipitation occurring during spring with hot and dry summer. The characteristics of each plot are shown in Table 3.2.

Table 3.2: Description of the experimental fields. Nugget effect could not be estimated because of the resolution of yield data.

Field	Area (ha)	Variety	Number of Potential Sampling Sites	Range of semi variogram Yield (m)	Pearson correlation coefficient (NDVI/yield)
P22	1.72	Syrah	45	21.14	0.13
P63	1.33	Syrah	42	7.37	0.28
P65	0.69	Syrah	33	27.71	0.86
P76	1.14	Carignan	37	22.20	0.39
P77	1.24	Syrah	19	9.25	0.48
P80	0.54	Syrah	40	20.34	0.63
P82	1.15	Syrah	53	21.69	0.47
P88	0.85	Syrah	21	14.08	-0.04
P104	0.81	Carignan	23	12.96	0.18

NDVI values were derived from a 1 m. resolution multi-spectral image taken the 31th of August 2008 by Avion Jaune (Montpellier, Hérault, France). The spectral regions captured in the images were: blue (445–520 nm) green (510–600 nm), red (632–695 nm) and near-infrared (757–853 nm). From these 1 m square image pixels, aggregation method described in (Acevedo-Opazo et al. 2008) was used to obtain 9m square image pixels reducing the effect of canopy discontinuity and bare soil on measured values. NDVI was finally computed from processed images according to Rouse et al. (1973). Mechanical or chemical weeding was performed over the inter-row spacing; therefore, row cover crop did not affect much NDVI values.

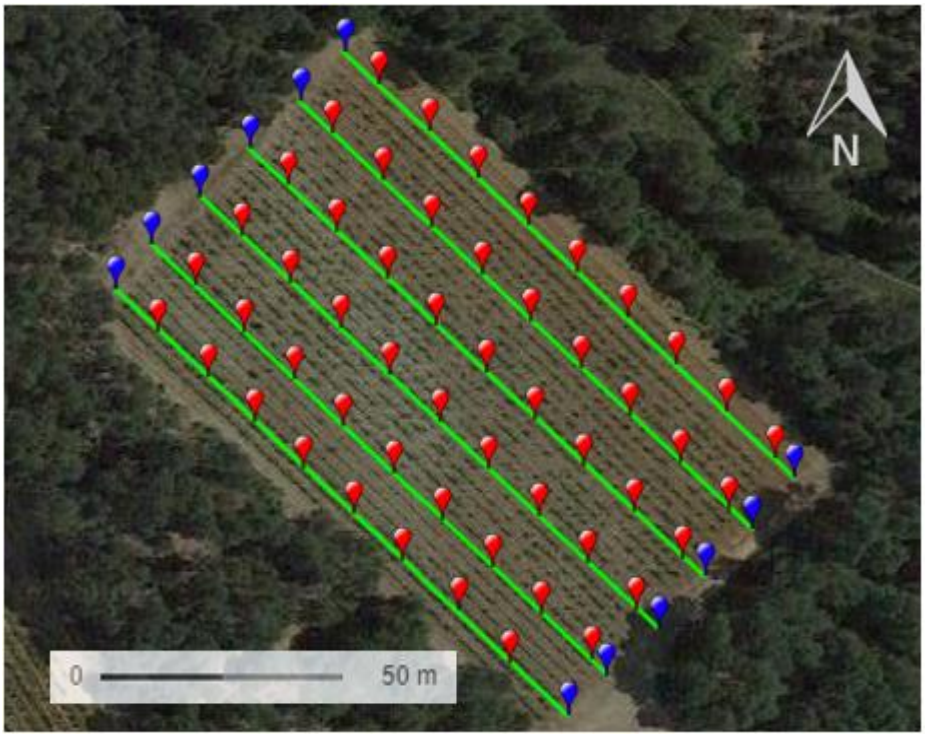

Figure 3.4: INRA Pech Rouge plot with row edges in blue and potential sampling sites in red

PSSs were selected regularly over the fields with measurement made on each node of a 15m² width sampling grid (Figure 3.4). At each node, yield was measured on 5 consecutive vines in the row and average yield was affected to the location corresponding to the central vine. The final data base was a set of 313 sites over the 9 different fields. It is noteworthy that, unlike simulated fields, the number of PSSs is reduced for real fields. Indeed, since it was not practically possible to measure the yield on each vine stock, the consequence is that the number of PSSs depends on the number of available measurement sites, therefore PSSs per site varied from 19 to 45. Each PSS was then characterized by a Grape Weight per vine value (GW) and a NDVI value.

For each field, the average of all available measured GW values was used to estimate the average yield of the field (Y).

3.3.7 Implementation

The core of the sampling approach was written in Java, the program used the Choco solver (Prud'homme et al. 2016). The calculations to obtain the distance matrix are made with Python. Theoretical fields, quantile classification according to their value for auxiliary data, estimation errors, and route distance were computed with R. Packages "gstats" and "sp" were used to generate Gaussian fields for the G variable.

As explained in the description of the constraints, the approach presented here takes into account the starting site of the practitioner to include it in the sampling circuit. Varying the starting site thus changes the result that will be obtained. In order to increase the number of situations tested for real data, this starting site is positioned on different ends of row across the vineyards. The approach was then applied to 86 different situations instead of 9. Results for the different starting sites were averaged for each field. For theoretical data, thirty simulations per set of parameters are tested (270 in total). CS was applied once to each simulated plot. The starting site is located on one of the corners of the plot. The two outer rows on each side and the first three vines of each row cannot be selected

as, in practice, they are subject to border effects. For both theoretical and experimental data, RS and MS were applied 1000 times. These repetitions were possible as they rely on random or partially random selection of SSs. The following figures are based on the average of the results obtained on each field, the result of each field being the average across repetitions with different starting site.

3.4 Results and discussion

3.4.1 Sampling & vine diversity

3.4.1.1 *Number of sampling sites*

Figure 3.5 presents the results obtained for the three sampling approaches, constrained sampling (CS), model sampling (MS) and random sampling (RS) for simulated fields with parameters set at their default values (Table 3.1). The different sampling approaches were tested on each simulated plot with a number of measurement sites ranging from N=5 to N=10. Results represent the averages obtained with thirty simulated plots.

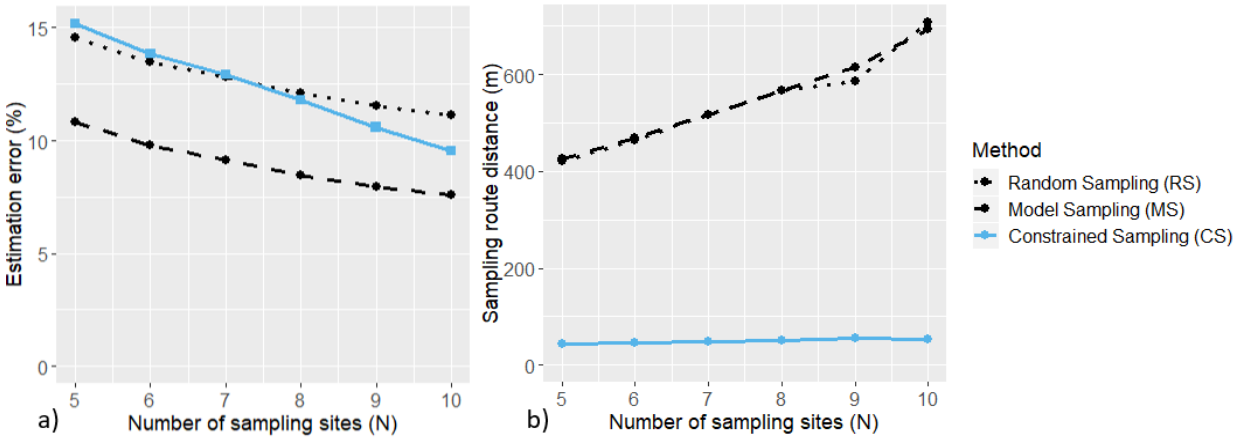

Figure 3.5 : Results for theoretical data with default parameters in function of the number of sampling sites; a) Estimation error, b) Sampling route distance.

All the sampling methods follow the same trend with a decreasing error as the number of sampled sites increases. This result is logical, and consistent with the literature. As Carrillo et al (2016) have already shown, taking into account auxiliary data, MS approach slightly improves the quality of yield estimation compared to a *RS* (Figure 3.5.a). With default values, it seems that the MS approach proposed by Carrillo et al. 2016 gives the best results in terms of estimation error. CS and RS both present similar errors whatever the number of SSs. Observed errors with CS are higher than for *MS* (i.e. without constraints). This result may be logical considering that the addition of the constraints may limit possibilities when choosing among the PSS.

Figure 3.5.b clearly illustrates the gains brought by *CS* in terms of travel distance across the vineyard. Logically, the travelled distance within the plot increases linearly with N, the number of SSs. Travel distances with CS are at least 85% better than with other sampling methods. Distances are also less sensitive to the increase in the number of SSs with CS. This is explained by the gain brought by considering this criterion when selecting SSs. In view of these first results, CS offers a compromise between MS and distance criterion optimization, with a higher estimation error in favour of a significant reduction in distance.

With the hypothesis of a walking speed of 0.9m/s and 60s needed per SS, Table 3.3 shows that MS and CS perform better than RS. For a given amount of time, CS allows a higher number of observations to be made compared to other sampling approaches and therefore the lowest estimation error.

Table 3.3: Sampling time and estimation error for RS, MS and CS on simulated data.

N	Random Sampling (RS)		Model Sampling (MS)		Constrained Sampling (CS)	
	Error (%)	Time (s)	Error (%)	Time (s)	Error (%)	Time (s)
5	14,6	767	10,8	773	15,2	347
6	13,5	877	9,8	881	13,9	411
7	12,8	993	9,1	994	12,9	474
8	12,1	1111	8,5	1111	11,8	537
9	11,5	1192	7,9	1223	10,6	603
10	11,1	1387	7,6	1371	9,5	660

3.4.1.2 *Impact of the semi-variogram range*

Figure 3.6 shows the estimation error and the distance results obtained with the 3 sampling approaches tested on theoretical fields with varying range (10m, 20m, 40m) of semi-variograms. Varying the range does not affect significantly the estimation error in function of the number of SSs compared to previous results. MS still presents the best estimation error compared to CS and RS (Figure 3.6.a). For a range of 40 m, CS seems more erratic with the highest error for a low number of SSs (N > 8) and an estimation error which tends to the error observed with MS for a higher number of SSs (N > 8) As for estimation error, results of distances associated with RS and MS do not seem to be affected by the range parameter either (Figure 3.6.b). The increase in range is nevertheless associated with a slight increase in distance with CS.

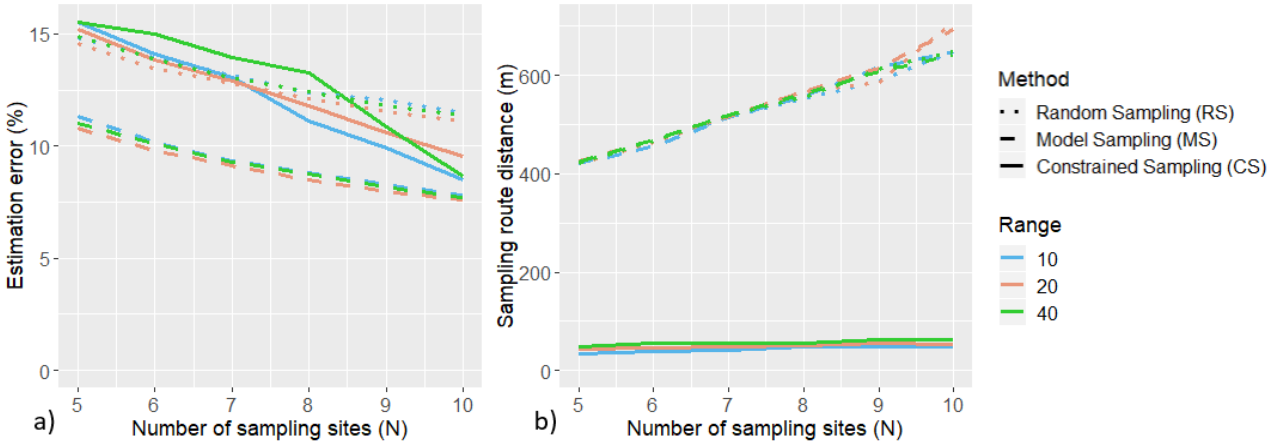

Figure 3.6: Effect of the semi-variogram range on sampling strategies.

Figure 3.7 gives an illustration of the sampling route obtained with the method for plots with different ranges (10, 20 and 40m).

The method always promotes the measurement sites in the immediate vicinity of the starting site (0.0 coordinates). This is expected as the method aims to minimize travel time. A robustness study (result not shown) showed that the results were similar regardless of the starting site, the method always promote measurement sites close to the starting point whatever the plot characteristics. Figure 3.7 shows that the average distance between sampling sites tends to increase with the plot range. On

these plots, the maximum distance between sampling points corresponds approximately to the range of the plot; it is of 40 m, 30 m and 10 m respectively for plot range of 40 m, 20 m and 10 m. This result is expected since as the range increases, the distance to find a higher diversity of values also increases leading to longer sampling routes for larger ranges. This result is consistent with that of Figure 3.6.b.

One surprising aspect, at first glance, is the close proximity of the sampling points. This characteristic is related to the nugget effect, which introduces erratic variance and high variability over short distances. This erratic variance makes it possible to find a high variability of values in the immediate vicinity of the starting point and other measurement sites. The sampling method takes advantage of this variability to minimize travel time. It should be remembered that in these simulations, the nugget effect was defined according to literature figures in precision viticulture. It is rather high since it represents more than 30 % of the plot variability.

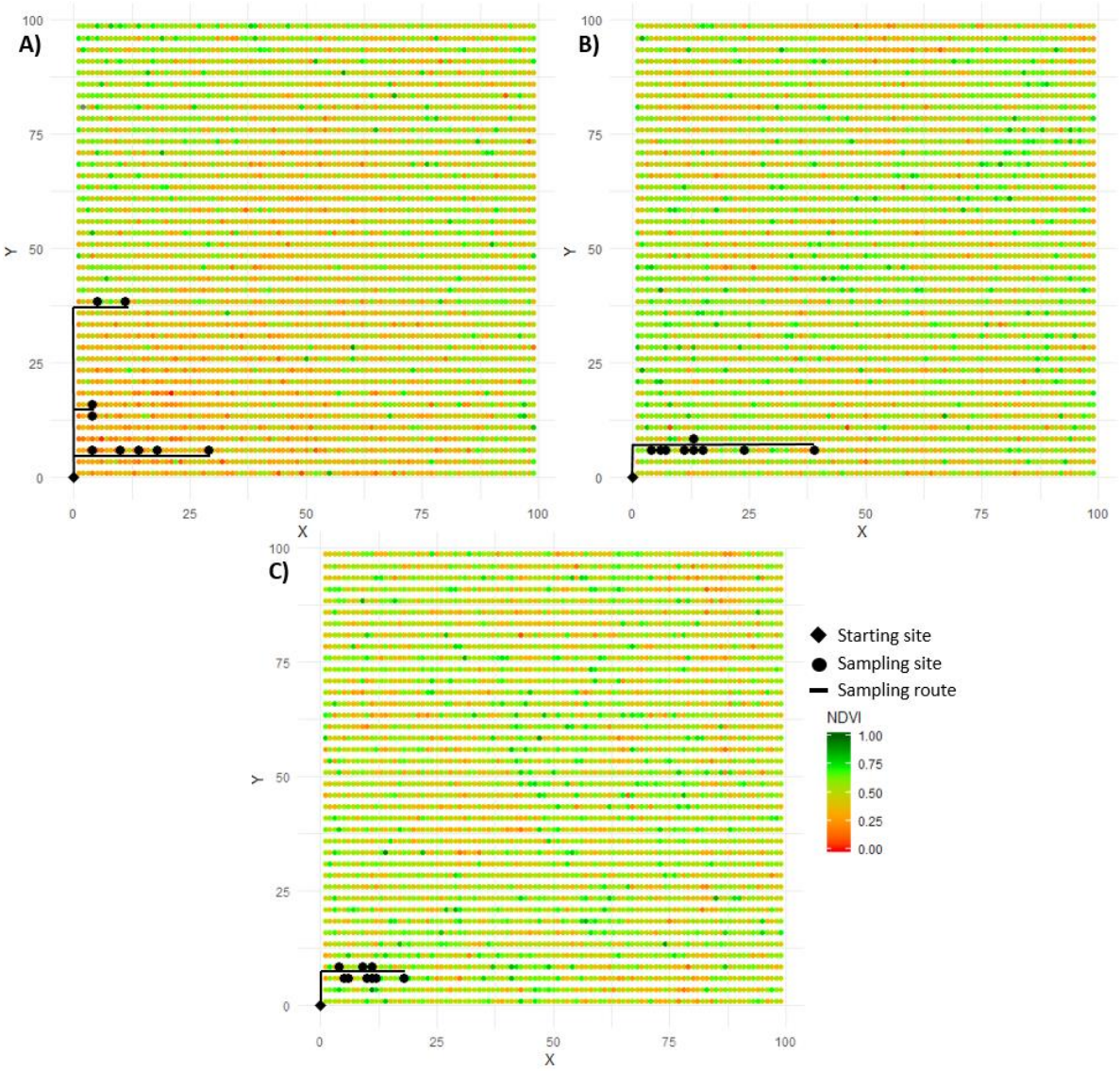

Figure 3.7: Illustration of sampling routes for N = 9 and three different ranges: A) Range = 40m; B) Range = 20m; C) Range = 10m

A decrease in this nugget effect leads to longer sampling route. For example, if the nugget effect is set to 0 (no erratic variability), then the sampling distance is longer since the distance needed to find representative values necessarily increases. Conversely in the case of no spatial autocorrelation, the method chooses contiguous independent measurement sites on a same row and the sampling distance

is in this case very short (result no shown). Figure 3.7 therefore represents the result of optimal sampling that takes into account the combined effect of the nugget effect and the range considering realistic figures of spatial variability in viticulture.

3.4.1.3 *Impact of the ratio nugget/sill*

Figure 3.8 shows the impact of the ratio for the three sampling strategies. Whatever the sampling method and the number of sampling points, an increasing ratio affect the estimation error (Figure 3.8.a). This result is logical since the increase in the ratio corresponds to an increase in the proportion of erratic (non-autocorrelated) variance in the total variance of the NDVI. Thus, sampling methods tend to select SSs whose NDVI value is not necessarily correlated to the expected yield value. Note that for a high ratio (ratio = 0.5), the estimation error with CS is more affected compared to other approaches. CS approach appears to be more sensitive to high ratio values. Indeed, the short-range variability introduced by higher erratic variance increase the chance to have sites in different quantile groups close to each other. The CS approach that minimizes the route from the starting site might select close SSs. Close SSs often provide redundant information, which leads to an increase in estimation error. On the other hand, the ratio does not seem to have any significant influence on the length of the sampling circuits (Figure 3.8.b).

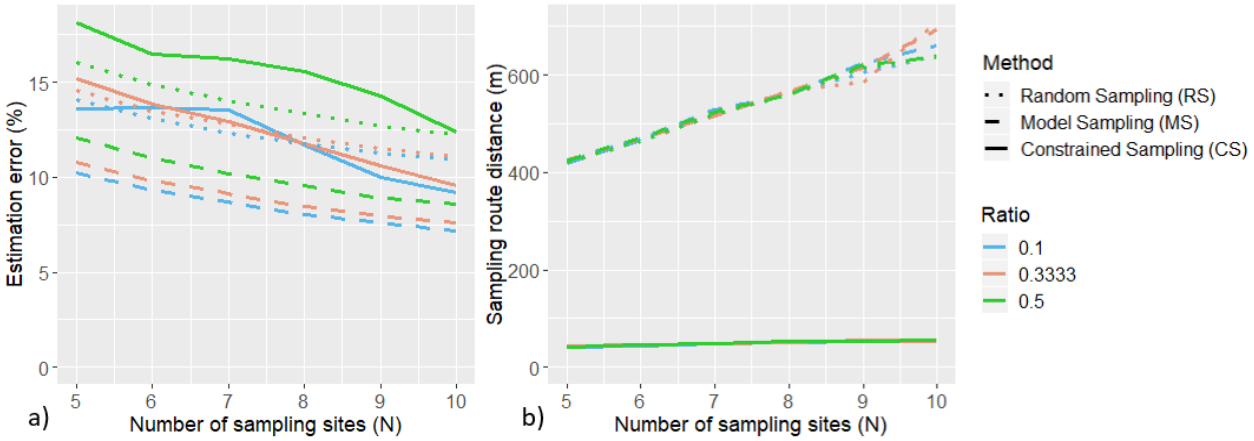

Figure 3.8: Effect of the proportion of the ratio nugget/sill on sampling strategies. a) Estimation error & b) Sampling route distance

3.4.1.4 *Impact of the correlation level with auxiliary data*

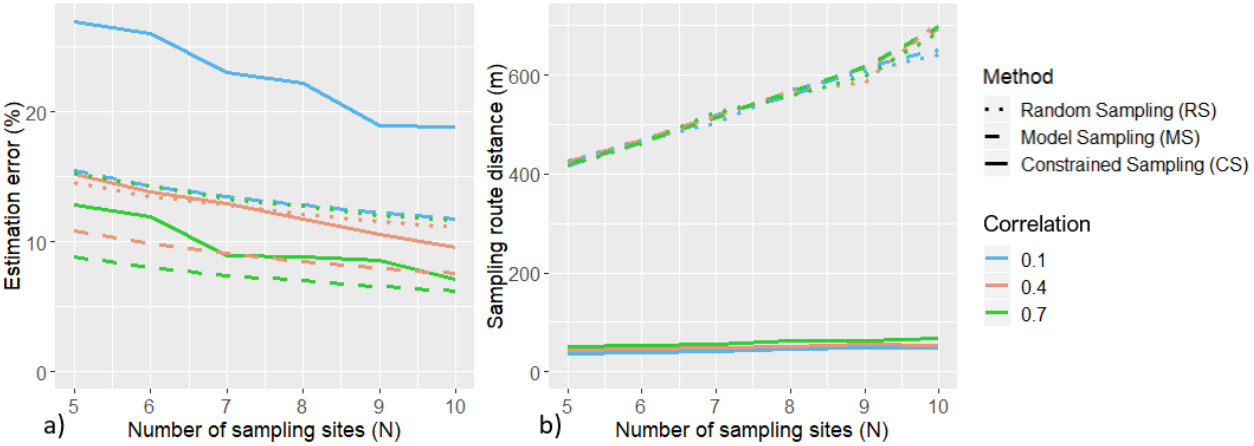

Figure 3.9: Effect of the correlation between NDVI and yield on sampling strategies. a) Estimation error & b) Sampling route distance

As expected, RS is not affected by a low level of correlation since SSs selection is not based on the NDVI data (Figure 3.9.a). This is not the case for MS and CS which both show lowest estimation errors when correlation between yield and NDVI is high while they show high estimation errors decreases when the correlation decreases. Since, MS and CS use the relationship between the two variables, it was expected that the quality of prediction decreases when this relationship is weakened. When the correlation is close to 0, MS tends to have the same results as RS while CS presents the worst estimation error. For the theoretical dataset, a very low correlation corresponds to a strongly noisy NDVI (Eq. 3.9). The resulting short range erratic variability in NDVI could explain this result for CS. It follows the same phenomenon as already observed in the previous section for an increasing ratio. These results highlight the sensitivity of CS to noise in NDVI for the selection of SSs aiming at optimising the distance. As for previous result, sampling distance does not seem much affected by the correlation, apart for CS for which it seems to slightly decrease with the correlation (Figure 3.9.b). This result supports the idea of a decreasing correlation allowing the CS approach to find SSs closer to each other.

3.4.1.5 *Impact of outlier zones*

The addition of local outlier (Figure 3.10.a) seems to slightly affect all methods, including the RS which should not be affected by changes in auxiliary data. The effect of outlier zones is still more important on the CS. It is difficult to draw conclusions as the effect is not proportional to the number of outlier zones. As it could be expected, the length of the sampling circuits is not affected by these local outlier zones (Figure 3.10.b).

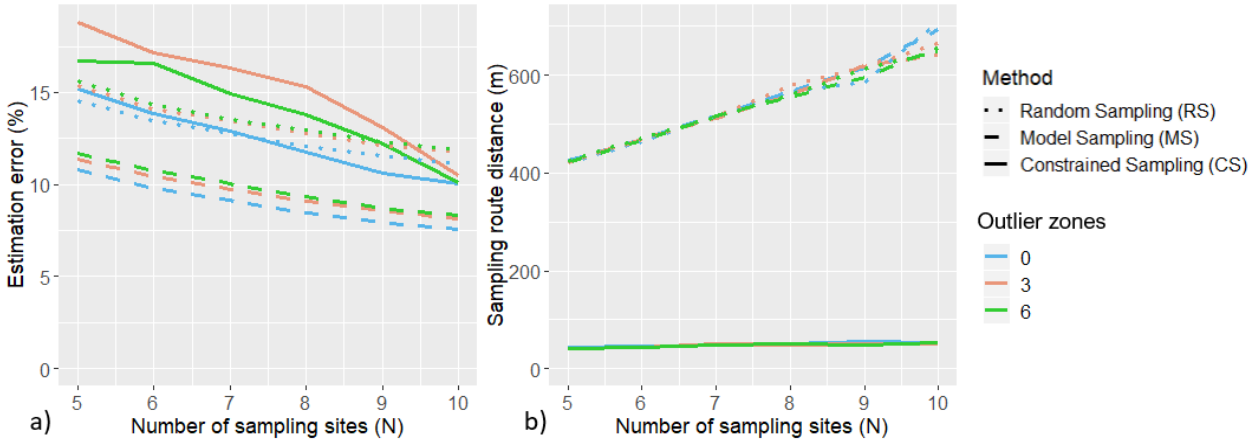

Figure 3.10 : Effect of outlier zones on sampling strategies.
a) Estimation error & b) Sampling route distance

3.4.2 Evaluation of sampling strategies on real data

Figure 3.11 shows the averaged result of the three sampling methods on real fields. The same logical decrease in error as the number of sampled sites increases is observed (Figure 3.11.a) The irregularity of the curves associated with CS can be explained by a smaller number of experiments compared to results obtained with theoretical fields. With N=6 put aside, CS estimation errors are just halfway between those of RS and MS. Considering the characteristic of real fields in terms of ratio and correlation between yield and NDVI (Table 3.2), these results are consistent with the results obtained on the simulated data. CS has been shown to be sensitive to the correlation between yield and NDVI and noise in NDVI values. The average correlation between yield and NDVI being quite large for real data (0.38 on average, Table 1.13.2) and NDVI values being smoothed, real fields present average characteristics close to the default values for theoretical fields. The reduction of erratic local variability in NDVI may have favoured results observed with real fields with CS. Figure 3.11.b illustrates the gains

in terms of travel distance across the vineyard when using CS compared to other sampling approaches. Distance is reduced by approx. 50 % with CS compared to *MS* and *RS*. This result is again consistent with those observed with theoretical data. However, the gain here is much smaller than with theoretical data because of the lower number of PSSs available for real fields Considering the minimum distance between two PSSs is 15m, the distance required to travel through the selected SS is necessarily higher for real fields because of the impossibility for the algorithm to find SSs closer than 15 m to each other, while this was possible for theoretical fields. Overall, this method offers a good compromise between the quality of the estimate and the travel constraint on the plot.

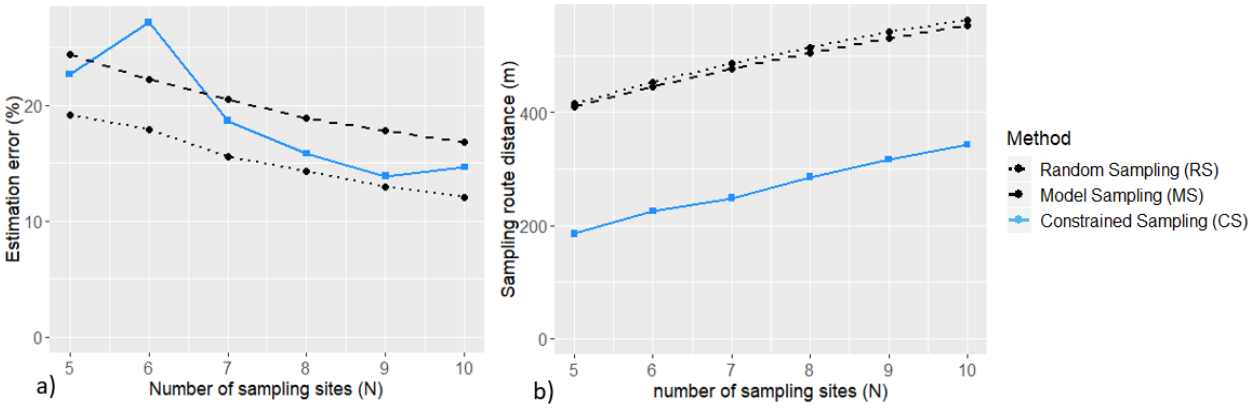

Figure 3.11 : Results on real data.
a) Estimation error & b) Sampling route distance

Table 3.4 compares the performance of the different approaches on sampling time and estimation error for real data. Walking speed is set at 0.9m/s and 60s are required for each SS. As for simulated data, CS perform better (up to 30%) than other approaches for a given amount of time.

Table 3.4 :Sampling time and estimation error for RS, MS and CS on real data.

N	Random Sampling (RS)		Model Sampling (MS)		Constrained Sampling (CS)	
	Error (%)	Time (s)	Error (%)	Time (s)	Error (%)	Time (s)
5	24,4	755	19,2	761	22,7	506
6	22,3	855	17,9	863	27,2	611
7	20,5	949	15,5	961	18,7	696
8	18,9	1042	14,3	1052	15,8	797
9	17,8	1129	12,9	1144	13,9	892
10	16,8	1214	12	1225	14,7	981

3.4.3 Further reflections

The method presented in this paper aims to select SSs accounting for variability highlighted by auxiliary data and the practitioner constraints simultaneously. It is intended to be general enough to be applicable to various combinations of auxiliary data and variable of interest. This type of strategy could be applied to any type of crop where the travel route of operators is constrained by the organization of the crop (trellised structure). It may be necessary to keep in mind the assumptions on which the approach is based on: the fact that there is a correlation between the variable of interest and the auxiliary variable and the relevant auxiliary variable is available with a high spatial resolution.

The choice of the auxiliary variable depends on the variable of interest to be estimated and the pedo-climatic context of the crop. Returning to the application case presented in this paper, it is based on preliminary studies that indicate that the NDVI measured at veraison was a relevant auxiliary variable to guide yield sampling in the specific context of non-irrigated Mediterranean vineyard. However, this same auxiliary variable may not be relevant in other soil and climatic contexts. As a result, the choice of the variables used requires prior knowledge and expertise for the correct implementation of the approach.

Overall, it seems that CS offers a strategy with relevant compromise between estimation error and sampling distance. It significantly improves the performance of the latter criterion for a small increase of the estimation error. For a given amount of time, CS presents better results than reference methods. For a given number of SSs, the gap for the average error between CS and *MS*, which also uses a model for estimating the variable of interest, could be explained by two points. First, the optimization of the distance tends to regroup the SSs around the starting point (Figure 3.6), and it has been shown that this could be a disadvantage when the auxiliary data is noisy, indeed in this case, CS favours the choice of different NDVI values very close to each other and not necessarily correlated with the variable of interest. Conversely, the other methods can be more representative as they are more likely to choose SSs anywhere over the plot. The second reason is that choosing points close to each other reduces the distance of the circuit but might increase the autocorrelation between measurements. Two close sites will provide more or less redundant information depending on the distance between them. It should be noted that on the plot, practitioners generally rely on a random sampling limited to a fraction of the plot (often a pair of rows) as they cannot cover the whole field. The results obtained would then correspond to a degraded version of the random sampling due to the same auto-correlation problem.

A potential solution to minimize the gap between MS and CS would be to set a minimum distance between two SSs. This miminum distance could be derived from spatial auto-correlation of available auxiliary data or historical yield maps. This distance would be determined with the range of the experimental semi-variogram obtained with the auxiliary data. Depending on the value, this could make it possible to avoid or limit the autocorrelation issue between the SSs and to better results of CS for slightly longer sampling times. Another practical option when the auxiliary data is particularly noisy (ratio > 0.33), would be to smooth the data using, for example, a moving window average. This was done on real fields of this study and gave particularly interesting results. The methodology proposed by Tisseyre et al. (2018) could be used to decide on the size of the smoothing window to be used according to the erratic variance that is to be eliminated. Each of these areas for consideration could improve the sampling method presented here with reduced estimation errors and more widely spaced measurement sites.

Other areas for improvement are possible, such as, for example, the ability to integrate multiple auxiliary data or to use more complex models.

3.5 Conclusion

The methodology presented in this paper describes a new approach, Constrained Sampling (CS), for yield sampling in viticulture. The originality of the approach comes from the association of method from Carrillo et al. based on auxiliary data and optimisation algorithms to propose relevant sampling routes in term of estimation error and travelled distance. While the model sampling principle guides sampling points choice considering auxiliary information, optimisation through constraint programming ensures the relevancy of the chosen route in term of walking distance for the practitioner. CS appears however sensitive to unfavourable situations (low correlation, poor spatial

structure, erratic variance of the auxiliary data) while other methods relying on random aspect may fare better. In favourable situations (good correlation between auxiliary data and yield and strong spatial structure), CS gives very good results. The estimation error is close to what is proposed by Carrillo et al. However, CS makes it possible to obtain much shorter sampling in distances and times. This saved time can then be used to increase the number of measurements and the reliability of the estimation.

3.6 Acknowledgements

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (#Digitag).

Chapitre 4 : Le recours à un modèle pour l'estimation de l'espérance d'une parcelle. Une comparaison entre Model sampling et Target sampling.

Dans le cadre de l'estimation du rendement en viticulture, les publications de Carillo et al. (2016) et Araya-Alman et al. (2017 & 2019) montrent que le *model sampling* donne de meilleurs résultats que le *target sampling* (voir section 1.1.5) sur leurs données respectives. En testant ces deux approches sur leurs données, ils font le constat que l'utilisation du *model sampling* aboutit à des erreurs d'estimation moindres. Pour rappel, la principale différence entre ces deux approches se situe dans la manière dont est inférée l'estimation finale, le *model sampling* mobilise un modèle d'estimation là où le *target sampling* base son estimation sur le calcul d'une moyenne des échantillons. Compte tenu des résultats obtenus dans le cadre des travaux de Carillo et al. (2016) et Araya-Alman et al. (2017 & 2019), l'approche d'estimation utilisée dans le cadre de cette thèse s'est focalisée sur le *model sampling* pour effectuer les estimations de rendement. Pour rappel, cette approche était permise car des observations à haute résolution spatiale permettant de décrire la structure spatiale des parcelles étaient disponibles. Toutefois, à notre connaissance, les propriétés respectives du *model sampling* et du *target sampling* n'ont jamais été rigoureusement étudiées dans le cadre de l'inférence de la moyenne d'une parcelle. A notre connaissance, il n'existe donc pas de travaux s'intéressant à cette question dans la littérature scientifique. L'objectif de ce chapitre a pour but d'étudier rigoureusement les propriétés respectives du *model sampling* et du *target sampling* afin d'être en mesure de mieux justifier le choix d'une approche plutôt qu'une autre dans le cadre d'une problématique d'échantillonnage pour l'estimation d'une moyenne parcellaire en production végétale. L'ensemble de ce chapitre s'appuie sur une décomposition statistique des estimateurs en biais et variance qui a été réalisée avec Sébastien Roux (INRAE). Ces travaux seront complémentés pour être soumis à la conférence ECPA (*European Conference on Precision Agriculture*) 2021.

4.1 Inférer la valeur d'un paramètre pour une population à partir d'un échantillon

4.1.1 Utilisation d'une moyenne arithmétique

Pour rappel, en échantillonnage, l'inférence désigne le procédé statistique mobilisé pour généraliser les propriétés de l'échantillon à l'ensemble de la parcelle. Le chapitre 1 fait état des avantages de l'utilisation d'une moyenne pour inférer une espérance. La moyenne est un estimateur efficace de l'espérance pour un grand nombre de fonctions de densité, sa variance correspond alors à la variance minimale que peut atteindre un estimateur. Par ailleurs, le théorème central limite démontre que la moyenne des observations converge vers l'espérance de la loi dont découlent les observations (1.1.3.3). Malgré ces avantages, en échantillonnage, la moyenne présente une certaine sensibilité aux valeurs aberrantes et extrêmes qui peuvent affecter l'estimation de manière importante. Elle est également sensible à la représentativité des mesures qui composent l'échantillon. Un exemple extrême consisterait en un échantillon présentant uniquement des valeurs fortes, l'estimation par la moyenne correspondrait alors à une surestimation de la valeur réelle. L'intérêt du *target sampling* est de proposer une procédure d'échantillonnage visant à limiter ce risque en assurant que les sites soient distribués dans toute la gamme des valeurs attributaires de la parcelle (Kerry et al., 2010). Le choix des sites est alors conditionné (ciblé) au regard d'une donnée auxiliaire disponible.
D'autres outils statistiques sont parfois mobilisés pour l'estimation d'une espérance. Utiliser la médiane permet par exemple de s'affranchir de la sensibilité aux valeurs extrêmes. Dans d'autre cas, le recours à des intervalles de confiance rend possible la prise prendre en compte de l'imprécision associée à l'estimation (section 1.1.3.4). Ces intervalles de confiance sont néanmoins soumis à des

hypothèses sur la distribution des données et ont recours à des estimateurs de la variance de ces distributions.

4.1.2 Utilisation d'un modèle

Les modèles constituent une autre catégorie d'outils pour l'inférence. Ils nécessitent pour fonctionner de disposer d'une donnée auxiliaire associée à chaque site de mesure potentiel de la parcelle. Le *model sampling* tel que défini dans la littérature est centré autour de ce concept. L'inférence par un modèle est réalisée en trois étapes. La première étape consiste à étalonner un modèle entre la variable d'intérêt et la donnée auxiliaire. La nature du modèle peut varier en fonction de la donnée auxiliaire et de l'expertise disponible. Afin de rester dans un cadre opérationnel réaliste, l'hypothèse de linéarité est généralement considérée (Carillo et al., 2016 ; Araya-Alaman et al., 2019). Le cadre linéaire permet en effet de mettre en œuvre des approches de moindre carrés classiques avec un effort d'acquisition de données de référence (prise d'échantillon) limité.

Une fois le modèle reliant variable d'intérêt et variable auxiliaire étalonné, la seconde étape du model sampling consiste à effectuer une estimation de la grandeur d'intérêt pour chaque site de la parcelle en mobilisant le modèle et les données auxiliaires. La troisième étape fait la synthèse de l'ensemble de ces estimations et des mesures par une moyenne. Chaque site étant désormais associé à une valeur, une pondération uniforme est appliquée. La force de cette approche réside dans l'intégration de l'information sur la structure spatiale contenu dans la donnée auxiliaire lorsqu'elle est disponible avec une bonne résolution spatiale. Lorsque la représentativité de l'échantillon fait défaut, la prédiction de la valeur en chacun des sites à partir de la données auxiliaire limite l'erreur d'estimation. Si on reprend le cas extrême présenté dans la sous-partie précédente avec un échantillon présentant uniquement des valeurs fortes, l'utilisation de la donnée auxiliaire et d'un modèle intègre dans l'estimation l'existence de sites présentant des valeurs plus faibles que ceux où a été effectués la mesure. A noter cependant que le modèle n'aura pas été étalonné pour la prédiction des valeurs faible, ou plus généralement pour effectuer des estimations sur un intervalle de valeurs qui n'est pas représenté par l'échantillon. Autrement dit, ce modèle entrainé sur un échantillon de valeurs fortes risque de se révéler moins efficace pour l'estimation de valeur plus faibles. Pour ces raisons, la sélection des sites de mesure s'appuie généralement aussi sur une procédure d'échantillonnage orientée en fonction des données auxiliaires afin de prendre en compte la distribution des valeurs attributaires pour le bon étalonnage du modèle.

4.2 Description statistique des méthodes d'inférence

4.2.1 Notations

On numérote de 1 à k, l'ensemble des K sites de la parcelle. Pour le site i, on note X_i la valeur de la donnée auxiliaire et Y_i la valeur pour la variable que l'on cherche à estimer. Pour chaque site $i \in K$, la valeur de X_i est disponible. En revanche, la valeur de Y_i n'est connu que pour les n sites échantillonnés ($i \in N$). On fait l'hypothèse que pour chaque Y_i, une valeur de X_i est disponible et qu'il existe un modèle linéaire reliant Y_i à X_i de la forme :

$$Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i \qquad Eq.4.1$$
$$où\ les\ \varepsilon_i\ suivent\ une\ loie\ normale\ N(0, \sigma^2)\ et\ sont\ indépendants$$

On note $\bar{Y}$ la valeur moyenne des Y sur la parcelle (Eq. 4.2) :

$$\bar{Y} = \frac{1}{k} \times \sum_{i=1}^{k} Y_i \qquad Eq.4.2$$

Dans un premier temps l'objectif de l'estimation est de fournir l'expression des estimateurs de $\bar{Y}$. L'expression de ces estimateurs se base sur l'hypothèse que l'échantillon est tiré d'une population infinie. On analyse ensuite les propriétés de ces estimateurs en calculant leur variance et leur espérance.

4.2.2 Estimation par moyenne arithmétique (approche sampling classique ou target sampling)

4.2.2.1 *Estimateur*

On note $\widehat{\bar{Y}_1}$, l'estimateur de $\bar{Y}$ en inférant l'estimation par moyenne. $\widehat{\bar{Y}_1}$ correspond simplement à la moyenne des valeurs de Y_i pour les n sites échantillonnés ($i \in N$) :

$$\widehat{\bar{Y}_1} = \frac{1}{n} \times \sum_{i \in N} Y_i \qquad Eq.4.3$$

4.2.2.2 *Espérance*

On s'intéresse tout d'abord à l'espérance de cet estimateur, c'est-à-dire la valeur moyenne qu'il prendrait en répétant l'estimation plusieurs fois. Un estimateur non-biaisé présenterait une espérance égale à $\bar{Y}$. L'espérance de l'estimateur s'écrit :

$$E\left[\widehat{\bar{Y}_1}\right] = E\left[\frac{1}{n} \times \sum_{i \in N} Y_i\right] \qquad Eq.4.4$$

Bien que l'estimateur par la moyenne n'ait pas recours aux données auxiliaires, on se place toute de même sous l'hypothèse que ces données existent afin de caractériser ses propriétés.

En reprenant la relation linéaire entre les Y_i et les X_i présentée dans l'Eq. 4.1, il est possible de remplacer Y_i par son expression à partir de X_i:

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} (\beta_0 + \beta_1 X_i + \varepsilon_i)\right]$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} \beta_0\right] + \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} (\beta_1 X_i)\right] + \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} \varepsilon_i\right]$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 \overline{X_{i \in N}} + 0$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 (\overline{X_{i \in N}} + \bar{X} - \bar{X})$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 \bar{X} + \beta_1 (\overline{X_{i \in N}} - \bar{X})$$

$$\boldsymbol{E\left[\widehat{\bar{Y}_1}\right] = \bar{Y} + \beta_1 (\overline{X_{i \in N}} - \bar{X})} \qquad Eq.4.5$$

L'espérance de $\widehat{\bar{Y}_1}$ n'est pas égale à $\bar{Y}$. On observe un biais qui dépend de β_1, un des coefficients du modèle, et de $(\overline{X_{i\in N}} - \bar{X})$ qui représente l'écart entre la moyenne des sites de mesure et la moyenne de l'ensemble de la parcelle pour la donnée auxiliaire. Ce second terme qui est associé à la représentativité de l'échantillon est facilement interprétable. Plus l'échantillon est représentatif de la parcelle, plus $\overline{X_{i\in N}}$ converge vers $\bar{X}$ et plus le biais tend à s'annuler.

4.2.2.3 *Variance*

Après avoir vu que l'estimateur basé sur la moyenne était biaisé, on s'intéresse maintenant à sa variabilité. On exprime pour cela sa variance :

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i\in N} Y_i\right] \qquad Eq.4.6$$

En remplaçant Y_i par son expression à partir de X_i :

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i\in N} (\beta_0 + \beta_1 X_i + \varepsilon_i)\right]$$

β_0, β_1 et X_i sont fixés, on a donc :

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i\in N} \varepsilon_i\right]$$

Or $\mathrm{V}ar(aX) = a^2 Var(X)$:

$$Var\left[\widehat{\bar{Y}_1}\right] = \frac{1}{n^2} \times Var\left[\sum_{i\in N} \varepsilon_i\right]$$

Et on obtient finalement :

$$Var\left[\widehat{\bar{Y}_1}\right] = \frac{1}{n^2} \times n\sigma^2$$

$$\boldsymbol{Var\left[\widehat{\bar{Y}_1}\right] = \frac{\sigma^2}{n}} \qquad Eq.4.7$$

La variance de l'estimateur a donc une expression simple dans laquelle interviennent deux termes. Elle dépend tout d'abord de n, le nombre de sites de mesure contenus dans l'échantillon. Ce résultat est logique et cohérent avec les propriétés des estimateurs présentées dans la partie 1.1.3.3 montrant qu'un plus grand nombre de sites de mesure permet de réduire la variabilité de l'estimation. La variance dépend également de σ^2 , la variance des résidus dans la relation linéaire liant la variable d'intérêt à la donnée auxiliaire. Ce résultat est intéressant puisqu'il montre que dans le cas où les données sont parfaitement corrélées ($\sigma^2 = 0$), la variance de l'estimateur devient nulle et toute l'erreur d'estimation provient du biais et de la représentativité des sites choisis. A l'inverse, si la corrélation entre les deux variables est nulle, alors le coefficient β_1 est égal à 0 le biais de l'estimation devient nul. Toute l'erreur d'estimation sera alors exprimée sous forme de variabilité.

4.2.3 Estimation par modèle

4.2.3.1 *Estimateur*

De la même façon que pour l'estimation à partir d'une moyenne, on note $\widehat{\overline{Y_2}}$ l'estimateur de $\overline{Y}$ basée sur un modèle. Comme détaillé dans la première sous-partie du chapitre, $\widehat{\overline{Y_2}}$ est construit comme la moyenne des k valeurs prédites avec le modèle.

Pour un site i de la parcelle, on note $\widehat{\beta_0}$ et $\widehat{\beta_1}$ les estimateurs respectifs de β_0 et β_1. Dans le cadre du modèle linéaire simple, ces estimateurs sont bien connus :

$$\widehat{\beta_1} = \frac{\sum_{i \in N}(X_i - \overline{X_N})(Y_i - \overline{Y_N})}{\sum_{i \in N}(X_i - \overline{X_N})^2} \qquad Eq.4.8$$

$$\widehat{\beta_0} = \overline{Y_N} - \widehat{\beta_1} \times \overline{X_N} \qquad Eq.4.9$$

Où $\overline{X_N}$ représente la moyenne des X_i pour $i \in N$et $\overline{Y_N}$ représente la moyenne des Y_i pour $i \in N$.

De même, on note $\widehat{Y}_i$ l'estimation de Y_i par le modèle. $\widehat{\overline{Y_2}}$ peut alors être décrit à partir de l'ensemble des valeurs prédites par le modèle:

$$\widehat{\overline{Y_2}} = \frac{1}{k} \times \sum_{i=1}^{k} \widehat{Y}_i \qquad Eq.4.10$$

Qui peut être également réécrit :

$$\widehat{\overline{Y_2}} = \frac{1}{k} \times \sum_{i=1}^{k} \left(\widehat{\beta_0} + \widehat{\beta_1} X_i\right) \qquad Eq.4.11$$

$$\widehat{\overline{Y_2}} = \widehat{\beta_0} + \widehat{\beta_1}\overline{X} \qquad Eq.4.12$$

4.2.3.2 *Espérance*

Comme dans la partie précédente, il est possible de décrire l'espérance de l'estimateur basé sur le modèle :

$$\mathrm{E}\left[\widehat{\overline{Y_2}}\right] = E\left[\widehat{\beta_0} + \widehat{\beta_1}\overline{X}\right] \qquad Eq.4.13$$

$$\mathrm{E}\left[\widehat{\overline{Y_2}}\right] = E\left[\widehat{\beta_0}\right] + E\left[\widehat{\beta_1}\right]\overline{X}$$

Or, une des propriété du modèle linéaire donne que les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$ de β_0 et β_1 ne sont pas biaisées (Wasserman, 2004). Autrement dit, que $E\left[\widehat{\beta_0}\right] = \widehat{\beta_0}$ et $E\left[\widehat{\beta_1}\right] = \widehat{\beta_1}$. On peut alors écrire :

$$\mathrm{E}\left[\widehat{\overline{Y_2}}\right] = \widehat{\beta_0} + \widehat{\beta_1}\overline{X}$$

Et finalement, on obtient :

$$\mathbf{E}\left[\widehat{\overline{\boldsymbol{Y_2}}}\right] = \overline{\boldsymbol{Y}} \qquad Eq.4.14$$

Ce résultat est important car il montre que l'estimation basée sur une inférence par modèle et données auxiliaire n'est pas biaisée, contrairement au cas où l'inférence est réalisée en utilisant une moyenne

directe. Afin de pouvoir pleinement comparer les deux estimateurs, il reste néanmoins nécessaire d'exprimer la variance de $\widehat{\bar{Y}_2}$.

4.2.3.3 *Variance*

On cherche donc à exprimer la variance de $\widehat{\bar{Y}_2}$:

$$Var\left[\widehat{\bar{Y}_2}\right] = Var\left[\widehat{\beta_0} + \widehat{\beta_1}\bar{X}\right] \qquad Eq.\,4.15$$

On rappelle que la variance d'une somme se décompose de la manière suivante : $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$, on peut alors écrire :

$$Var\left[\widehat{\bar{Y}_2}\right] = Var\left[\widehat{\beta_0}\right] + Var\left[\widehat{\beta_1}\bar{X}\right] + 2Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\bar{X}\right] \qquad Eq.\,4.16$$

$$Var\left[\widehat{\bar{Y}_2}\right] = \sigma^2_{\widehat{\beta_0}} + \sigma^2_{\widehat{\beta_1}}\bar{X}^2 + 2\bar{X}Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right]$$

Or dans le cadre du modèle linéaire simple, les expressions des composantes $\sigma^2_{\widehat{\beta_0}}$, $\sigma^2_{\widehat{\beta_1}}$ et $Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right]$ sont connues et de la forme (Wasserman, 2004).:

$$\sigma^2_{\widehat{\beta_1}} = \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \qquad Eq.\,4.17$$

$$\sigma^2_{\widehat{\beta_0}} = \sigma^2 \times \left(\frac{1}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\right) \qquad Eq.\,4.18$$

$$Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right] = -\frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 \qquad Eq.\,4.19$$

On remplace donc ces composantes par leurs expressions (Eq. 4.17 à Eq. 4.19) dans l'expression de $\widehat{\bar{Y}_2}$ présentée dans l'équation 4.16 :

$$Var\left[\widehat{\bar{Y}_2}\right] = \sigma^2 \times \left(\frac{1}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\right) + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\bar{X}^2 - \frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 2\bar{X}$$

$$Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\sigma^2 + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\bar{X}^2 - \frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 2\bar{X}$$

On factorise ensuite par $\frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}$:

$$Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\left(\overline{X_N}^2 + \bar{X}^2 - 2\bar{X} \times \overline{X_N}\right)$$

$$\boldsymbol{Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2} \qquad Eq.\,4.20$$

On observe que la variance de l'estimateur $\widehat{\bar{Y}_2}$ est plus grande que celle de l'estimateur $\widehat{\bar{Y}_1}$. Un terme supplémentaire s'y ajoute, celui-ci est dépendant de :

- $(\overline{X_N} - \bar{X})^2$, l'écart quadratique de la moyenne des n sites choisis à la moyenne de la parcelle de la variable auxiliaire ;
- (ii) $\sum_{i\in N}(X_i - \overline{X_N})^2$, la somme des carrés des écarts (dispersion) des n sites autour de leur propre moyenne ;

- (iii) σ^2, la variance des résidus du modèle reliant variable d'intérêt et données auxiliaires.

Chacun de ces termes s'interprète assez facilement. Le premier (i) correspond à la représentativité des sites de mesure pour la donnée auxiliaire. De manière analogue au biais de $\widehat{\bar{Y}_1}$, l'intervention de ce terme dans l'erreur d'estimation est cohérent puisqu'il met en évidence la nécessité de sélectionner des sites de mesure dont la distribution des valeurs est centrée autour de la moyenne de la parcelle.

Le second (ii) s'explique par les propriétés du modèle linéaire. Il met en évidence la nécessité de sélectionner des sites de mesures dont les valeurs (données auxiliaire) présentent la dispersion la plus importante possible. La variabilité des paramètres estimés du modèle est plus importante lorsqu'il est étalonné à partir d'un ensemble de points similaires et peu dispersés que lorsque les points sont très dispersés autour de leur propre moyenne. Cette caractéristique est inhérente à la régression linéaire, elle est illustrée par la Figure 4.1 qui représente l'intervalle de confiance d'un modèle linéaire selon les points choisis pour son étalonnage.

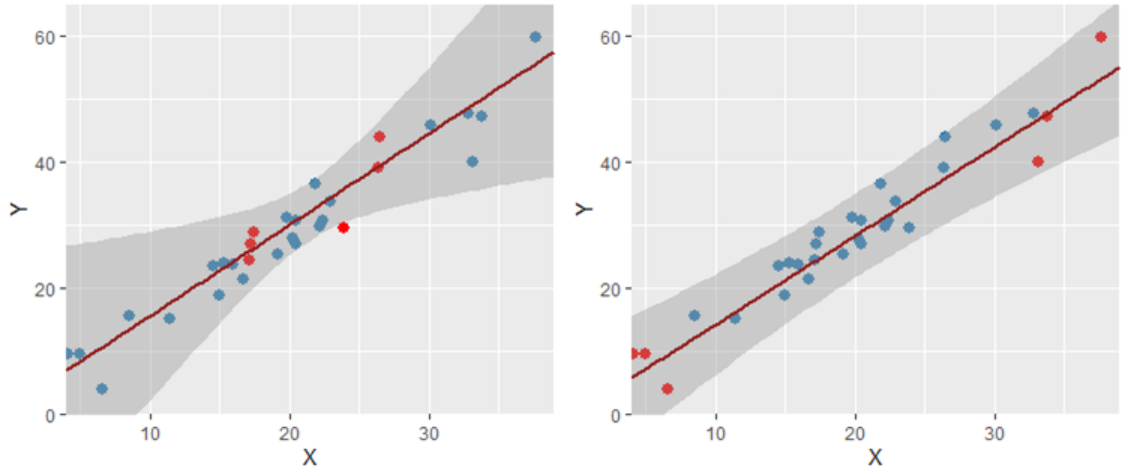

Figure 4.1 : Intervalle de confiance d'une droite de régression selon la distribution des données utilisées pour l'étalonnage. Sur ces deux figures, les six points rouges correspondent aux données utilisées pour l'étalonnage du modèle linéaire, la droite représente le modèle linéaire inféré et la zone grisée l'intervalle de confiance à 95%. L'utilisation de mesures plus dispersées par rapport à leur propre moyenne réduit la variabilité des paramètres du modèle estimé et son intervalle de confiance.

Enfin, (iii) la variance des résidus du modèle, qui est liée à la corrélation entre la variable d'intérêt et la variable auxiliaire, intervient également, l'utilisation d'un modèle sera logiquement pénalisée lorsque l'aptitude des données auxiliaires à prédire la variable d'intérêt décroît.

4.2.1 Comparaison et considérations pratiques

La Table 4.1 synthétise les résultats pour chaque estimateur :

Table 4.1 : Récapitulatif des propriétés des estimateurs basé sur une moyenne et sur un modèle.

Estimateur	Expression	Esperance	Variance
$\widehat{\bar{Y}_1}$: moyenne	$\frac{1}{n} \times \sum_{i \in N} Y_i$	$\bar{Y} + \beta_1(\overline{X_{i \in N}} - \bar{X})$	$\frac{\sigma^2}{n}$
$\widehat{\bar{Y}_2}$: modèle	$\widehat{\beta_0} + \widehat{\beta_1}\bar{X}$	$\bar{Y}$	$\frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i \in N}(X_i - \overline{X_N})^2} \times \sigma^2$

Sur la base de l'information apportée par la donnée auxiliaire, il apparait que les deux estimateurs présentent des propriétés différentes. L'estimateur $\widehat{\bar{Y}}_1$ basé sur la moyenne se révèle être biaisé alors que $\widehat{\bar{Y}}_2$ apparaît comme étant un estimateur non-biaisé. En contrepartie, la variabilité de $\widehat{\bar{Y}}_2$ est supérieure à la variabilité de $\widehat{\bar{Y}}_1$. Le choix de l'estimateur revient à un choix entre une erreur portée sur la variabilité ou sur le biais. La section suivante étudie comment se répercute ces propriétés sur l'erreur d'estimation afin de proposer quelques considérations pratiques pour leur utilisation.

4.1 Inférence et erreur d'estimation

Dans un contexte opérationnel, la question du choix entre une estimation basée sur un modèle et une estimation basée sur une moyenne doit prendre en compte l'expression de l'erreur d'estimation. En se basant sur l'information portée par la donnée auxiliaire, Il s'agit de déduire des informations sur l'erreur qu'il est possible d'attendre de l'estimation afin de retenir l'approche qui correspond à l'erreur d'estimation la plus faible.

4.1.1 Erreurs théoriques

L'expression de l'erreur théorique peut se faire à travers la formule de l'erreur quadratique moyenne (en anglais, MSE : mean square error). Elle représente l'espérance élevée au carré de l'écart théorique entre l'estimateur et la valeur estimée. Celle-ci s'exprime simplement à partir du biais et de la variance :

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] \qquad Eq.4.21$$

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

Après avoir exprimé la variance et l'espérance de nos estimateurs dans la partie précédente, on peut donc exprimer leurs erreurs quadratiques moyennes respectives. Pour $\widehat{\bar{Y}}_1$:

$$MSE\left(\widehat{\bar{Y}}_1\right) = Var\left[\widehat{\bar{Y}}_1\right] + Biais\left[\widehat{\bar{Y}}_1\right]^2 \qquad Eq.4.22$$

$$MSE\left(\widehat{\bar{Y}}_1\right) = \left[\frac{\sigma^2}{n}\right] + [\beta_1(\overline{X_{\iota\in N}} - \bar{X})]^2$$

$$MSE\left(\widehat{\bar{Y}}_1\right) = \frac{\sigma^2}{n} + {\beta_1}^2 \times (\overline{X_{\iota\in N}} - \bar{X})^2 \qquad Eq.4.23$$

Et pour $\widehat{\bar{Y}}_2$:

$$MSE\left(\widehat{\bar{Y}}_2\right) = Var\left[\widehat{\bar{Y}}_2\right] + Biais\left[\widehat{\bar{Y}}_2\right]^2$$

$$MSE\left(\widehat{\bar{Y}}_2\right) = \left[\frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2\right] + [0]^2$$

$$MSE\left(\widehat{\bar{Y}}_2\right) = \frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 \qquad Eq.4.24$$

L'expression de $MSE\left(\widehat{\bar{Y}}_2\right) - MSE\left(\widehat{\bar{Y}}_1\right)$ permet d'identifier quel estimateur parmi $\widehat{\bar{Y}}_2$ et $\widehat{\bar{Y}}_1$ permet d'obtenir la plus faible MSE. Si la différence est positive, $MSE\left(\widehat{\bar{Y}}_2\right)$ est plus petite que $MSE\left(\widehat{\bar{Y}}_1\right)$ et

l'estimation par le modèle linéaire doit être privilégiée. Si la différence est négative, il vaut mieux privilégier l'estimation par la moyenne.

$$MSE\left(\widehat{\bar{Y}_2}\right) - MSE\left(\widehat{\bar{Y}_1}\right) = \left[\frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2\right] - \left[\frac{\sigma^2}{n} + {\beta_1}^2 \times (\overline{X_{\iota\in N}} - \bar{X})^2\right] \quad Eq.4.25$$

$$MSE\left(\widehat{\bar{Y}_2}\right) - MSE\left(\widehat{\bar{Y}_1}\right) = \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 - {\beta_1}^2 \times (\overline{X_{\iota\in N}} - \bar{X})^2$$

$$MSE\left(\widehat{\bar{Y}_2}\right) - MSE\left(\widehat{\bar{Y}_1}\right) = (\overline{X_N} - \bar{X})^2 \times \left[\frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} - {\beta_1}^2\right] \quad Eq.4.26$$

On a donc $MSE\left(\widehat{\bar{Y}_2}\right) < MSE\left(\widehat{\bar{Y}_1}\right)$ si et seulement si :

$$\frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} < {\beta_1}^2 \quad Eq.4.27$$

Le choix de la méthode dépend donc explicitement de la pente de la régression linéaire faisant le lien entre donnée auxiliaire et variable d'intérêt ainsi que de la variance des résidus du modèle (associé à leur corrélation). L'autre facteur intervenant dans ce choix correspond à la dispersion des valeurs de données auxiliaires pour les sites de mesure. Une dispersion importante de ces valeurs permet en effet de mieux estimer les paramètre d'un modèle (Eq. 4.17, 4.18 et 4.19), ce qui favorise son utilisation pour l'inférence.

Cette connaissance permet de raisonner le choix de l'estimateur. La section suivante propose de la mettre en œuvre pour caractériser la meilleure approche possible à partir des propriétés de chaque parcelle.

4.1.2 Application numérique sur des données de rendement viticole

Chaque parcelle viticole est spécifique en terme de localisation, de taille de forme, de surface, d'âge, etc. Il en résulte que l'expression du rendement et de sa variabilité provient d'une interaction unique entre les facteurs de l'environnement (et leur variabilité spatiale et temporelle) en interaction avec les pratiques culturales (Taylor et al., 2005). Le rendement, sa variabilité spatiale et les liens potentiels avec des variables auxiliaires susceptibles d'en estimer la variabilité sont nécessairement spécifiques à chaque parcelle. Ce constat, amène nécessairement à raisonner l'échantillonnage (nombre et position des sites de mesure) de manière spécifique à chaque parcelle. Jusqu'à présent, ce document a cherché à adapter la position des sites de mesure à l'intérieur de la parcelle en fonction de la connaissance décrite par la variable auxiliaire disponible, le nombre d'échantillons étant laissé à la discrétion de l'opérateur en fonction de ses contraintes opérationnelles. Les développements proposés précédemment montrent qu'il est possible d'adapter la méthode d'inférence en fonction des propriétés de la parcelle et en particulier en fonction de la relation qui lie la variable d'intérêt (le rendement) avec la donnée auxiliaire.

L'objectif de cette section est de montrer comment la connaissance établie dans les sections précédentes peut être valorisée d'un point de vue opérationnel pour choisir la meilleure méthode d'estimation en considérant une alternative : *model sampling* ou *target sampling*. Cette section fait les hypothèses que (i) une variable auxiliaire pertinente est disponible avec une résolution spatiale suffisante pour renseigner sur la variabilité intra-parcellaire de la parcelle, que (ii) la variabilité est organisée spatialement et que (iii) quelques mesures de la variable d'intérêt (le rendement) ont été

effectuées afin de caractériser la qualité de la relation qui lie la variable d'intérêt à la variable auxiliaire pour chaque parcelle.

Table 4.2: Valeurs numériques des grandeurs influençant le choix de la méthode d'inférence sur données réelles.

Field	Area (ha)	Variety	β_1	σ	Correlation coefficient
P22	1.72	Syrah	570.4	995.8	0.13
P63	1.33	Syrah	4511.6	672.3	0.28
P65	0.69	Syrah	17140.0	491.2	0.86
P76	1.14	Carignan	6501.2	617.9	0.39
P77	1.24	Syrah	12932.3	926.0	0.48
P80	0.54	Syrah	12002.6	690.9	0.63
P82	1.15	Syrah	10126.3	547.8	0.47
P88	0.85	Syrah	-1184.9	852.0	-0.04
P104	0.81	Carignan	3758.0	1100.0	0.18

Le tableau 4.2 présente les valeurs numériques des grandeurs impliquées dans cette décision pour les données réelles sur lesquelles est basée une partie des résultats présentés au chapitre 3. Ces données ont été présentés en détail par Carillo et al. (2016).

De manière générale sur ces exemples, les rapports $\frac{\beta_1}{\sigma}$ prennent des valeurs élevées. Seules les parcelles 22 et 88 présentent des rapports $\frac{\beta_1}{\sigma}$ inférieures à 1. Il s'agit des parcelles présentant les indices de corrélation les plus faibles. A noter que dans les équations (Eq. 4.27), ces termes sont élevés au carré, les écarts sont alors d'autant plus importants. Il est difficile de caractériser la fraction $\frac{1}{\sum_{i\in N}(X_i-\overline{X_N})^2}$ sans réaliser d'échantillonnage. L'augmentation de la dispersion apparaît néanmoins favorable à l'utilisation d'un modèle. Ces applications numériques vont dans le sens des résultats obtenus par Carillo et al. (2015) sur ces mêmes données. Il semblerait que l'inférence par un modèle soit intéressante quand le coefficient de corrélation dépasse une certaine valeur de l'ordre de 0.2 mais ce résultat reste à confirmer.

4.1.1 Comparaison des méthodes d'inférence sur données

La Figure 4.2 compare les méthodes d'inférence pour quatre des parcelles présentées ci-dessus. Il s'agit des parcelles 22, 65, 76 et 82. Celles-ci sont choisies car elles représentent la diversité des 9 parcelles initiales en terme de corrélation entre la donnée auxiliaire et la variable d'intérêt. Pour chaque parcelle, 10000 échantillons de 8 sites de mesure sont réalisés selon la méthode des quantiles (pour plus de détails, voir chapitre 3). Pour chaque échantillon, l'inférence est réalisée deux fois en utilisant respectivement la moyenne et le modèle. Les échantillons sont ensuite placés sur la figure en fonction des erreurs obtenus avec le modèle (ordonnée) et obtenus avec la moyenne (abscisse). L'ordonnée correspond donc à l'erreur obtenue pour un *model sampling* et l'abscisse à l'erreur obtenue pour un *target sampling*. La droite rouge ($Y = X$) sépare les échantillons pour lesquels l'inférence par la moyenne est la meilleure (en dessous de la droite) de ceux pour lesquels l'inférence par la moyenne est la meilleure (au-dessus de la droite). Pour des raisons de lisibilité, les 10000 points sont représentés par une densité en deux dimensions. Les zones bleus clairs correspondent à une importante densité de points et les zones plus foncées à une densité plus faible. La figure fait

également apparaitre à titre indicatif la proportion d'échantillons qui tombent au-dessus et en dessous de la droite.

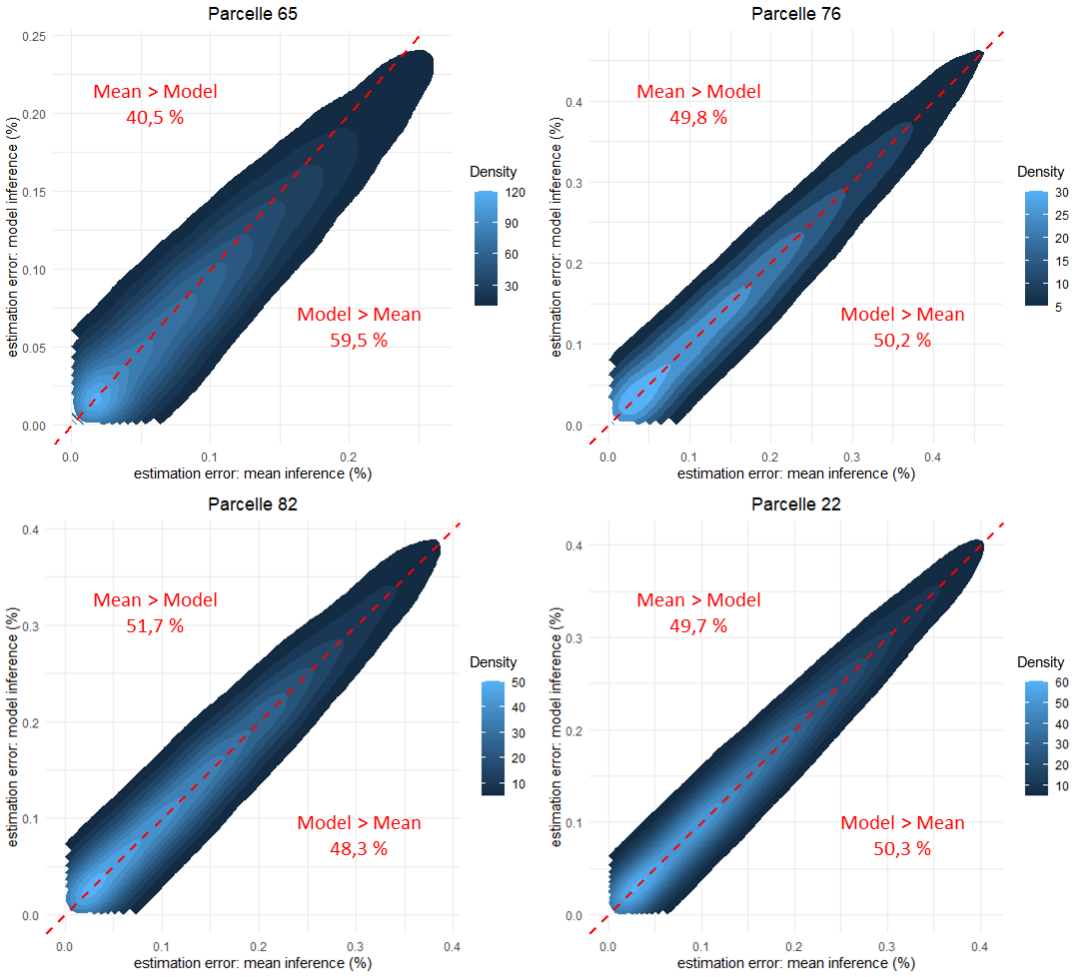

Figure 4.2 : Densité des erreurs d'estimation avec un modèle (ordonnées) et avec une moyenne (abscisses) pour 10 000 échantillon à N = 8 sites de mesure. La droite Y=X départage les échantillons pour lesquels le modèle est meilleur par rapport aux échantillons pour lesquels la moyenne est meilleure.

La différence entre inférence par un modèle et inférence par une moyenne apparait très clairement pour la parcelle 65, la densité de point est plus concentrée dans la partie inférieure droite du graphique. Environ 60% des échantillons donnent de meilleurs résultats avec un modèle. Cette parcelle se démarque des autres par la forte corrélation qui lie le rendement au NDVI ce qui est cohérent avec le résultat observé. Pour les autres parcelles, la différence entre les deux méthodes d'inférence est plus faible, en témoigne la quasi symétrie autour de la droite d'équation $Y = X$. Pour des corrélations intermédiaires, la parcelle 77 ne discrimine pas clairement les deux méthodes d'inférence. La parcelle 82 bien que présentant une corrélation de 0.47 entre rendement et NDVI, donne même des résultats légèrement meilleurs avec l'utilisation d'une moyenne simple. Dans le cas défavorable d'une faible corrélation comme celle de la parcelle 22, le recours au modèle n'est pas particulièrement pénalisé par rapport à la moyenne et reste privilégié. L'inférence par un modèle linéaire donne globalement de meilleurs résultats sur ces exemples et confirme les résultats obtenus par Carillo et al. (2015). Ces

résultats pratiques apparaissent cependant moins tranchés que les résultats théoriques de la partie précédente.

Plusieurs explications peuvent expliquer cet écart. Le *model sampling* est tout d'abord construit sur plusieurs hypothèses. Il est notamment supposé qu'il existe une corrélation entre les variables, que l'équation de régression est linéaire, que les résidus du modèle sont indépendants et sont issue d'une même loi normale. Ces hypothèses peuvent être remises en cause en pratique, la relation entre le NDVI, pris ici comme exemple, peut s'éloigner de la linéarité selon les parcelles. Le fait que la vigne soit taillée ou non est un exemple de facteur qui pourrait altérer le type de modèle liant ces deux variables.

De même les phénomènes de structure spatiale pourraient affecter l'indépendance des résidus du modèle. Pour s'affranchir de cette hypothèse importante, une solution pourrait être d'espacer suffisamment les sites de mesure afin de limiter leur autocorrélation spatiale. En utilisant la portée d'un variogramme (Annexe : A propos du semi-variogramme), il devient possible de considérer des observations indépendantes. Il serait également possible d'inclure ces problèmes d'indépendance en utilisant un modèle linéaire généralisé intégrant ces spécificités. Les limites de l'inférence par un modèle peuvent également s'expliquer par la fraction théorique $\frac{1}{\sum_{i\in N}(X_i-\overline{X_N})^2}$ qui reste difficile à quantifier. Enfin, le facteur $\frac{\sigma^2}{N}$ dans l'expression des erreurs quadratiques moyennes peut rendre négligeable les facteurs identifiés pour la décision sur la méthode d'inférence.

4.2 Conclusion

Les résultats théoriques et pratiques présentés dans ce chapitre rejoignent les observations plus ou moins empiriques faites par Carillo et al. (2015) et Araya-Alman et al. (2017 & 2019) et vont dans le sens de l'utilisation d'un modèle pour l'inférence lorsqu'une donnée auxiliaire est disponible. En pratique, l'inférence par la moyenne semble bénéfique lorsque l'on dispose de données auxiliaires fortement corrélées (coefficient de corrélation de Pearson > 0.5). Lorsque la corrélation est plus faible, le bénéfice apporté par ces méthodes d'inférence n'est pas visible mais l'estimation ne semble pas pénalisée, même dans les cas les plus défavorables. Remarquons que le nombre de parcelles sur lesquelles sont basées ces conclusions reste très limité. Des expérimentations sur des jeux de données plus importants restent nécessaires pour valider les résultats théoriques présentés ici et les hypothèses sur lesquelles ils s'appuient. Le développement des capteurs de rendement dans les machines à vendanger permettra à l'avenir de disposer de donnée systématique précieuse pour mieux étudier cette question dans des contextes techniques et pédoclimatiques plus diversifiés.

Le chapitre suivant prend le parti de l'utilisation d'un modèle linéaire pour l'inférence. Celui-ci prend la forme d'un article et propose un nouvel estimateur pour répondre à la question du choix des sites de mesure. Il présente un critère permettant de discriminer les sites de mesure d'une parcelle dans l'objectif de minimiser l'erreur d'estimation finale. L'approche mise en œuvre reprend les concepts statistiques exposés dans ce chapitre en se penchant à nouveau sur un estimateur basé sur un modèle linéaire. L'expression de celui-ci est adaptée à la sélection d'un échantillon parmi un ensemble de taille finie à travers un formalisme matriciel.

Chapitre 5 : A new criterion based on estimator variance for model sampling in precision agriculture

B. Oger[1-2], G. Le Moguedec[3], P. Vismara[2-4] and B. Tisseyre[1]
[1]*ITAP, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[2]*MISTEA, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[3]*AMAP, Univ. Montpellier, INRAE, CIRAD, CNRS, IRD France*
[4]*LIRMM, Univ. Montpellier, CNRS, France*

5.1 Introduction :

En production végétale, l'échantillonnage est une pratique indispensable en vue d'estimer les caractéristiques agronomiques d'une parcelle, que ce soit relativement aux cultures, au sol, etc. Les estimations résultant de l'échantillonnage sont des connaissances locales incontournables qui renseignent sur l'état du système de production et permettent aux agriculteurs d'ajuster leurs prises de décisions. Au cours du processus d'estimation, un échantillon d'observations est réalisé sur un nombre limité de sites de mesure. La grandeur d'intérêt est ensuite caractérisée à partir de cet échantillon d'observations par des techniques d'inférence basées sur un estimateur.

De nouvelles méthodes permettant l'acquisition rapide de données parcellaires se sont développées avec l'essor des technologies de l'information et de la communication en agriculture. Les méthodes de télédétection sont de plus en plus utilisées pour caractériser la vigueur des couverts végétaux grâce aux indices de végétation (Liaghat & Balasundram, 2010 ; Venkataratnam, 2001 ; Barnes & Baker, 2000) et d'autres capteurs permettent de récolter des données directement sur les parcelles (Rehman et al, 2014).

Malgré le développement de ces nouvelles méthodes de collecte de données, certaines prises de décision nécessitent encore le recours à un échantillonnage sur la parcelle, car certaines mesures restent inaccessibles par l'intermédiaire des capteurs aujourd'hui disponibles. Toutefois, ces nouvelles sources d'information sont précieuses car elles permettent, lorsqu'elles sont accessibles avec une haute résolution spatiale, de caractériser la variabilité et la structure spatiale des parcelles. Par ailleurs, même si la grandeur de mesure souhaitée n'est pas directement accessible, les observations issues des capteurs peuvent présenter un lien plus ou moins fort avec la grandeur d'intérêt ; c'est par exemple le cas entre le rendement et des observations de vigueur obtenues par télédétection. Dans ce contexte, des approches d'échantillonnage mobilisant ces nouvelles sources d'observation sont apparues. Les approches de *target sampling* et *stratified sampling* utilisent notamment ces observations afin de guider le choix des sites de mesure sur la parcelle (Miranda et al. 2018, Uribeetxebarria et al., 2019, Arnó et al., 2017.) D'autres méthodes proposent d'aller plus loin en mobilisant également ces observations au moment d'inférer l'estimation de la variable d'intérêt. L'estimateur est alors construit sur la base d'un modèle reliant la grandeur échantillonnée à l'information auxiliaire (observation) disponible. Ces approches dites de *model sampling* ont montré des résultats prometteurs en agriculture (Carillo et al, 2016 ; Araya-Alman et al. 2019 ; Murthy et al., 1995).

Les méthodes utilisées par les approches de *model* et *target sampling* pour guider le choix des sites de mesure restent cependant très empiriques. Dans ce contexte, cet article propose une réflexion plus approfondie sur le choix des sites de mesure. L'étude s'intéresse à l'estimation d'une espérance ou d'une valeur cumulée sur l'ensemble de la parcelle. Elle fait l'hypothèse que la grandeur d'intérêt est

plus ou moins fortement reliée linéairement à une donnée auxiliaire disponible. Les propriétés statistiques d'un estimateur basé sur un modèle linéaire sont alors décrites en s'appuyant sur un formalisme matriciel.

Ces travaux mettent en lumière comment le choix des sites de mesure affecte l'estimation finale. Ils montrent également comment les approches empiriques de *target sampling* permettent de réduire les erreurs d'estimation. Ces résultats théoriques sont mis en œuvre sur des données expérimentales pour l'estimation du rendement en viticulture et débouchent sur quelques considérations pour le choix des sites lors de l'utilisation d'une approche de *model sampling*.

5.2 Matériel et méthode :

5.2.1 Hypothèses et notations :

Dans cette partie, les notations en gras représentent les matrices et vecteurs.

Pour une parcelle donnée, l'objectif est d'identifier les n sites de mesure qui composeront l'échantillon pour l'estimation de la somme des valeurs prises par la grandeur d'intérêt sur la parcelle. Ces sites sont choisis parmi l'ensemble K des possibles sites de mesure, on parle aussi de sites de mesure potentiels. Pour chaque possible site de mesure ($i \in K$), numéroté de 1 à k, il existe une valeur pour la grandeur d'intérêt noté Y_i. Cette valeur est uniquement accessible pour les n sites échantillonnés ($i \in N$). Une deuxième variable, notée X_i, correspondant à une donnée auxiliaire est disponible pour chaque site de mesure potentiel ($i \in K$). L'hypothèse est faite qu'une relation linéaire unit la grandeur d'intérêt à la donnée auxiliaire. Il est possible d'écrire les valeurs de X_i sachant Y_i sous la forme :

$$\boldsymbol{Y_K}|\boldsymbol{X_K} = \beta_0 \boldsymbol{I_K} + \beta_1 \boldsymbol{X_K} + \boldsymbol{\varepsilon_K} \qquad Eq.5.1$$

Avec

$$\boldsymbol{\varepsilon_K} \sim N(\boldsymbol{0_K}, \sigma^2 \boldsymbol{I_K}) \qquad Eq.5.2$$

Où $\boldsymbol{Y_K}$ et $\boldsymbol{X_K}$ sont deux vecteurs de longueur k contenant respectivement les valeurs de la grandeur d'intérêt et de la donnée auxiliaire. Le vecteur $\boldsymbol{0_K}$ est un vecteur nul de longueur k et $\boldsymbol{I_K}$ la matrice identité de dimension k. Enfin β_0, β_1 et σ^2 représente les paramètres du modèle reliant $\boldsymbol{Y_K}$ à $\boldsymbol{X_K}$.

On suppose également que $\boldsymbol{Y_K}$ et $\boldsymbol{X_K}$ sont des vecteurs multinormaux. En particulier $\boldsymbol{X_K}$ suit une loi multinormale d'espérance $\boldsymbol{\mu_K}$ et de variance $\boldsymbol{V_K}$. Il est possible d'écrire l'espérance et la variance de la loi conditionnelle des observations de $\boldsymbol{Y_K}|\boldsymbol{X_K}$:

$$\mathbb{E}(\boldsymbol{Y_K}|\boldsymbol{X_K}) = \beta_0 \boldsymbol{I_K} + \beta_1 \boldsymbol{X_K} \qquad Eq.5.3$$

$$\mathbb{V}(\boldsymbol{Y_K}|\boldsymbol{X_K}) = \sigma^2 \boldsymbol{I_K} \qquad Eq.5.4$$

Par conséquent, le vecteur déconditionné $\boldsymbol{Y_K}$, suit lui-même une distribution multinormale d'espérance et de variance :

$$\mathbb{E}(\boldsymbol{Y_K}) = \mathbb{E}\big(\mathbb{E}(\boldsymbol{Y_K}|\boldsymbol{X_K})\big) = \mathbb{E}(\beta_0 \boldsymbol{I_K} + \beta_1 \boldsymbol{X_K})$$

$$\mathbb{E}(\boldsymbol{Y_K}) = \beta_0 \boldsymbol{I_K} + \beta_1 \boldsymbol{\mu_K} \qquad Eq.5.5$$

Et :

$$\mathbb{V}(\boldsymbol{Y_K}) = \mathbb{V}\big(\mathbb{E}(\boldsymbol{Y_K}|\boldsymbol{X_K})\big) + \mathbb{E}\big(\mathbb{V}(\boldsymbol{Y_K}|\boldsymbol{X_K})\big) = \mathbb{V}\big(\mathbb{E}(\beta_0 \boldsymbol{I_K} + \beta_1 \boldsymbol{X_K})\big) + \mathbb{E}\big(\mathbb{V}(\sigma^2 \boldsymbol{I_K})\big)$$

$$\mathbb{V}(\boldsymbol{Y_K}) = \beta_1 \boldsymbol{V_K} + \sigma^2 \boldsymbol{I_K} \qquad Eq.5.6$$

L'ensemble N, constitué des sites retenus dans l'échantillon, et l'ensemble R, constitué des sites non sélectionnés dans l'échantillon, forment une partition de l'ensemble K : $K = N \cup R$ et $N \cap R = \emptyset$. On peut ainsi décomposer les vecteurs $\boldsymbol{Y_K}$ et $\boldsymbol{X_K}$:

$$\boldsymbol{Y_K} = \begin{bmatrix} \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} \qquad et \qquad \boldsymbol{X_K} = \begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \end{bmatrix} \qquad Eq.5.7$$

On peut également décomposer les paramètres de la loi multinormale de $\boldsymbol{X_K}$:

$$\boldsymbol{\mu_K} = \begin{bmatrix} \boldsymbol{\mu_N} \\ \boldsymbol{\mu_R} \end{bmatrix} \qquad Eq.5.8$$

$$\boldsymbol{V_K} = \begin{bmatrix} \boldsymbol{V_{NN}} & \boldsymbol{V_{NR}} \\ \boldsymbol{V_{RN}} & \boldsymbol{V_{RR}} \end{bmatrix} \qquad Eq.5.9$$

5.2.2 Estimation des paramètres de la régression à partir de l'échantillon

La régression est construite à partir des observations qui sont choisies pour l'échantillonnage pour les variables X et Y, celles-ci sont contenues dans les vecteurs $\boldsymbol{Y_N}$ et $\boldsymbol{X_N}$. L'équation suivante reprend l'équation 5.1 pour l'ensemble N :

$$\boldsymbol{Y_N}|\boldsymbol{X}_N = \beta_0 + \beta_1 \boldsymbol{X_N} + \boldsymbol{\varepsilon_N} \quad avec \quad \boldsymbol{\varepsilon_N} \sim N(\boldsymbol{0_N}, \sigma^2 \boldsymbol{I_N})$$

$$\boldsymbol{Y_N}|\boldsymbol{X}_N = [\boldsymbol{1_N} \quad \boldsymbol{X_N}] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon_N}$$

$$\boldsymbol{Y_N}|\boldsymbol{X}_N = [\boldsymbol{1_N} \quad \boldsymbol{X_N}][\boldsymbol{\beta}] + \boldsymbol{\varepsilon_N} \qquad Eq.5.10$$

L'estimation des valeurs de $\boldsymbol{\beta}$ à partir de l'ensemble N par les moindres carrés conduit à l'estimateur suivant :

$$\widehat{\boldsymbol{\beta}} = ([\boldsymbol{1_N} \quad \boldsymbol{X_N}]^t [\boldsymbol{1_N} \quad \boldsymbol{X_N}])^{-1} . [\boldsymbol{1_N} \quad \boldsymbol{X_N}] \boldsymbol{Y_N} \qquad Eq.5.11$$

En notant $\overline{X_N} = \sum_{i \in N} \frac{X_i}{N}$, $\overline{Y_N} = \sum_{i \in N} \frac{Y_i}{N}$ et $\overline{X_N Y_N} = \sum_{i \in N} \frac{X_i \times Y_i}{N}$, il est possible de réécrire l'expression de $\widehat{\boldsymbol{\beta}}$ pour obtenir l'expression suivante :

$$\widehat{\boldsymbol{\beta}} = \frac{n}{\sum_{i \in N} (X_i - \overline{X_N})^2} \begin{bmatrix} \frac{1}{n} \times \sum_{i \in N} (X_i - \overline{X_N})^2 + \overline{X_N}^2 & -\overline{X_N} \\ -\overline{X_N} & 1 \end{bmatrix} \begin{bmatrix} \overline{Y_N} \\ \overline{X_N Y_N} \end{bmatrix} \qquad Eq.5.12$$

On peut alors établir que le vecteur $\widehat{\boldsymbol{\beta}}$ suit une loi binormale d'espérance et de variance :

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad Eq.5.13$$

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\sum_{i \in N} (X_i - \overline{X_N})^2} \begin{bmatrix} \frac{1}{n} \times \sum_{i \in N} (X_i - \overline{X_N})^2 + \overline{X_N}^2 & -\overline{X_N} \\ -\overline{X_N} & 1 \end{bmatrix} \qquad Eq.5.14$$

Enfin on s'intéresse à l'estimateur de σ^2, le dernier paramètre du modèle linéaire. Cette estimation est réalisée avec $n - 2$ degrés de liberté :

$$\widehat{\sigma^2} = \frac{\left(\boldsymbol{Y_N} - [\boldsymbol{1_N} \quad \boldsymbol{X_N}] . \widehat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{Y_N} - [\boldsymbol{1_N} \quad \boldsymbol{X_N}] . \widehat{\boldsymbol{\beta}}\right)}{n - 2} \qquad Eq.5.15$$

5.2.3 Loi conditionnelle

On s'intéresse dans cette partie au vecteur conjoint $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$ que l'on souhaite décomposer en utilisant les notations présentées dans l'Eq. 5.7. On obtient alors :

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \\ \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} \qquad Eq.5.16$$

A partir des équations 5.5 et 5.8, il est possible de décrire l'espérance de la loi conjointe :

$$\mathbb{E}\begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \\ \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu_N} \\ \boldsymbol{\mu_R} \\ \beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{\mu_N} \\ \beta_0 \boldsymbol{I_R} + \beta_1 \boldsymbol{\mu_R} \end{bmatrix} \qquad Eq.5.17$$

De même, à partir des équations 5.6 et 5.9, il est possible de décrire la variance de la loi conjointe :

$$\mathbb{V}\begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \\ \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V_N} & \boldsymbol{V_{NR}} & \beta_1 \boldsymbol{V_N} & \beta_1 \boldsymbol{V_{NR}} \\ \boldsymbol{V_{RN}} & \boldsymbol{V_R} & \beta_1 \boldsymbol{V_{RN}} & \beta_1 \boldsymbol{V_R} \\ \beta_1 \boldsymbol{V_N} & \beta_1 \boldsymbol{V_{NR}} & \beta_1{}^2 \boldsymbol{V_N} + \sigma^2 \boldsymbol{I_N} & \beta_1{}^2 \boldsymbol{V_{NR}} \\ \beta_1 \boldsymbol{V_{RN}} & \beta_1 \boldsymbol{V_R} & \beta_1{}^2 \boldsymbol{V_{RN}} & \beta_1{}^2 \boldsymbol{V_R} + \sigma^2 \boldsymbol{I_R} \end{bmatrix} \qquad Eq.5.18$$

En distinguant les valeurs de $\boldsymbol{X_N}$, $\boldsymbol{X_R}$ et $\boldsymbol{Y_N}$ qui sont connues (1) de celles de $\boldsymbol{Y_R}$ qui sont inconnues (2), les notations $\boldsymbol{m_1}$, $\boldsymbol{m_2}$, $\boldsymbol{\Sigma_{11}}$, $\boldsymbol{\Sigma_{12}}$, $\boldsymbol{\Sigma_{21}}$ et $\boldsymbol{\Sigma_{22}}$ sont introduites :

$$\mathbb{E}\begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \\ \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{m_1} \\ \boldsymbol{m_2} \end{bmatrix}$$

Avec :

$$\boldsymbol{m_1} = \begin{bmatrix} \boldsymbol{\mu_N} \\ \boldsymbol{\mu_R} \\ \beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{\mu_N} \end{bmatrix} \quad et \quad \boldsymbol{m_2} = [\beta_0 \boldsymbol{I_R} + \beta_1 \boldsymbol{\mu_R}] \qquad Eq.5.19$$

Et :

$$\mathbb{V}\begin{bmatrix} \boldsymbol{X_N} \\ \boldsymbol{X_R} \\ \boldsymbol{Y_N} \\ \boldsymbol{Y_R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}$$

Avec :

$$\boldsymbol{\Sigma_{11}} = \begin{bmatrix} \boldsymbol{V_N} & \boldsymbol{V_{NR}} & \beta_1 \boldsymbol{V_N} \\ \boldsymbol{V_{RN}} & \boldsymbol{V_R} & \beta_1 \boldsymbol{V_{RN}} \\ \beta_1 \boldsymbol{V_N} & \beta_1 \boldsymbol{V_{NR}} & \beta_1{}^2 \boldsymbol{V_N} + \sigma^2 \boldsymbol{I_N} \end{bmatrix} \quad et \quad \boldsymbol{\Sigma_{12}} = \begin{bmatrix} \beta_1 \boldsymbol{V_{NR}} \\ \beta_1 \boldsymbol{V_R} \\ \beta_1{}^2 \boldsymbol{V_{NR}} \end{bmatrix} \qquad Eq.5.20$$

$$\boldsymbol{\Sigma_{21}} = [\beta_1 \boldsymbol{V_{RN}} \quad \beta_1 \boldsymbol{V_R} \quad \beta_1{}^2 \boldsymbol{V_{RN}}] \quad et \quad \boldsymbol{\Sigma_{22}} = [\beta_1{}^2 \boldsymbol{V_R} + \sigma^2 \boldsymbol{I_R}]$$

5.2.4 Formalisation d'un estimateur

L'objectif est d'estimer T, la somme des valeurs unitaires de la variable d'intérêt Y_i sur la parcelle. En séparant les valeurs pour lesquelles une observation est disponible (S), des valeurs non-observées (R) comme défini dans l'équation 5.7 :

$$T = \sum_{i \in K} Y_i \qquad Eq.5.21$$

$$T = \mathbf{1}_{\boldsymbol{K}}^{t} \boldsymbol{Y_K}$$

$$T = \mathbf{1}_{\boldsymbol{N}}^{t} \boldsymbol{Y_N} + \mathbf{1}_{\boldsymbol{R}}^{t} \boldsymbol{Y_R} \qquad Eq.5.22$$

On note $\hat{T}$ l'estimateur de T. Les valeurs du vecteur $\boldsymbol{Y_N}$, qui correspondent aux valeurs mesurées de la grandeur d'intérêt, étant connues, le problème revient à estimer les valeurs de $\boldsymbol{Y_R}$. $\mathbf{1}_{\boldsymbol{R}}^{t}\, \mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_N}, \boldsymbol{X_N}, \boldsymbol{X_N})$ est choisi comme estimateur de $\mathbf{1}_{\boldsymbol{R}}^{t} \boldsymbol{Y_R}$ car celui-ci permet de minimiser le risque quadratique.

$$\hat{T} = \mathbf{1}_{\boldsymbol{N}}^{t} \boldsymbol{Y_N} + \mathbf{1}_{\boldsymbol{R}}^{t}\, \mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_N}, \boldsymbol{X_N}, \boldsymbol{X_N}) \qquad Eq.5.23$$

En décomposant $\mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_N}, \boldsymbol{X_N}, \boldsymbol{X_N})$ à l'aide des notations introduites dans la sous-section précédente :

$$\hat{T} = \mathbf{1}_{\boldsymbol{N}}^{t} \boldsymbol{Y_N} + \mathbf{1}_{\boldsymbol{R}}^{t} \left(\boldsymbol{m_2} + \boldsymbol{\Sigma_{21}}.\boldsymbol{\Sigma_{11}^{-1}}.\begin{bmatrix} \boldsymbol{X_N} - \boldsymbol{\mu_N} \\ \boldsymbol{X_R} - \boldsymbol{\mu_R} \\ \boldsymbol{Y_N} - \beta_0 \boldsymbol{I_N} - \beta_1 \boldsymbol{\mu_N} \end{bmatrix} \right) \qquad Eq.5.24$$

En développant, il est possible de réécrire l'expression de $\hat{T}$ sous la forme :

$$\hat{T} = n\overline{Y_N} + (k-n)\beta_0 + \beta_1 \mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}} \boldsymbol{X_R} \qquad Eq.5.25$$

Cette formulation fait intervenir les coefficient β_0 et β_1. En pratique, ceux-ci ne sont pas connus et remplacés par leurs estimateurs respectifs :

$$\hat{T} = n\overline{Y_N} + (k-n)\widehat{\beta_0} + \widehat{\beta_1} \mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}} \boldsymbol{X_R} \qquad Eq.5.26$$

5.2.5 Propriétés de l'estimateur

Pour cet estimateur, on s'intéresse à l'expression de son espérance :

$$\mathbb{E}(\hat{T}) = \mathbb{E}\left(n\overline{Y_N} + (k-n)\widehat{\beta_0} + \widehat{\beta_1} \mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}} \boldsymbol{X_R}\right) = n\overline{Y_N} + (k-n)\beta_0 + \beta_1 \mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}} \boldsymbol{X_R} \qquad Eq.5.26$$

Il s'agit ici d'un estimateur sans biais et de variance :

$$\mathbb{V}(\hat{T}) = \mathbb{V}\left(n\overline{Y_N} + (k-n)\widehat{\beta_0} + \widehat{\beta_1} \mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}} \boldsymbol{X_R}\right)$$

$$\mathbb{V}(\hat{T}) = \begin{bmatrix} (k-n) & \sum_{i \in R} X_i \end{bmatrix}.\mathbb{V}(\hat{\beta}).\begin{bmatrix} (k-n) \\ \sum_{i \in R} X_i \end{bmatrix} \qquad Eq.5.27$$

Après quelques transformations, cette variance peut s'écrire :

$$\mathbb{V}(\hat{T}) = (k-n)^2 \times \left(\frac{1}{n} + \frac{(\overline{X_R} - \overline{X_N})^2}{\sum_{i \in N}(X_i - \overline{X_N})^2}\right) \times \sigma^2 \qquad Eq.5.28$$

La variance de l'estimateur dépend donc de :

- k, le nombre de sites de la parcelle ;
- n, le nombre de sites de mesure ;
- σ^2, la variance de la résiduelle du modèle ;
- $X_{i\epsilon N}$, les valeurs prises individuellement par les sites de mesure pour la donnée auxiliaire ;
- $\overline{X_N}$, la valeur moyenne des sites de mesure pour la donnée auxiliaire ;
- $\overline{X_R}$, la valeur moyenne des sites de mesure pour la donnée auxiliaire.

Cette variance tend naturellement vers 0 lorsque n tend vers k.

Le raisonnement tenu a conduit à la construction d'un estimateur de l'espérance de T. S'il s'agit de faire une prévision, de la même manière que pour une prévision en régression linéaire, il faut tenir compte de la variance individuelle ε_i pour chacun des Y_i non observés ($i \in R$). Si on appelle $\tilde{T}$ cette prévision, elle a pour variance :

$$\mathbb{V}(\tilde{T}) = \mathbb{V}(\hat{T}) + (k-n).\mathbb{V}(\mathbf{1}_{\boldsymbol{R}}^{t}\varepsilon_R) = \mathbb{V}(\hat{T}) + (k-n) \times \sigma^2$$

$$\mathbb{V}(\tilde{T}) = (k-n)^2 \times (\frac{1}{n} + \frac{1}{k-n} + \frac{(\overline{X_R} - \overline{X_N})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}) \times \sigma^2 \qquad Eq.5.29$$

$\tilde{T}$ est une prévision de T, la somme Y_i. Le raisonnement précédent est applicable à $\frac{\tilde{T}}{k}$ qui est un estimateur de l'espérance des $Y_{i\in K}$. La variance de $\frac{\tilde{T}}{k}$ est de la forme $\frac{\mathbb{V}(\tilde{T})}{k^2}$ et présente des propriétés similaires.

5.2.6 Critère de variance pour le choix des sites de mesure

Par la suite, on désignera par « *critère de variance* », la part de variance de l'estimateur (Eq. 5.27) ou de la prévision (Eq. 5.29) associée aux valeurs de donnée auxiliaire des sites de mesure :

$$\frac{(\overline{X_R} - \overline{X_N})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \qquad Eq.5.30$$

Pour un échantillon de taille n donné, le *critère de variance* définit la part de variance qui dépend du choix des sites de mesure. Au numérateur apparaît $(\overline{X_R} - \overline{X_N})^2$: la différence entre la moyenne de l'échantillon et celle de l'ensemble de la population. Cette différence traduit la représentativité des sites de mesure. Le dénominateur $\sum_{i\in N}(X_i - \overline{X_N})^2$ correspond à la somme des écarts au carré entre les sites de mesure et leur propre moyenne, il représente la dispersion des valeurs de l'échantillon selon la donnée auxiliaire.

5.2.7 Obtention d'un échantillon

Deux méthodes sont mises en œuvre pour la sélection des n sites de mesure.

La première est un *random sampling* (Wulfsohn, 2010). Dans cette approche, l'ensemble N des sites échantillonnés est tiré parmi l'ensemble K des sites disponibles par un tirage sans remise.

La deuxième méthode repose sur le principe du *target sampling*. Celle-ci propose de partitionner l'ensemble K en n sous-ensembles selon les valeurs pour la donnée auxiliaire (variable X). Un unique site de mesure est ensuite sélectionné aléatoirement dans chacun des n sous-ensembles (Carillo et al., 2016 ; Oger et al. 2019). Deux partitionnements sont également testés :

- La méthode des quantiles où l'ensemble K est découpé par les percentiles $\frac{k}{n}$, $\frac{2k}{n}$,…, $\frac{(n-1)\times k}{n}$.

- L'algorithme des k-means.

5.2.8 Mesure de la qualité de l'estimation

La qualité de l'erreur d'estimation est mesurée par l'erreur d'estimation. Celle-ci est définie comme l'écart relatif absolu entre la valeur prise par l'estimateur et la grandeur estimée. Sa valeur est exprimée en pourcentage par rapport à la grandeur estimée :

$$Erreur\ (\%) = \frac{\left|\tilde{T} - T\right|}{T} \qquad Eq.5.31$$

L'erreur quadratique moyenne (RMSE : *root mean square error*) est une mesure de la qualité d'une estimation au regard d'un nombre important d'estimation. En définissant $Samples$ comme un ensemble d'échantillons, elle est calculée comme suit :

$$RMSE = \sqrt{\sum_{i \in Samples} \frac{\left(\tilde{T}_{Samples} - T\right)^2}{Cardinal(Samples)}} \qquad Eq.5.32$$

En théorie, la RMSE est également définie comme la somme du biais au carré et de la variance (Wasserman, 2004) :

$$RMSE = \sqrt{\left(\mathbb{E}\left(\tilde{T}\right) - T\right)^2 + \mathbb{V}\left(\tilde{T}\right)} \qquad Eq.5.33$$

5.2.9 Données

Les parcelles réelles utilisées pour tester la méthode appartiennent à l'INRAE Pech-Rouge (Narbonne, France). L'expérimentation et les données qui en découlent sont détaillées par Carrillo et al. (2016). Celles-ci sont brièvement résumées ci-après. La donnée auxiliaire correspond à un indice de végétation : le NDVI. Neuf parcelles sont représentées dans ce jeu de données. Toutes sont non irriguées et exposées au climat méditerranéen avec des précipitations se produisant au printemps et un été chaud et sec. Les caractéristiques de chaque parcelle sont présentées dans le tableau 1.

Table 5.1: Caractéristiques des parcelles expérimentales.

Field	Area (ha)	Variety	Total Number of Sites (k)	Pearson correlation coefficient (NDVI/yield)
P22	1.72	Syrah	45	0.13
P63	1.33	Syrah	42	0.28
P65	0.69	Syrah	33	0.86
P76	1.14	Carignan	37	0.39
P77	1.24	Syrah	19	0.48
P80	0.54	Syrah	40	0.63
P82	1.15	Syrah	53	0.47
P88	0.85	Syrah	21	-0.04
P104	0.81	Carignan	23	0.18

Les valeurs NDVI sont dérivées d'une image multispectrale de résolution de 1 pixel = 1 m² prise le 31 août 2008 par Avion Jaune (Montpellier, Hérault, France). Les régions spectrales saisies dans les images

étaient les suivantes : bleu (445-520 nm), vert (510-600 nm), rouge (632-695 nm) et proche infrarouge (757-853 nm). À partir de cette image, la méthode d'agrégation décrite dans (Acevedo-Opazo et al. 2008) a été utilisée pour obtenir des pixels d'image de 9m², réduisant l'effet de la discontinuité du couvert végétal et du sol nu sur les valeurs mesurées. Le NDVI a finalement été calculé à partir des images traitées selon Rouse et al. (1973). Le désherbage mécanique ou chimique a été effectué sur l'espacement entre les rangs ; par conséquent, l'enherbement n'a que peu affecté les valeurs de NDVI.

Les mesures de rendement local sur les parcelles ont été réalisées régulièrement selon une grille d'échantillonnage de 15x15 m. À chaque nœud de la grille, le rendement a été mesuré sur 5 pieds de vigne consécutifs sur le rang et le rendement moyen a été affecté à la localisation correspondant au pied de vigne central. La base de données finale est constituée d'un ensemble de 313 sites répartis sur les 9 parcelles différentes. Pour chaque site, une valeur de rendement et de NDVI est disponible.

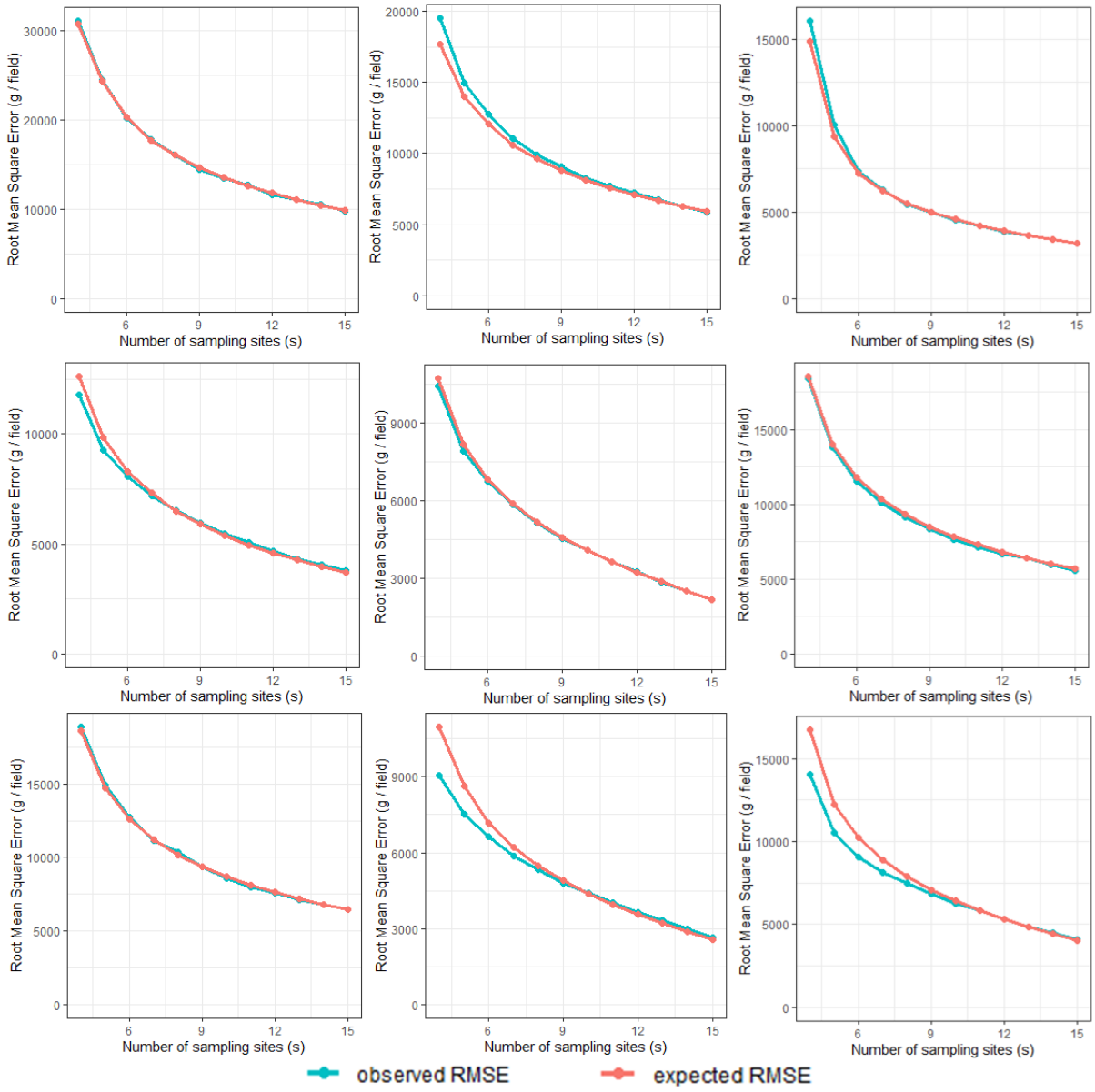

Figure 5.1 : RMSE observées (en bleu) et théoriques (en rouge) ; moyennes effectuées pour 9 parcelles de vignes (de gauche à droite et de haut en bas : P22, P63, P65, P76, P77, P80, P82, P88, P104) avec un nombre variable de sites d'échantillonnage. Les RMSE observées résultent d'un échantillonnage aléatoire et du calcul de l'écart entre le rendement des parcelles et la moyenne donnée par l'échantillon. Les RMSE théoriques sont déduites de l'équation XX à partir des valeurs de NDVI des sites échantillonnés.

5.3 Résultats

La Figure 5.1 compare les RMSE théoriques et observées des estimations de rendement en fonction du nombre de sites de mesure (s) pour chacune des neuf parcelles considérées (Table 5.1). Le nombre de sites de mesure varie de 4 à 15 pour chaque parcelle. La courbe bleue correspond à la RMSE observée. Chaque point représente la RMSE des estimations de 10000 tirages. La courbe rouge donne la moyenne des RMSE théorique calculée avec l'équation de la variance théorique de la prévision telle que proposée (eq. 5.29)

La Figure 5.1 montre que Les valeurs de RMSE moyennes observées (bleu) et théoriques (rouge) sont très similaires. En moyenne, l'écart entre les valeurs prédites et observées est de 2.6%, toutes parcelles confondues. Seules quelques parcelles (P63, P88, P104) avec un coefficient de corrélation faible présentent un décalage allant jusqu'à 10% entre valeurs observées et prédites pour les valeurs de n les plus faibles (inférieures à 9). Un modèle linéaire reliant la RMSE attendue à la RMSE construite à partir de l'ensemble des points présenté dans les neufs graphiques de la Figure 5.1 correspond à un coefficient de corrélation multiple de $R^2 = 0.9897$ (résultat non montré). Cette adéquation entre résultats théoriques et observés apporte une première validation de la formule de la variance théorique et du critère de variance ainsi que des hypothèses sur lesquelles leurs expressions sont basées (absence de biais).

Pour toutes les parcelles, la RMSE diminue logiquement lorsque le nombre de sites de mesure augmente, cette diminution est dégressive. Ce phénomène s'explique logiquement, car le gain apporté par un site de mesure additionnel diminue avec leur nombre. Par exemple, passer de 4 à 5 sites de mesure a plus d'impact sur la qualité de l'estimation du rendement que de passer de 14 à 15.

La Figure 5.2 montre le résultat de 9 000 *random sampling* sur les données disponibles, toutes parcelles confondues (1 000 random sampling par parcelle). Chaque estimation de rendement est le résultat d'une moyenne calculée à partir de 8 sites de mesure choisis aléatoirement. Les résultats d'erreur d'estimation de chacun des 9000 échantillons sont représentés par les pixels de la Figure 5.2 en fonction de la valeur du critère de variance constaté ; chaque pixel est associé à un intervalle de valeurs pour le critère de variance et pour l'erreur d'estimation, sa couleur dépend du nombre d'estimations tombant entre ces deux intervalles.

Les valeurs du critère de variance prises pour ces *random samplings* se concentrent autour de la médiane (0.012) avec 45% des valeurs comprises entre 10^{-2} et 10^{-1} et une dispersion allant de 10^{-10} à 10^{1}. La courbe en rouge présente une régression locale (Jacoby 2000) de l'évolution de l'erreur d'estimation moyenne en fonction du critère de variance observé. L'intervalle de confiance à 95 % de la courbe est représentée par une enveloppe grise. Pour des valeurs faibles de critère de variance, l'erreur d'estimation correspond à un plateau avec des valeurs d'erreur proches de 15%, puis l'erreur d'estimation commence à augmenter lorsque le critère de variance dépasse 10^{-1}.

Sur une échelle logarithmique, on observe une augmentation de l'erreur d'estimation en fonction du critère de variance, celle-ci est d'abord lente puis s'accélère. Cette observation est cohérente avec l'équation théorique de la variance de l'estimation (Eq 5.28). En effet, dans cette formule, le critère de variance s'ajoute aux termes $\frac{1}{n}$ et $\frac{1}{k-n}$. Pour les valeurs les plus faibles (inférieures à 10^{-2}), la valeur du critère de variance représente une part négligeable de la somme de ces trois termes. Ses variations n'impactent alors que très peu la variance de l'estimation. Quand le critère de variance atteint des valeurs de l'ordre de $\frac{1}{n}$, ses variations affectent significativement la variance de l'estimation. Une

augmentation du critère de variance se répercute alors sur la variance de l'estimation ce qui a pour conséquence d'augmenter l'erreur d'estimation.

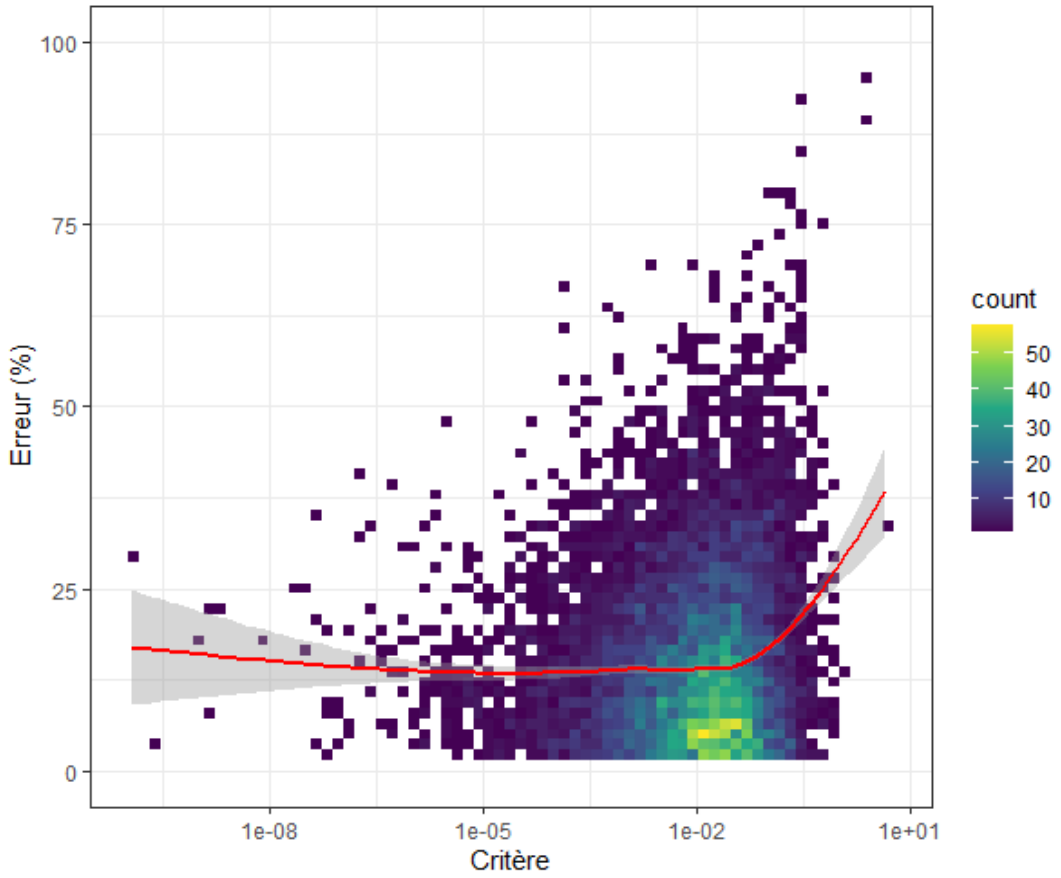

Figure 5.2 : Relation entre critère de variance et erreur d'estimation. L'erreur d'estimation moyenne (en rouge) augmente pour lorsque les estimations sont effectuées avec un échantillon qui présente un critère de variance élevé.

Par un procédé analogue à celui de la Figure 5.2, les trois graphiques de la Figure 5.3 montrent les résultats obtenus sur un échantillon de trois parcelles. Les illustrations 3A, 3B et 3C correspondent respectivement aux parcelles P22, P82 et P65 choisies pour leurs différences de corrélation entre l'indice de NDVI et le rendement (respectivement 0.13 ,0.47 et 0.86) (Tableau 1).

Les résultats obtenus pour chacune des trois parcelles sont très similaires à ceux présentés à la Figure 5.2: une forte proportion d'échantillons avec une valeur de critère de variance comprise entre 10^{-2} et 10^{-1} et une augmentation de l'erreur d'estimation pour les échantillons dont le critère de variance dépasse 10^{-2}. Les erreurs d'estimation sont plus faibles pour la parcelle P65 (2C) qui présente le coefficient de corrélation entre variable d'intérêt et variable auxiliaire le plus élevé. L'erreur minimale de la parcelle P65 (2C) est en effet de 10 % contre 15 à 20% pour les deux autres parcelles. L'effet de la corrélation entre données d'intérêt et données auxiliaires s'illustre également par le nombre d'échantillons pour lesquels une erreur d'estimation supérieure à 30% a été observée. Ce nombre est logiquement plus élevé pour la parcelle P22 (Figure 5.3.A) et la parcelle P82 (Figure 5.3.B) que pour la parcelle P65 (Figure 5.3.C). Ce résultat valide l'effet de la corrélation entre variable d'intérêt et variable auxiliaire ; la corrélation au sens de Pearson est une mesure de la qualité de la relation linéaire liant les variables. Ce dernier est relié à la variance de la résiduelle du modèle, σ^2, qui intervient directement dans l'expression théorique de la variance de l'estimateur. La distinction entre les parcelles P22 (Figure 5.3.A) et P83 (Figure 5.3.B) avec des corrélations respectives de 0.13 et 0.47 n'apparaît cependant pas clairement.

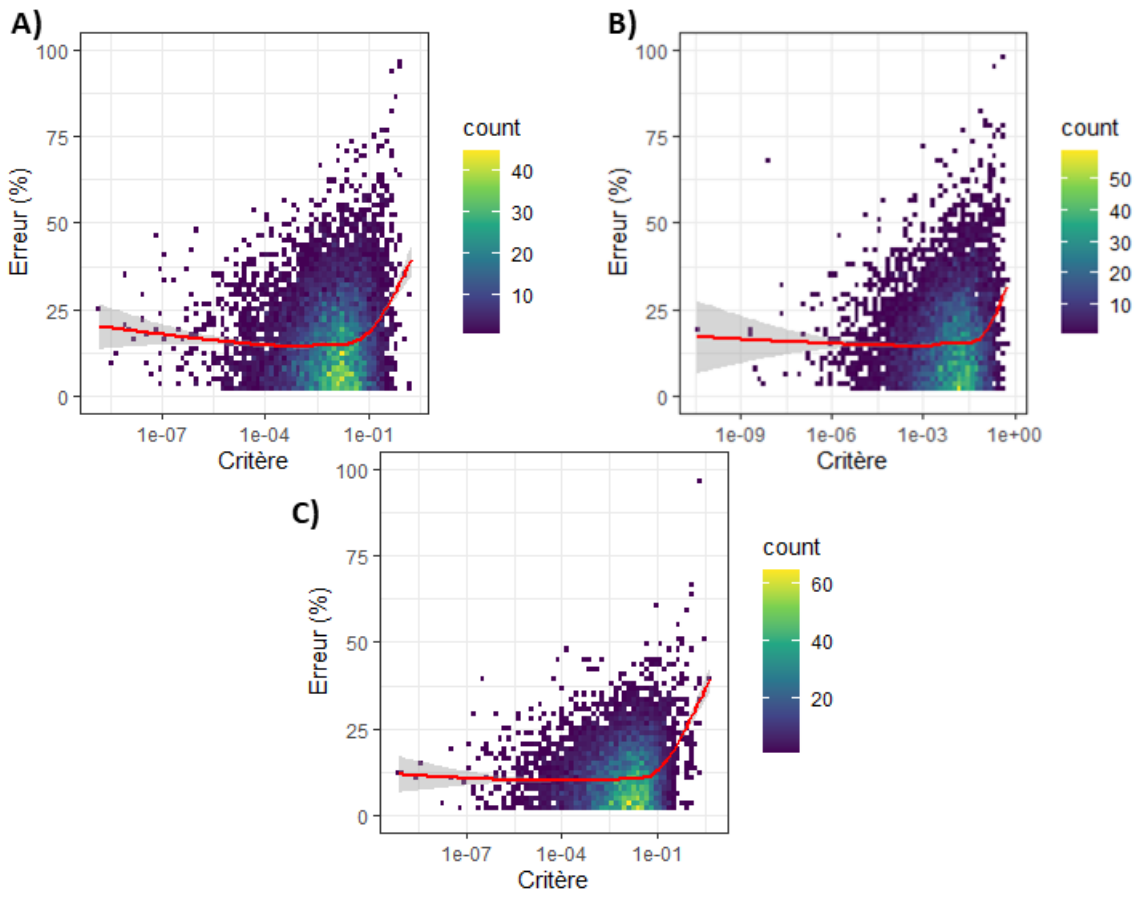

Figure 5.3: Evolution de l''erreur d'estimation moyenne en fonction du critère de variance pour trois parcelles présentant des corrélations différentes entre la donnée auxiliaire et la variable d'intérêt : la parcelle avec une corrélation élevée (3C) correspond à des erreurs d'estimation plus faibles que les parcelles présentant des corrélations moyennes (3B) à faible (3A).

La Figure 5.4 met en évidence l'intérêt des approches de target sampling pour réduire l'erreur d'estimation. La Figure 5.4.A présente les résultats obtenus par la méthode des quantiles (4A) tandis que la figure (4B) présente les résultats obtenus par la méthode des kmeans. Ces deux résultats sont obtenus avec 9 000 *target sampling* toutes parcelles confondues (1 000 par parcelle) avec des échantillons de 8 sites de mesure. A titre de comparaison, la Figure 5.4.C reprend les résultats de la Figure 5.2 obtenus par randoms sampling.

La comparaison des Figure 5.4.A et Figure 5.4.B avec la Figure 5.4.C montre que les erreurs d'estimation avec les approches de *target sampling* sont systématiquement plus faibles que celles obtenus avec le *random sampling.* Pour les deux approches, l'erreur d'estimation moyenne se situe autour de 13% et ne dépend que très peu des valeurs prises pour le critère de variance. Contrairement au random sampling, la courbe ne présente pas un minimum à partir de laquelle les erreurs d'estimation augmentent rapidement.

Ce résultat est à associer aux valeurs prises pour le critère de variance. Pour le *target sampling* basé sur les quantiles, ces valeurs sont dispersées entre 10^{-10} et 10^{-2} , et entre 10^{-9} et 10^{-1} pour celui basé sur les kmeans. Pour les deux approches, les valeurs maximales du critère de variance restent inférieures aux valeurs à partir desquelles l'erreur d'estimation augmente de manière importante pour le *random sampling*.

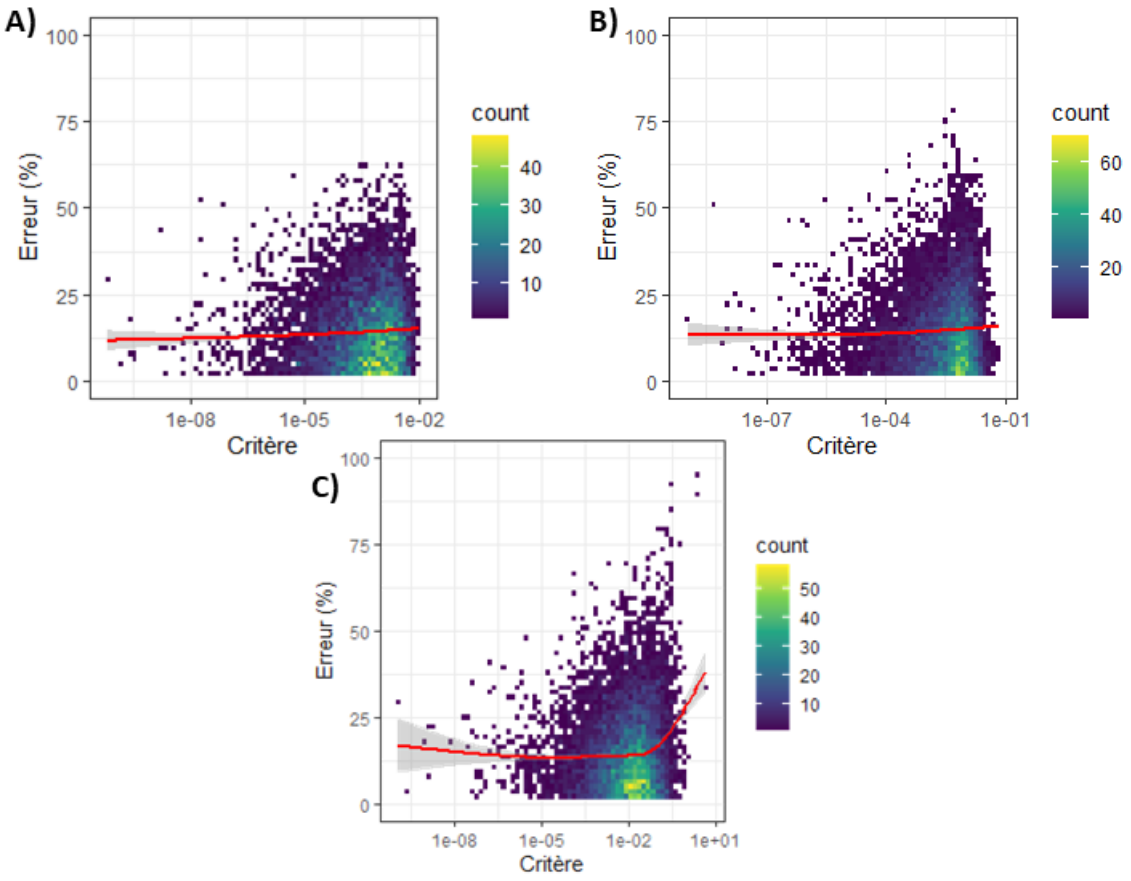

Figure 5.4 : Les approches de target sampling sont associées à des valeurs de critère de variance moindres, limitant ainsi l'erreur d'estimation. La figure compare un target sampling basé sur l'approche des quantiles (4A) et celle des kmeans (4B) au random sampling (4C).

Ce résultat explique d'un point de vue théorique, l'intérêt des approches mises en place de manière plus ou moins empiriques dans la littérature existante (Meyers et al. 2020, Oger et al. 2020, Carillo et al. 2016, Araya-Alman et al. 2017). Celles-ci font appel à des méthodes basées sur les données auxiliaires telles que les intervalles entre les quantiles pour le choix des sites de mesure. En contraignant les valeurs attributaires que peuvent prendre les sites de mesure, ces approches tendent (i) à réduire la différence entre la moyenne de l'échantillon et celle de l'ensemble de la population, qui correspond au numérateur du critère de variance $(\overline{X_R} - \overline{X_N})^2$, et (ii) à augmenter la dispersion des valeurs de l'échantillon, soit le dénominateur du critère de variance $\sum_{i \in S}(X_i - \overline{X_N})^2$. Ce sont ces deux phénomènes associés qui limitent les valeurs du critère de variance et donc la variance de l'estimation.

5.4 Réflexions complémentaires

Les résultats présentés dans la Figure 5.2 mettent en évidence, sur des données réelles, que les erreurs d'estimation peuvent être reliées au critère de variance *le model sampling*. Le choix des sites de mesure selon leurs valeurs attributaires pour la donnée auxiliaire apparaît donc comme un outil permettant de contrôler une partie de l'erreur d'estimation. Toutefois, la Figure 5.3 montre que la donnée auxiliaire n'a d'intérêt que s'il existe une corrélation élevée entre la donnée auxiliaire et la variable d'intérêt.

Cette connaissance rend possible la mise en place d'un plan d'échantillonnage adapté aux objectifs poursuivis par les professionnels de la viticulture. En admettant que la corrélation entre variable

d'intérêt et variable auxiliaire est connue, il est alors possible de mieux raisonner le nombre de sites d'échantillonnage sur la base d'un compromis entre la qualité de la prédiction souhaitée et l'effort d'échantillonnage nécessaire. La méthode de sélection des sites de mesure est choisie au regard des contraintes opérationnelles (temps, distance, pénibilité …), soit en cherchant directement à minimiser le critère de variance et la variance de l'estimation, soit en utilisant les méthodes du target sampling qui permettent de contrôler indirectement cette variance (Figure 5.4).

Pour un échantillonnage donné, l'utilisation directe de l'équation du critère de la variance de l'estimation permet de prédire l'erreur attendue pour l'estimation à partir du nombre de sites de mesure et des valeurs attributaires sélectionnées. La confiance qu'il est possible de placer dans une estimation est ainsi rendue quantifiable. Il s'agit d'un enjeu majeur des problématiques d'échantillonnage en production végétale. La caractérisation de cette variabilité reste néanmoins dépendante de la connaissance de l'écart type des résidus du modèle. Celui-ci, propre à chaque parcelle, peut apparaitre difficilement estimable selon les cultures et les grandeurs considérées. La mise en place de références permettant de connaitre les paramètres attendus d'un modèle unissant les variables d'intérêt et données auxiliaires utilisées en production végétale représente un enjeu pour le développement des approches de model sampling.

Le critère proposé est basé sur des hypothèses relativement simples qui, même si elles ne sont pas toujours totalement vérifiées sur des données réelles, assurent que son utilisation soit applicable à des parcelles réelles. Les tests présentés sur un nombre limité de parcelles correspondant à des conditions différentes permettant d'être plutôt confiant quant à son utilisation concrète. Cependant la robustesse de la méthode et la validité des hypothèses sur lesquelles elle s'appuie sont autant de points nécessitant d'être testés dans une plus grande diversité de situations. Le modèle linéaire se base notamment sur l'hypothèse d'indépendance des résidus, cette hypothèse revient à considérer que la structure spatiale de la variable d'intérêt est entièrement expliquée par la donnée auxiliaire. Ces travaux pourraient être étendus à un cadre plus général adaptant l'expression de la variance de la résiduelle du modèle intégrant une structure spatiale. Par ailleurs, les fonctionnalités de l'approche et les considérations théoriques pourraient également être élargies à d'autres types de modèles ou à des données de plus grandes dimensions pour la rendre plus adaptables à la diversité des systèmes de production végétale.

5.5 Conclusion :

Cet article démontre l'intérêt de raisonner le choix de sites de mesure en adéquation avec la méthode d'inférence utilisée. Dans le cas d'une inférence basée sur l'étalonnage d'un modèle linéaire, l'article fait le lien entre les propriétés statistiques de l'estimation et la stratégie d'échantillonnage. L'impact du nombre de sites de mesure ainsi que des valeurs de la donnée auxiliaire sur ces sites sur la variance de l'estimation est quantifié. Idéalement, les sites de mesure doivent ainsi présenter une moyenne proche de celle de l'ensemble des sites de la parcelle. La dispersion de ces mêmes valeurs autour de leur moyenne permet dans le même temps de mieux estimer les coefficients du modèle. L'article propose un critère permettant de contrôler la qualité attendue pour l'estimation. Sur un exemple appliqué à l'estimation du rendement viticole, les résultats montrent que les approches de target sampling basées sur des algorithmes de classification tels que proposées dans la littérature tendent à sélectionner des échantillons ayant des propriétés intéressantes au regard de ce critère.

Chapitre 6 : Résolution du problème d'optimisation du parcours d'échantillonnage

Dans les chapitres précédents (Chapitres 2, 3 et 5), ce document s'est principalement focalisé sur la définition de règles pour la sélection des sites de mesure qui constitueront l'échantillon final. Ce chapitre présente plus en détails les algorithmes développés pour trouver des solutions respectant ces règles tout en optimisant le parcours d'échantillonnage.

6.1 Définition du problème d'optimisation des parcours d'échantillonnage

Les données disponibles constituent le point de départ des approches mises en place. Initialement, on dispose d'un jeu de données contenant la liste des k **ceps** (ou ensembles de ceps) envisagés pour l'échantillonnage.

Chaque *cep de vigne* est décrit par deux variables :

- Un numéro identifiant (allant de 1 à k) ;
- La valeur de la donnée auxiliaire associée à ce cep.

Un deuxième jeu de données présente la liste des l **sites de mesure**. Un site de mesure correspond à un cep et à un inter-rang. Chaque cep étant accessible depuis deux inter-rangs, on a généralement $l = 2k$. Il est cependant possible de retirer préalablement certains sites (sites aberrants, bordures de parcelle, sites inaccessibles, etc.).

Chaque site y est décrit par cinq variables :

- Un numéro de site (allant de 1 à l) ;
- Ses coordonnées spatiales sur la parcelle selon X et selon Y
- Le numéro de l'inter-rang d'accès pour la mesure
- Le numéro identifiant du cep auquel il correspond.
- La longueur du plus court trajet à pied vers chacun des autres sites de mesure. Ces distances sont préalablement calculées selon les principes présentés dans le chapitre 1 et stockées dans une matrice de distance donnant la distance de marche associée à chaque couple de sites donné.

Le **problème d'optimisation** à résoudre peut se résumer ainsi :

> *Trouver un ensemble de n (n correspondant au nombre d'échantillons souhaité sur la parcelle) sites de mesure sélectionnés parmi l'ensemble des l sites candidats tels que (i) l'estimation obtenue grâce à cet ensemble de sites respecte un certain nombre de critères, principalement associés à la donnée auxiliaire, et que (ii) la longueur du parcours reliant les sites de mesure soit minimale.*

Les méthodes, les contraintes et les critères envisagés pour atteindre ce double objectif ont évolué au cours de la thèse et plusieurs approches ont été considérées. Plusieurs variantes de ce problème d'optimisation ont été implémentées et testées au cours de la thèse. Par souci de simplification, seule la dernière version du problème sera détaillée dans ce chapitre. Les différentes variantes du problème ont concerné deux principaux aspects :

a) Le critère permettant de contrôler la variance de l'estimation résultant de l'échantillon sélectionné. Les différentes solutions testées ont porté sur les méthodes de partitionnement

des sites potentiels, soit en considérant des groupes basés sur les quantiles, soit en définissant un critère général basé sur la variance de l'échantillon sur la donnée auxiliaire (Eq. 5.30, section 5.2.6)

b) La distance minimum, notée d_{min}, entre chaque couple de sites de mesure. Cette contrainte est une réponse au questionnement évoquées dans les chapitres 1 et 3 autour du risque d'autocorrélation entres sites de mesure trop proches. Des sites de mesure trop rapprochés et corrélés spatialement tendent en effet à apporter une information redondante ce qui tend à augmenter les erreur d'estimation (Figure 6.1).

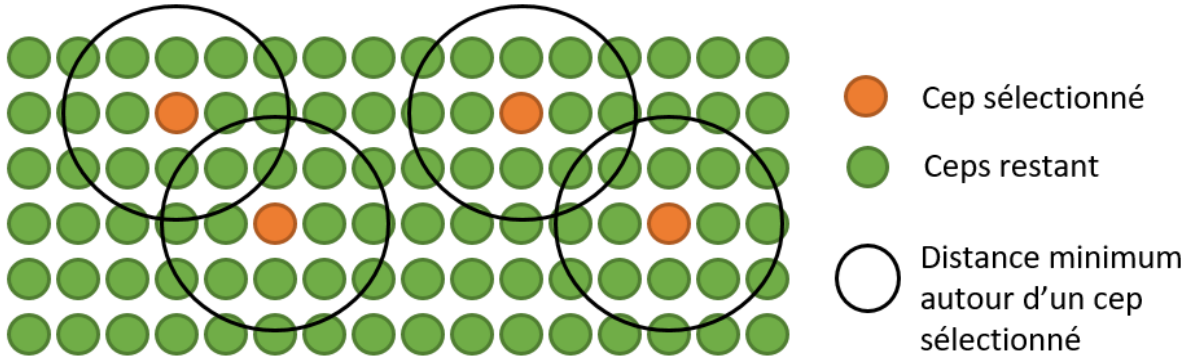

Figure 6.1 : Deux points ou sites de mesure sélectionnés dans un échantillon doivent être séparé par une distance minium afin de garantir leur relative indépendance.

La longueur du parcours est définie comme la longueur totale du chemin passant au moins une fois par chaque site de mesure de l'échantillon et qui commence et termine au même point de départ. L'objectif est donc de trouver l'échantillon respectant les contraintes sur la précision de l'estimation et qui minimise la longueur du parcours. Cette minimisation présente un double enjeu, il est en effet nécessaire de choisir les bons sites de mesure, mais également que ceux-ci soient parcourus dans le bon ordre. Ce problème se rapproche du problème du voyageur de commerce (ou TSP pour Travelling Salesman Problem) (Lawler et al. 1985 ; Papadimitriou 1977) pour lequel l'objectif est d'ordonner le passage dans un ensemble de villes afin de trouver le plus court chemin visitant une fois chaque ville. La différence avec le problème d'échantillonnage réside dans le fait que les sites à visiter ne sont pas imposés mais doivent être choisis parmi un ensemble de sites potentiels.

Trouver une solution à ce problème n'est pas aisé car il présente une forte combinatoire. En prenant en compte l'ordre dans lequel sont parcourus les sites, il existe $\frac{l!}{(l-n)!}$ manières différentes d'échantillonner une parcelle. n variant entre 5 et 10 et l prenant généralement des valeurs comprises entre 1000 et 10000, le nombre d'échantillonnages possibles varie entre 10^{15} et 10^{40}.

Pour atteindre les objectifs définis, deux méthodes ont été mises en place au cours de la thèse. Au départ, sachant que la formulation du problème allait évoluer et qu'il était combinatoire, la programmation par contraintes a été considérées. Ce choix était justifié à la fois pour les performances et la grande expressivité (définie en informatique comme la capacité d'un langage à décrire des problèmes ou des solutions). Des travaux récents ont d'ailleurs montré l'apport de la programmation par contrainte pour des problématiques la vendange sélective (Briot et al., 2015) et la gestion durable des terres (Justeau-Allaire et al., 2019).

Toutefois, des limites liées au temps nécessaire pour résoudre les grosses instances propres à notre problème sont vites apparues. Cette limite explique pourquoi, dans une deuxième phase, des méthodes heuristiques ont été préférées bien que ne garantissant pas une solution optimale, elles

permettent de proposer une solution convenable dans un temps raisonnable. Nous allons présenter brièvement ces deux approches.

6.2 Programmation par contraintes et recherche d'un parcours d'échantillonnage optimal

6.2.1 La programmation par contraintes, paradigme informatique adapté à la résolution de problèmes d'optimisation

6.2.1.1 *Principes généraux*

Est désigné par programmation par contraintes (ou CP en anglais pour *Constraint Programming*) un paradigme de programmation faisant appel à des techniques issues traditionnellement de l'intelligence artificielle. La programmation par contraintes, conçue pour résoudre une grande variété de problèmes combinatoires repose sur une modélisation particulière sous forme de problème de satisfaction de contraintes (ou CSP pour Constraint Satisfaction Problem). Les problèmes y sont définis par un ensemble de variables, de domaines et de contraintes (Rossi, 2006).

6.2.1.2 *Problèmes de satisfaction de contraintes*

Les **variables** $X = \{x_1, x_2, x_3, \dots\}$ représentent les différentes grandeurs dont l'attribution à une valeur permet la résolution du problème. Chacune des variables peut prendre sa valeur dans un ensemble qui lui est propre, appelé **domaine**. On note $D(x_1)$ le domaine de la variable x_1. Si $D(x_2) = \{1,2\}$ alors la variable x_2 peut seulement prendre les valeurs $x_2 = 1$ ou $x_2 = 2$.

Les **contraintes** représentent des liens logiques entre les différentes variables. Elles portent sur des sous-ensembles de variables et restreignent les valeurs que peuvent prendre les variables impliquées.

La résolution d'un CSP se fait en associant à chaque variable, une valeur présente dans son domaine, on parle d'**instanciation**. Une instanciation est dite **cohérente** si et seulement si elle satisfait les contraintes concernées par les variables qu'elle implique. Dans le cas contraire, l'instanciation est qualifiée d'incohérente et les contraintes non satisfaites sont dites violées. Par ailleurs, une instanciation sera dite **totale** si elle concerne toutes les variables du problème ou partielle lorsque toutes les variables ne sont pas encore associées à une valeur. La résolution d'un CSP revient alors à trouver, si elle existe, une instanciation qui soit à la fois totale et cohérente.

La force de la programmation par contraintes réside dans le fait qu'elle est permet, soit de trouver une solution à un problème donné, soit de prouver qu'il n'existe pas de solution.

Dans le cas d'un problème d'optimisation, la modélisation du problème comporte une variable représentant l'objectif à minimiser (ou maximiser). La résolution de ce type de problème consiste à trouver une première solution affectant une valeur cohérente à l'objectif à optimiser. La phase d'optimisation se poursuit par la résolution successive de variantes du problème intégrant une contrainte additionnelle qui impose de trouver une valeur de l'objectif inférieure à celle de la précédente solution. La recherche s'arrête lorsqu'il n'y a pas de solution à la dernière variante du problème, ce qui prouve que la valeur de l'objectif dans la dernière solution était bien minimale.

6.2.1.3 *Solveur et résolution d'un problème de satisfaction de contraintes*

La résolution de problèmes de satisfaction de contraintes est confiée à des solveurs dont le principe est d'instancier les variables tour à tour. Une partie de l'efficacité de ces solveurs provient de mécanismes de filtrage et de propagation. Le filtrage, toujours associé à une contrainte, retire du

domaine des variables les valeurs ne respectant pas la contrainte. Prenons comme exemple le cas où on dispose de deux variables x_1 et x_2. La variable x_1 est instanciée et prend comme valeur $x_1 = 1$ et le domaine de x_2 contient trois valeurs : $D(x_2) = \{1,2,3\}$. Pour une contrainte imposant $x_1 \neq x_2$, le filtrage selon cette contrainte retirerait la valeur 1 de $D(x_2)$. On obtient alors $D(x_2) = \{2,3\}$.

La propagation consiste à appliquer à bon escient les filtrages associés aux différentes contraintes, par exemple chaque fois que les domaines des variables sont modifiés (par un précédent filtrage ou une instanciation). Un mécanisme de retour en arrière -- appelé *backtracking* -- permet de revenir sur l'instanciation d'une variable lorsque son domaine est entièrement vidé et qu'il n'est par conséquent plus possible de trouver une instanciation totale et cohérente. Sur la base de ces principes, les solveurs sont capables de résoudre automatiquement un CSP en ayant recours à des techniques complémentaires que nous n'aborderont pas ici. Si bon nombre de solveurs sont aujourd'hui capables de résoudre efficacement beaucoup de problèmes, une utilisation avancée de ces outils nécessite néanmoins une expertise sur l'ordre d'instanciation des variables ou le choix des algorithmes de filtrage afin d'améliorer le temps de résolution.

6.2.1.4 *Utilisation dans le contexte de la thèse*

La programmation par contraintes a été la première approche choisie pour la résolution du problème d'échantillonnage. Face à la combinatoire de ces problèmes et à la nécessité de tester différentes variantes, ce paradigme est légitimement apparu comme un outil de choix. L'expressivité de la programmation par contraintes a facilité l'évolution des modèles qui revenait simplement à changer des contraintes et variables du CSP.

Dans ce cadre, plusieurs modélisations de différentes versions du problème ont vu le jour tout au long de la thèse. La première a été présentée dans un papier de conférence présentée aux quatorzièmes journées francophones de la programmation par contraintes (Oger et al. 2018). La modélisation présentée dans la sous-partie suivante, une des dernières en date, mobilise le critère de variance présenté dans le Chapitre 5 :.

6.2.2 Implémentation de la recherche d'un plan d'échantillonnage optimal avec la programmation par contraintes

6.2.2.1 *Variables et domaines*

Dans le cas du problème d'échantillonnage en viticulture, l'objectif est d'identifier les n sites de mesure qui constitueront l'échantillon. On définit donc un ensemble de $n + 1$ variables $S = \{S_0, S_1, \dots, S_n\}$, où pour tout $i \in \{1, \dots, n\}$, S_i représente le $i^{ème}$ site de mesure et S_0 correspond au point de départ, fixé par l'utilisateur en dehors des rangs. L'échantillonnage passant par les sites numéro 3, 10, 27 et 38 correspondra à l'instanciation : $\{S_0 = Départ, S_1 = 3, S_2 = 10, S_3 = 27, S_4 = 38\}$.

Chaque S_i peut prendre comme valeur l'ensemble des sites présents sur la parcelle numérotés de 1 à l et on a $D(S_i) = \{1,2, \dots, l\}$.

Un deuxième ensemble de variables $C = \{C_1, C_2, \dots, C_n\}$, correspond aux ceps qui seront échantillonnés. Ces variables, très proche de celles de l'ensemble S auxquelles elles sont directement associées, prennent pour valeur les numéros de cep. Le domaine de ces variables correspond à la liste des ceps : $D(C_i) = \{1,2, \dots, k\}$

Un troisième ensemble de variables contient les distances entre chaque couple de sites successifs, en incluant le point de départ. Il est noté $cost = \{cost_0, cost_1, \dots, cost_n\}$. Chaque variable $cost_i$

représente la distance entre les sites S_i et S_{i+1}. Par ailleurs, $cost_n$ représente la distance entre les sites S_n et S_0.

Une variable notée $objective$ stocke la somme de ces coûts individuels. Cette variable représente la longueur du parcours d'échantillonnage et correspond à la grandeur que l'on cherche à minimiser.

Enfin une dernière variable appelée $critère$ correspond au critère de variance calculé pour l'échantillon. Cette variable prend ses valeurs dans un ensemble de nombres réels, son domaine va de 0 à $seuil_{critère}$. La valeur de $seuil_{critère}$, la valeur maximum du critère de variance, est choisie par l'opérateur. Pour les expérimentations, en fonction des résultats présentés au chapitre 5, nous avons utilisé une valeur du critère égale à 10^{-4} de manière à limiter au mieux la variance de l'estimation finale.

6.2.2.2 *Modélisation des contraintes*

Plusieurs contraintes limitent les valeurs que peuvent prendre les variables lors de l'instanciation. En premier lieu, deux contraintes imposent que les valeurs au sein de l'ensemble S et de l'ensemble P soient différentes :

$$AllDiff(S_0, S_1, S_2, \dots, S_n) \qquad Eq.6.1$$

Et

$$AllDiff(C_1, C_2, \dots, C_n) \qquad Eq.6.2$$

Une contrainte dite de « *channeling* » fait le lien entre les variables S et les variables C à partir des valeurs présentent dans le jeu de données :

$$\forall i \in \{1, \dots, n\}, C_i = Cep(S_i) \qquad Eq.6.3$$

Deux autres contraintes similaires lient les variables S aux variables $cost$ et $objective$ à l'aide des distances préalablement calculées:

$$\forall i \in \{0, \dots, n-1\}, cost_i = dist(S_i,\ S_{i+1})\ \ \&\ \ cost_n = dist(S_n,\ S_0) \qquad Eq.6.4$$

Et

$$objective = \sum_{i=0}^{n} cost_i \qquad Eq.6.5$$

La variable associée au critère de variance est gérée de façon similaire à partir des valeurs de données auxiliaires associées aux ceps. On note DA la fonction qui associe à chaque cep sa valeur pour la donnée auxiliaire en utilisant les données ; $\overline{DA(C)}$ la moyenne des valeurs de données auxiliaires sur l'échantillon et $\overline{DA(R)}$ la moyenne des valeurs des données auxiliaires des ceps exclus de l'échantillon. La contrainte s'exprime alors selon la formule du critère de variance :

$$critere = \frac{(\overline{DA(C)} - \overline{DA(R)})^2}{\sum_{C_i \in C}\ (DA(C_i) - \overline{DA(C)})^2} \qquad Eq.6.6$$

Il est important de noter qu'une instanciation définit non seulement les sites à visiter mais aussi l'ordre dans lequel ils seront visités. L'instanciation $\{S_1 = 3, S_2 = 10, S_3 = 27, S_4 = 38\}$ correspond au même échantillon que $\{S_1 = 3, S_2 = 27, S_3 = 10, S_4 = 38\}$ mais l'ordre dans lequel les sites sont visités n'est pas le même et la longueur des parcours qui résulte de ces ordres de visite est différente. On remarque

en particulier que $\{S_1 = 3, S_2 = 10, S_3 = 27, S_4 = 38\}$ et $\{S_1 = 38, S_2 = 27, S_3 = 10, S_4 = 3\}$ correspondent au même échantillon et au même parcours, les sites sont simplement visités dans l'ordre inverse. Ces deux solutions sont dites symétriques. Pour éviter de considérer deux fois chaque solution à cause de cette symétrie, on ajoute une autre contrainte :

$$S_1 > S_n \qquad Eq.6.7$$

La dernière contrainte concerne l'aspect b) évoqué dans la définition du problème au paragraphe 6.1. Pour un cep donné, la fonction $forbidden$ renvoie la liste des ceps qui ne peuvent être inclus avec lui dans l'échantillon car géographiquement trop proches (séparés par une distance euclidienne inférieure à d_{min}). Il en découle une contrainte garantissant la distance minimum entre chaque couple de ceps :

$$\forall (i,j) \in \{1, \dots, n\}^2 \; et \; i \neq j, C_i \notin forbidden(C_j) \qquad Eq.6.8$$

Les huit contraintes présentées ici permettent de modéliser le problème énoncé.

Le modèle a été implémenté grâce à la librairie Choco (Prud'homme et al. 2016) dans le langage de programmation Java. Les contraintes standards auraient permis d'implémenter le modèle mais, par souci d'efficacité, nous avons été amenés à développer des propagateurs spécifiques pour quelques contraintes. Par ailleurs, comme il s'agissait principalement de valider les modèles nous n'avons pas essayer d'optimiser son implémentation ni d'ajuster les paramètres de résolution. Le développement d'une application opérationnelle nécessiterait donc des investigations plus poussées.

Pour la contrainte 6.6, plus complexe, le filtrage reste à l'heure actuelle très limité et nécessiterait des travaux additionnels s'inspirant par exemples des travaux menés sur la contrainte « deviation » (Schauss et al. 2007).

Grace à l'expressivité de la programmation par contraintes, il suffit de modifier quelques variables et contraintes pour modéliser une variante du problème. Par exemple, pour sélectionner l'échantillon optimal selon l'approche des quantiles plutôt que selon le critère de variance, 3 modifications suffisent. La variable $critere$ est délaissé au profit d'un nouvel ensemble de n variables $Q_i, i \in \{1, \dots, n\}$. Ces variables prennent pour valeurs les quantiles représentés par chacun des sites de mesure. Elles sont reliées aux variables ceps par une contrainte de *channeling*. Enfin, une contrainte *AllDifferent* impose que les $Q_i, i \in \{1, \dots, n\}$ soient toutes différentes.

6.3 Recherche opérationnelle et indentification d'un parcours d'échantillonnage en temps limité

6.3.1 De la programmation par contraintes aux outils de la recherche opérationnelle

A travers la programmation par contrainte, présentée dans la partie précédente, on dispose d'une méthode efficace pour trouver une solution optimale. Plusieurs modifications apparues au cours de la thèse comme l'ajout d'une distance minimale entre chaque couple de site ou encore la recherche d'un échantillon en fonction de son critère de variance ont augmenté la complexité du problème. Dans les situations les plus combinatoires, cette méthode perd de son efficacité et la recherche d'une solution se compte en heure ou en jours. Nous avons alors mis en place une approche adaptée à la recherche d'une solution dans un temps limité. Dans cette partie, on se propose de résoudre le problème en utilisant des méthodes appartenant à la recherche opérationnelle (Werra & al. 2003). Ces méthodes se basent sur des heuristiques qui, à défaut de garantir le meilleur échantillon, permettent de trouver de bonnes solutions dans un temps raisonnable.

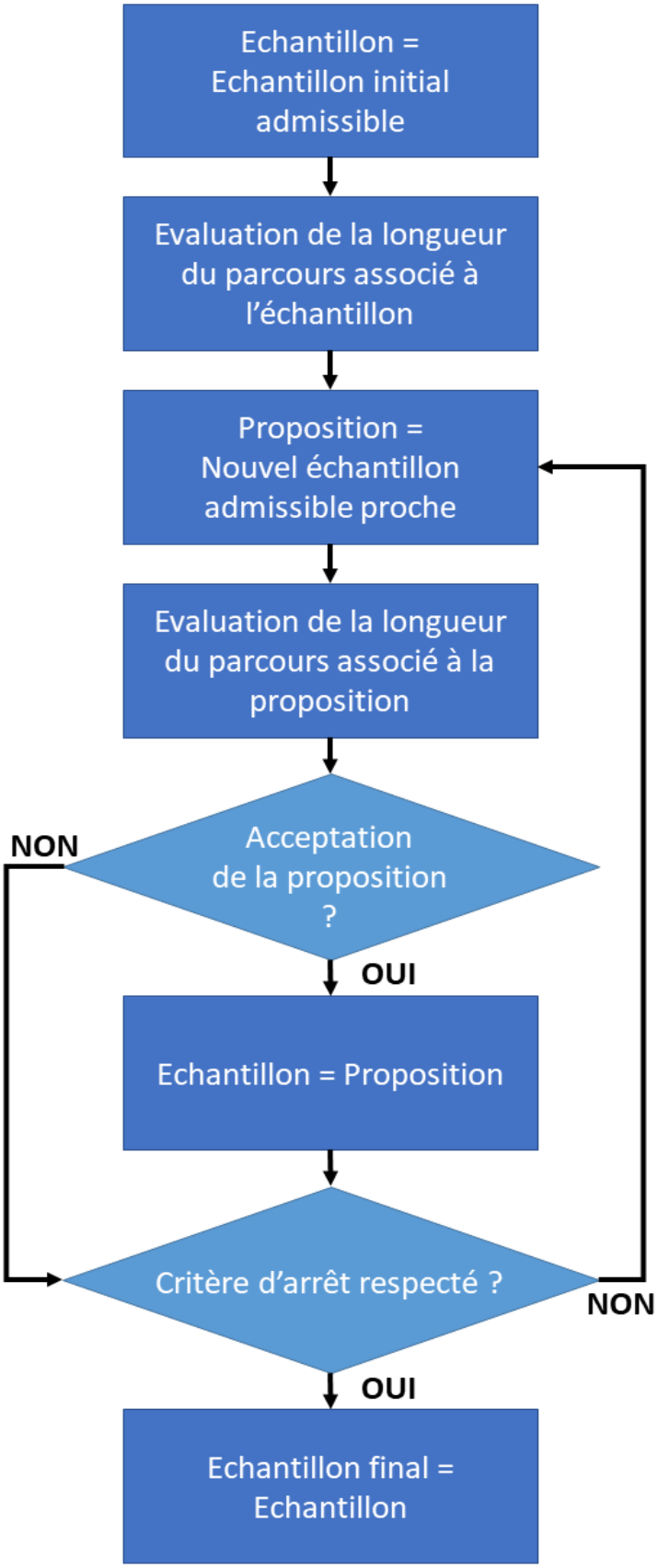
Echantillon = Echantillon initial admissible
Evaluation de la longueur du parcours associé à l'échantillon
Proposition = Nouvel échantillon admissible proche
Evaluation de la longueur du parcours associé à la proposition
Acceptation de la proposition ?
NON
OUI
Echantillon = Proposition
Critère d'arrêt respecté ?
NON
OUI
Echantillon final = Echantillon

Figure 6.2 : Logigramme du fonctionnement général de l'approche mettant en œuvre les outils de la recherche opérationnelle.

6.3.2 Mise en place d'une approche pour la recherche d'un plan d'échantillonnage

6.3.2.1 *Fonctionnement général*

L'approche d'optimisation générale est réalisée selon le principe du recuit simulé (Lutton & Bonomi, 1998). Les algorithmes suivant ce principe ont pour but d'optimiser une grandeur (ici la longueur du parcours) par modifications élémentaires itératives (Figure 6.2). Pour résoudre les problèmes de minimum local liées aux méthodes de recherche locale, le recuit simulé repose sur une méthode originale pour contrôler les modifications basées avec la notion de température.

Par la suite, on définit comme **échantillon admissible** tout échantillon respectant les contraintes présentées dans la partie 6.1. Celles-ci imposent que les données auxiliaires associées aux ceps constituant cet échantillon respectent un critère de variance inférieur au seuil fixé par l'utilisateur ($seuil_{critère}$) et que les ceps de l'échantillon soient séparés deux à deux par une distance minimum d'au moins d_{min}.

Algorithm 1 Algorithme du recuit simule

$Echantillon \leftarrow Echantillon\ initale\ admissible$
$Longueur \leftarrow LongueurParcours(Echantillon)$
$Temperature \leftarrow Temperature\ initiale$
$i \leftarrow 0$
while $i \leq seuil\ d'iteration$ **do**
  $Proposition \leftarrow Modification(Echantillon)$
  $LongueurProp \leftarrow LongueurParcours(Proposition)$
  if $LongueurProp < Longueur$ **then**
    $Echantillon \leftarrow Proposition$
    $Longueur \leftarrow LongueurProp$
  else if $rand(0,1) \leq exp(-(LongueurProp - Longueur)/Temperature)$ **then**
    $Echantillon \leftarrow Proposition$
    $Longueur \leftarrow LongueurProp$
  end if
  $Temperature \leftarrow Temperature\ initiale/(1 + log(1 + i))$
  $i \leftarrow i + 1$
end while

Figure 6.3 : Algorithme du recuit simulé appliqué à l'optimisation de la longueur des parcours d'échantillonnage.

A un instant donné, l'algorithme dispose d'un certain échantillon admissible. L'objectif est de proposer un nouvel échantillon admissible légèrement différent de l'échantillon courant. Si la nouvelle proposition améliore la longueur du parcours, alors le nouvel échantillon est systématiquement accepté. Si au contraire elle dégrade cette longueur, la nouvelle proposition est acceptée avec une certaine probabilité. L'acceptation d'un « mauvais » échantillon tend à éviter de s'enfermer trop vite dans un minimum local. La probabilité d'accepter un « mauvais » échantillon est soumise à deux facteurs :

- Elle est inversement proportionnelle à la différence de longueur entre le parcours de l'échantillon courant et celui de la nouvelle proposition, ainsi plus un nouvel échantillon s'écarte de la distance de parcours obtenue avec l'échantillon courant, moins il a de chances d'être accepté
- Suivant une approche de *recuit simulé*, la probabilité d'accepter un « mauvais » échantillon est également proportionnelle à une température qui diminue avec le temps. Au fur et à mesure des itérations, l'algorithme est moins à même d'accepter une mauvaise proposition.

Selon la fonction de décroissance de la température utilisée, ce phénomène assure la convergence de l'algorithme vers la solution optimale dans un temps infini (Aarts & Laarhoven, 1985). Pour un temps fini, elle permet d'obtenir un échantillon dont les propriétés s'approchent de cet optimum. Le critère d'arrêt de cet algorithme, décrit dans la Figure 6.3, est exprimé en nombre d'itérations ou en temps de calcul.

Le fonctionnement général de cet algorithme se base sur deux hypothèses. La première est que l'on est capable de proposer de nouveaux échantillons admissibles proche de l'échantillon actuel. La seconde est de pouvoir évaluer la qualité des nouveaux échantillons en calculant la longueur de parcours associée à un échantillon.

6.3.2.1 *Recherche d'échantillons admissibles*

On doit trouver un nouvel échantillon admissible qui soit proche du précédent. Pour être admissible, cet échantillon doit correspond à un critère de variance inférieur au seuil fixé et une distance minimale doit séparer chaque couple de ceps qui le compose. Comme l'approche générale, cette recherche fonctionne de manière itérative avec des modifications de proche en proche à partir de l'échantillon courant de l'algorithme général. Les propositions de modification se font sur un cep à la fois en le remplaçant par un de ses voisins directs (Figure 6.4).

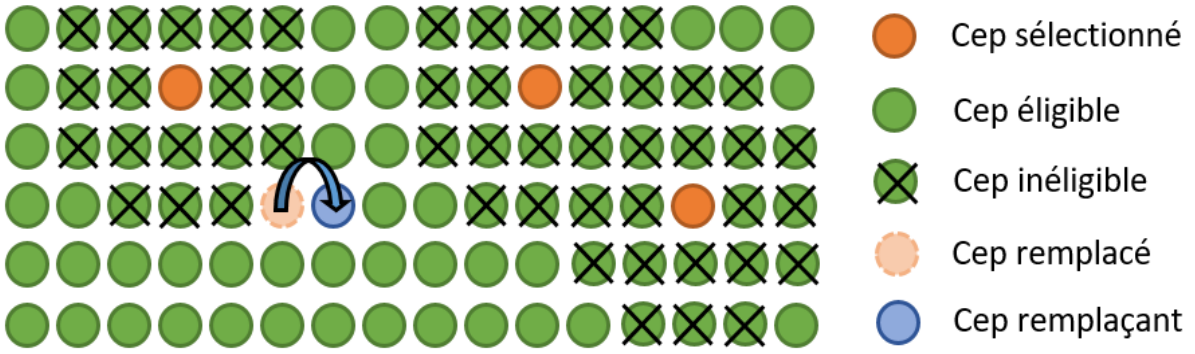

Figure 6.4 : Modification d'un échantillon en remplaçant un cep par un de ses voisins directs. Le nouveau cep ne peut pas se trouver à une distance inférieure à d_{min} d'un autre cep constituant l'échantillon.

Le cep modifié peut aussi être remplacé par un cep plus éloigné de façon aléatoire, avec une faible probabilité, ou lorsqu'il n'est pas possible de le remplacer par un de ses voisins. Ces remplacements exceptionnels sont appelés mutations (Figure 6.5).

Le nouvel échantillon ainsi obtenu respecte la distance devant séparer deux ceps (ii) Figure 6.4 & Figure 6.5). Si le nouvel échantillon correspond à un critère de variance supérieur à $seuil_{critère}$, alors l'échantillon n'est pas admissible et l'algorithme itère une nouvelle modification. L'échantillon non admissible peut tout de même être conservé ou non comme point départ de l'itération suivante selon une probabilité qui dépend de l'écart entre son critère de variance calculé et le $seuil_{critère}$. Les itérations s'enchainent par modifications successives jusqu'à obtenir un échantillon dont le critère de

variance est inférieur à $seuil_{critère}$. L'échantillon admissible ainsi obtenu correspondra à la proposition d'échantillon du recuit, sous réserve que celui-ci soit différent de l'échantillon courant.

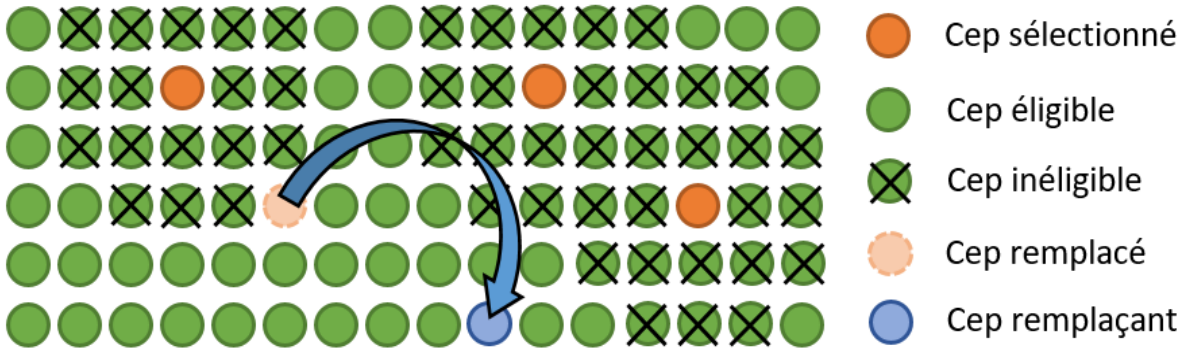

Figure 6.5 : Modification d'un échantillon par mutation, évènement rare au cours duquel un cep est remplacé par un autre, indépendamment de la distance qui les sépare.

6.3.2.1 *Algorithme des fourmis et longueur du parcours associé à un échantillon*

L'algorithme général repose également sur la capacité à déterminer la longueur du parcours associée à un échantillon donné. L'objectif est de trouver le plus court parcours permettant de relier tous les ceps. Le parcours est contraint par la structure de la parcelle.

Choisir le parcours de longueur minimale revient à ordonner correctement les ceps composant l'échantillon, à la manière d'un problème du voyageur de commerce. On cherche à trouver le parcours de longueur minimale reliant un ensemble de ceps où chacun d'entre eux possède deux accès différents (un par inter-rang). Plusieurs méthodes appartenant à la recherche opérationnelle permettent de résoudre les problèmes de ce type.

L'approche retenue ici pour résoudre ce problème se base sur l'algorithme des colonies de fourmis, déjà appliqué au problème du voyageur de commerce (Dorigo & Gambardella, 1997 ; Stutzle & Hoos, 1997). L'approche par colonie nous permet de prendre en compte la singularité du problème où chaque cep est accessible depuis deux inter-rangs.

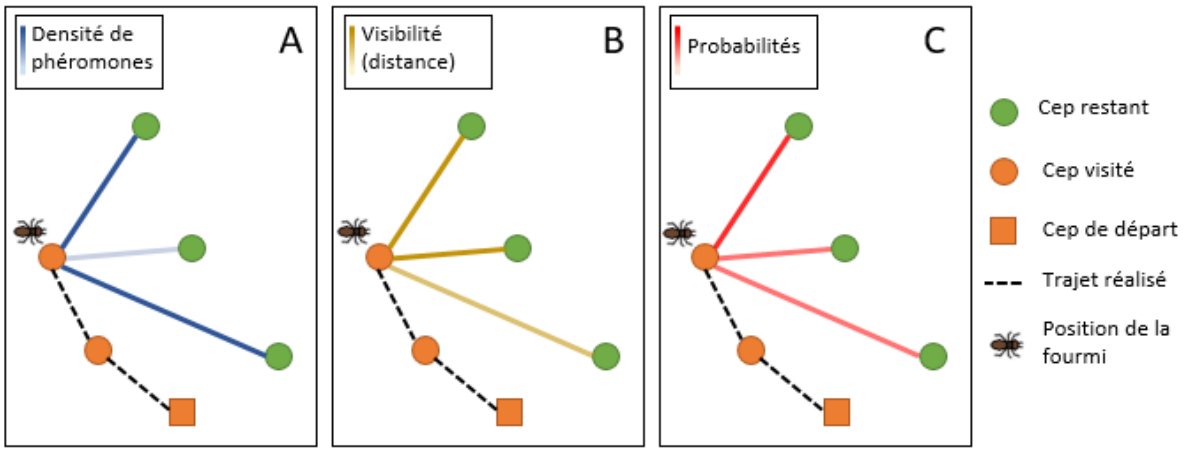

Figure 6.6 : Le choix du prochain cep dans l'algorithme des fourmis (C) dépend de la quantité de phéromones déposées (A) et de la visibilité (B).

Cet algorithme met en œuvre une population de fourmis virtuelles, il est basé sur un fonctionnement itératif. A chaque itération, chaque fourmi doit visiter un des deux sites de mesure associés à chaque cep. On définit comme **arrête** le chemin reliant deux sites de mesure. Chaque fourmi dépose une

quantité de phéromones sur les arêtes qu'elle a parcourues. La quantité de phéromone étant la même pour chaque fourmi, plus le parcours est court plus la fourmi dépose de phéromones par arête. A chaque itération, une partie des phéromones présentes sur les arêtes s'évapore. Lorsqu'une fourmi est sur un site de mesure, le choix du site suivant dépend de deux facteurs. Plus une arête menant à un site est longue, moins elle a de chance d'être choisie par la fourmi. On parle de visibilité (Figure 6.6). Par ailleurs, plus il y a de phéromones disposées sur une arête, plus celle-ci aura de chance d'être choisie.

Au cours des itérations le trajet des fourmis converge vers une solution qui sera retenue comme le parcours associé à l'échantillon. L'algorithme des colonies de fourmis ne garantit en rien de trouver de parcours de longueur minimale mais permet d'obtenir rapidement une solution qui s'approche de cette optimalité.

6.3.2.2 *Paramétrage et implémentation*

Les algorithmes mis en place dépendent de nombreux paramètres qui vont jouer sur la convergence vers une solution. Ces paramètres interviennent à différentes étapes de la méthode.

Pour l'algorithme de recuit simulé, on retrouve quatre principaux paramètres :

- Le nombre d'itérations ou le temps de calcul pour le critère d'arrêt ;
- La température initiale ;
- La fonction de décroissance de la température ;
- La fonction de probabilité qui permet de calculer les chances d'accepter un mauvais échantillon.

Le choix de ces valeurs est déterminant dans la qualité de la solution et la convergence de l'algorithme général. Le nombre d'itérations doit être suffisamment grand pour permettre d'atteindre des solutions intéressantes mais augmente la durée des calculs. La température initiale et la fonction de décroissance doivent permettre de converger vers une solution mais aussi d'éviter de rester bloqué dans un minimum local.

Il existe deux paramètres liés à la recherche d'échantillons admissibles :

- La probabilité de mutation ;
- Le choix de la fonction de probabilité pour la conservation des échantillons non-admissibles à une étape de l'itération.

Expérimentalement, il apparait que ces paramètres sont moins impactant que les précédents. Le choix de la probabilité de mutation doit être suffisamment faible pour ne pas rendre l'approche trop instable, il est ici fixé à 1%. Le choix de la probabilité de conservation des échantillons doit permettre de faire converger ceux-ci vers l'admissibilité sans pour autant limiter l'exploration de la parcelle aux mêmes ensembles de ceps.

Pour l'algorithme des colonies de fourmis, les paramètres principaux sont :

- Le nombre de fourmis ;
- Le nombre d'itérations ;
- La quantité de phéromones déposées ;
- L'évaporation des phéromones ;
- La fonction de probabilité pour le choix des arêtes à partir de la visibilité et la quantité de phéromone.

Comme pour le recuit, le choix du nombre de fourmis et d'itérations doit assurer d'atteindre des solutions intéressantes mais ne doit pas être excessif. Cet algorithme est appelé autant de fois qu'il y a d'itérations dans le processus général. Les temps de calcul augmentent donc très vite. Les paramètres jouant sur les probabilités et les quantités de phéromones sont ajustés par analyse de sensibilité. Pour tous ces paramètres il est possible de trouver individuellement les valeurs qui peuvent permettre de tendre vers les résultats attendus. Cependant, de nombreuses interactions entre paramètres existent et il est souvent difficile de les quantifier.

L'ensemble des algorithmes présentés dans cette partie ont été implémentés dans le langage R sans nécessiter le recours à des librairies extérieures. Le recuit simulé et l'algorithme des colonies constituent en effet des méthodes faciles à implémenter.

6.4 Comparaison des approches

Les deux approches ont été testées sur 6 parcelles :

- 3 parcelles de l'INRAe Pech Rouge (Narbonne, France), présentés dans le chapitre 2 pour lesquels les NDVI sont extraits pour chaque cep à partir d'une imagerie aérienne de résolution 1 pixel = 0.25m²
- 3 parcelles de l'institut Agro – Montpellier (Villeneuve-lès-Maguelone, France), pour lesquels un indice de GLCV (*Green Leaf Cover Vegetation*) est extrait pour chaque cep.

Les deux implémentations sont testées sur chacune des parcelles pour des valeurs de n variant entre 5 et 10. Les deux approches s'arrêtent après 48h de temps CPU. Cependant plusieurs expérimentations étant réalisé en même temps sur la même machine, les temps de calcul ont probablement biaisé par les problèmes de mémoire (swap). Il s'agit donc de résultats préliminaires sur des implémentations non-optimisées. Les calculs sont réalisés sur le serveur de calcul Meso@LR (Intel(R) Xeon(R) CPU E5-2680 v4 2.4GHz).

Le tableau 6.1 présente les résultats de ces expérimentations. Pour chaque approche est donné la longueur du parcours de l'échantillon obtenu, sa valeur du critère de variance et le temps qu'il a fallu au programme pour identifier cette solution. Les distances en gras représentent les valeurs dont l'optimalité a pu être démontrée par la programmation par contraintes.

La programmation par contrainte est capable de trouver la solution optimale pour les plus petites instances ($n = 5 / 6$) dans un temps relativement limité (de l'ordre de la dizaine de minutes). Les approches de recherche opérationnelle proposent dans la plupart des cas des échantillons proches de cette optimalité mais le temps nécessaire pour arriver à ces solutions est bien plus élevé. A partir de $n = 7$, la combinatoire augmente rapidement et l'optimalité n'est plus prouvée. Les solutions proposées par les méthodes heuristiques (colonies de fourmis) apparaissent alors meilleures.

Le Tableau 6.2 compare les deux approches de programmation par contraintes, une utilisant le critère de variance, l'autre utilisant la méthode des quantiles. Dans les deux cas, l'optimalité des solutions trouvées est garantie par l'approche considérée. Les distances optimales sont alors généralement plus courtes pour l'approche basée sur le critère, qui permet une plus grande souplesse dans le choix des sites. Le temps nécessaire pour trouver la solution optimale apparaît également systématiquement plus court pour la méthode des critères. En revanche lorsqu'aucune des deux approches ne trouve la solution optimale, la méthode des quantiles trouve généralement une meilleure solution que l'approche basée sur le critère. Il semble donc falloir privilégier l'approche du critère lorsque l'objectif est de trouver la solution optimale au risque d'obtenir de moins bons résultats pour les instances les plus complexes.

Tableau 6.1 : Comparaison des approches par programmation par contraintes et recherche opérationnelle pour l'identification d'un échantillon avec un critère inférieur à 10^{-4}. Pour chaque échantillon solution, ce tableau donne la distance, le critère de variance de l'échantillon et le temps nécessaire pour obtenir cette solution. Les calculs sont stoppés après 48h, les solutions optimales apparaissent en gras.

		Prog. par contraintes			Recherche opérationnelle		
Parcelle	N	Dist. (m)	Critère	Temps(s)	Dist. (m)	Critère	Temps(s)
Pech_rouge_1	5	**118,7**	1,57E-05	1522	161,3	1,30E-05	22286
Pech_rouge_1	6	326,7	1,24E-05	1181	140,3	2,42E-05	52848
Pech_rouge_1	7	312,8	9,76E-05	32561	172,1	1,83E-06	117963
Pech_rouge_1	8	370,7	5,12E-05	49028	228,5	8,20E-05	130303
Pech_rouge_1	9	354,0	4,97E-06	80691	233,4	6,55E-05	129130
Pech_rouge_1	10	395,2	3,45E-06	10710	233,5	7,51E-05	128599
Pech_rouge_2	5	207,9	1,07E-05	36089	235,8	1,62E-05	22061
Pech_rouge_2	6	288,9	1,25E-05	104565	241,0	8,02E-05	9701
Pech_rouge_2	7	338,6	6,42E-05	4911	250,4	8,73E-05	129443
Pech_rouge_2	8	439,8	1,18E-05	160370	250,6	8,30E-05	41814
Pech_rouge_2	9	446,5	4,41E-07	72503	265,5	3,30E-06	161568
Pech_rouge_2	10	424,0	4,46E-08	155404	290,3	4,47E-05	17064
Pech_rouge_3	5	**129,0**	1,32E-05	1166	138,6	7,56E-05	38842
Pech_rouge_3	6	368,8	1,12E-05	191	158,7	2,40E-05	53108
Pech_rouge_3	7	352,3	8,07E-05	17383	214,3	9,84E-06	171192
Pech_rouge_3	8	418,4	1,63E-06	104301	261,8	5,48E-07	76143
Pech_rouge_3	9	393,5	1,39E-05	73544	265,5	7,81E-05	93730
Pech_rouge_3	10	435,1	7,95E-06	19574	265,6	2,28E-05	73633
Arnel	5	**393,1**	4,00E-05	441	412,7	7,82E-05	67312
Arnel	6	**438,7**	1,11E-05	81287	439,9	6,51E-05	141546
Arnel	7	799,1	2,47E-05	147610	472,9	1,61E-05	104008
Arnel	8	948,3	1,81E-05	21014	512,7	1,21E-07	13388
Arnel	9	1014,3	3,59E-05	71339	534,3	1,49E-05	98650
Arnel	10	1177,9	4,07E-05	9982	558,6	7,70E-05	137526
Estagnol	5	**218,9**	2,77E-05	567	219,1	5,78E-07	4473
Estagnol	6	**228,7**	9,87E-06	28185	229,0	5,51E-05	2004
Estagnol	7	483,5	2,25E-05	15505	281,7	2,45E-05	15567
Estagnol	8	528,9	4,14E-05	1403	301,2	1,52E-05	17302
Estagnol	9	609,7	4,22E-05	80137	304,5	7,62E-05	10231
Estagnol	10	561,7	4,51E-05	135277	340,0	2,66E-05	88050
Larzat	5	**271,0**	4,69E-05	4418	606,4	1,46E-07	2582
Larzat	6	**320,9**	1,28E-05	95285	419,5	6,17E-06	13086
Larzat	7	454,9	3,61E-05	169569	431,0	4,67E-05	59926
Larzat	8	528,8	2,57E-05	42684	485,7	1,05E-05	87444
Larzat	9	1391,7	9,20E-05	108267	722,9	9,70E-05	118283
Larzat	10	1761,2	1,48E-05	31235	783,0	4,76E-05	76166

Tableau 6.2 : Comparaison de deux approches utilisant la programmation par contraintes : l'une basée sur le critère de variance (devant être inférieur à 10^{-4},l'autre basée sur la méthode des quantiles Pour chaque échantillon solution, ce tableau donne la distance, le critère de variance de l'échantillon et le temps nécessaire pour obtenir cette solution. Les calculs sont stoppés après 48h, les solutions optimales apparaissent en gras.

		Prog. par contr. (critères)			Prog. par contr. (quantiles)		
Parcelle	N	Dist. (m)	Critère	Temps(s)	Dist. (m)	Critère	Temps(s)
Pech_rouge_1	5	**118,7**	1,57E-05	1522	**124,3**	1,53E-02	3986
Pech_rouge_1	6	326,7	1,24E-05	1181	151,3	5,29E-03	37608
Pech_rouge_1	7	312,8	9,76E-05	32561	180,3	1,18E-03	19589
Pech_rouge_1	8	370,7	5,12E-05	49028	205,9	3,93E-03	126777
Pech_rouge_1	9	354,0	4,97E-06	80691	239,5	2,11E-03	42204
Pech_rouge_1	10	395,2	3,45E-06	10710	322,9	6,20E-04	21556
Pech_rouge_2	5	207,9	1,07E-05	36089	207,4	6,02E-03	81091
Pech_rouge_2	6	288,9	1,25E-05	104565	254,6	2,07E-03	7641
Pech_rouge_2	7	338,6	6,42E-05	4911	255,5	8,74E-03	45811
Pech_rouge_2	8	439,8	1,18E-05	160370	261,1	7,05E-04	90497
Pech_rouge_2	9	446,5	4,41E-07	72503	317,2	2,49E-03	10545
Pech_rouge_2	10	424,0	4,46E-08	155404	360,3	2,09E-03	39108
Pech_rouge_3	5	**129,0**	1,32E-05	1166	**133,8**	2,25E-04	2031
Pech_rouge_3	6	368,8	1,12E-05	191	158,8	2,86E-03	156161
Pech_rouge_3	7	352,3	8,07E-05	17383	195,5	3,10E-04	149281
Pech_rouge_3	8	418,4	1,63E-06	104301	222,7	1,93E-04	86271
Pech_rouge_3	9	393,5	1,39E-05	73544	330,5	9,09E-04	159017
Pech_rouge_3	10	435,1	7,95E-06	19574	306,0	5,98E-04	53463
arnel	5	**393,1**	4,00E-05	441	**398,1**	8,67E-02	1544
arnel	6	**438,7**	1,11E-05	81287	**452,9**	7,09E-02	1767
arnel	7	799,1	2,47E-05	147610	488,2	3,75E-02	115245
arnel	8	948,3	1,81E-05	21014	573,4	1,88E-02	133674
arnel	9	1014,3	3,59E-05	71339	645,0	1,86E-02	10766
arnel	10	1177,9	4,07E-05	9982	819,8	2,34E-02	147220
estagnol	5	**218,9**	2,77E-05	567	**232,6**	2,79E-01	1765
estagnol	6	**228,7**	9,87E-06	28185	**237,3**	2,95E-01	7854
estagnol	7	483,5	2,25E-05	15505	328,8	2,66E-01	129341
estagnol	8	528,9	4,14E-05	1403	401,4	4,09E-01	2115
estagnol	9	609,7	4,22E-05	80137	474,5	6,42E-02	23129
estagnol	10	561,7	4,51E-05	135277	523,0	6,69E-02	25526
larzat	5	**271,0**	4,69E-05	4418	**293,3**	8,12E-02	940
larzat	6	**320,9**	1,28E-05	95285	**366,5**	7,86E-03	124978
larzat	7	454,9	3,61E-05	169569	454,2	2,25E-02	129039
larzat	8	528,8	2,57E-05	42684	505,0	2,14E-02	109181
larzat	9	1391,7	9,20E-05	108267	580,9	9,78E-03	79786
larzat	10	1761,2	1,48E-05	31235	774,9	1,52E-02	17552

6.5 Conclusions

Ce chapitre présente la résolution du problème d'optimisation de l'échantillonnage. Le défi associé à cette résolution d'un problème combinatoire réside dans l'utilisation simultanée de méthodes statistiques et informatiques. Deux approches complémentaires sont ainsi proposées. L'une propose de trouver la solution optimale dans le cas d'une combinatoire faible à moyenne grâce à la programmation par contraintes, l'autre propose de trouver une solution intéressante dans un temps plus court sans en garantir l'optimalité.

Bien que préliminaires, ces expérimentations semblent montrer que la programmation par contraintes permet d'obtenir la solution optimale sur les petites instances. Pour ce qui est des instances les plus combinatoires, les outils de la recherche opérationnelle semblent proposer de meilleurs résultats. Les Tableaux 6.1 et 6.2 ne sont cependant le résultat que d'une seule expérimentation, des répétitions de ces expérimentations doivent confirmer ces résultats afin de prendre en compte la variabilité des méthodes heuristiques. L'obtention d'une méthode qui soit applicable dans un contexte de production nécessiterait des travaux supplémentaires sur ces approches afin d'optimiser les temps de calculs et la gestion de la mémoire.

Une approche hybride utilisant une approche à la combinatoire du problème pourrait constituer une solution opérationnelle. Cet aspect n'est naturellement pas étudié dans la thèse mais soulève des questions intéressantes. Ce chapitre démontre la faisabilité de la résolution du problème posé.

Le chapitre suivant présente les types de parcours d'échantillonnage obtenus avec la résolution du problème d'échantillonnage et les compare aux pratiques actuelles en viticulture. Dans cette note technique publiée dans la revue Oeno-One, l'approche mise en place repose sur la méthode des quantiles à l'aide de la programmation par contraintes.

Chapitre 7 : Is the optimal strategy to decide on sampling route always the same from field to field using the same sampling method to estimate yield?

B. OGER[1-2]*, C. LAURENT[1-3], P. VISMARA[2-4] and B. TISSEYRE[1]

[1] *ITAP, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[2] *MISTEA, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[3] *Fruition sciences, France*
[4] *LIRMM, Univ. Montpellier, CNRS, France*

7.1 Abstract

Aim: This short communication aims at providing insights to verify whether common yield sampling protocols (i.e. one round trip within the fields over two representative rows) are optimal whatever the considered fields. In addition, it aims to show how factors like the spatial organisation of the within-field yield variability, the length of the rows, the erratic variance, etc. may affect the optimal sampling route and the error of the yield estimation.

Methods and Material: A new algorithm based on constraint programming and stochastic approaches was used to provide optimal sampling routes for vineyards. This algorithm guarantees the representativeness of the measurement sites and a minimization of the walking distance. Practical constraints (trellised structure, starting point, etc.) are considered by the algorithm to optimise the walking distance and the resulting sampling route. The algorithm has been applied to 60 simulated vineyards with known yield variability. Characteristics like yield spatial structure, row length and proportion of erratic variance were controlled during the simulation process and were used to study how they affect the optimal sampling route derived from the algorithm.

Results: The row length as well as the spatial organization of the within-field yield variability are the main factors that determine the optimality of a sampling route. Spatial organisation of the yield happens to have a strong incidence; fields with small yield patterns (Range of the semi-variogram = 25 m) showed a yield estimation error of less than 2 % with an optimal sampling route of three minutes with 7 sampling sites, whereas it takes more than 5 minutes (with 9 sampling sites) to achieve the same estimation error for fields with larger spatial patterns (range > 50 m). Results also highlight the relevance of original sampling routes which intend to sample only the beginnings of rows or mixed approaches based on a round trip in two inter-rows and complementary samples on the beginnings of one or more rows.

Conclusions: This study shows that an optimal sampling route strongly depends on the field characteristics. The optimal sampling route should therefore be tailored to each field. This approach is a first step which shows how this methodology could be used to identify other factors of influence. It could also apply to real fields to optimise other logistic operations in viticulture.

Significance and Impact of the Study: This short communication demonstrates the necessity to tailor sampling strategy to characteristics of each field to provide both an optimised sampling route (minimum walking distance with minimum samples) and the best possible estimate. It also proposes an original approach based on field simulations and an optimal sampling route generation algorithm. This approach makes it possible to produce new insights (and also to validate empirical practices) that

can help the wine industry to better manage the logistics at harvest. This paper also gives considerations when it comes to the choice of a sampling route for a given field.

7.2 Introduction

Precise knowledge of field yields is critical for the wine industry, mainly for the logistical organisation of the harvest among other reasons (Clingeleffer et al., 2001). Field yield estimation is often carried out by sampling. Observations of yield components are then made on a limited number of vines (Carrillo et al., 2016; Arnó et al. 2017; Wolpert and Vilas, 1992) and the mean value is generally used to provide an estimation of the field yield. In order to have a relevant estimation of the final yield, sampling is often carried out a few days before harvest, at a critical period in terms of work load. As a result, the implementation of a yield estimation protocol is often the result of a balance between the accepted error on the estimation and the time required to carry out the observations (Wolpert and Vilas, 1992). Vine fields (even small) presents a high yield variance (Taylor et al., 2005); average coefficient of variation was found to be around 40 %. This variability is mostly explained by the spatial variation of environmental factors (soil, water availability, fertility, etc.) but also to other biotic factors (disease, weed or inter-vine competition, etc.).

This within-field variability affects the quality of estimates resulting from sampling. Indeed, when estimating yield by sampling, the error of an estimation (and resulting confidence associated with the estimate) is a function of the number 'n' of observations and the variance of the sample (observed variables). For a given field (with a given yield variance), the higher the 'n', the more confident the estimate is, but the longer the time required to carry out the sampling. Note the sampling time may present high variations depending on the location of the observations over the field. Indeed, the sampling time is directly related to 'n', but also to the time needed to travel from one observation site to another. Optimising sampling time therefore requires to optimise both the number 'n' of observations according to the field variability and the location of the observation sites to limit the travel time.

In the scientific literature, there are very few papers which have focused on optimizing the location of sampling sites for yield estimation in viticulture. Most of the studies focused on: i) the number 'N' of observations to be considered in order to reach a reliable estimation and to minimise the error of estimation (Wolpert and Vilas, 1992, Carrillo et al., 2016), ii) the type of observations and sensing systems to limit measurement time and/or measurement errors (Diago et al. 2012; Reis et al. 2012; Nuske et al. 2011; Serrano et al. 2005; Dunn and Martin 2004), iii) the optimization of the representativeness of the observation sites by targeting samples based on an auxiliary variable (i. e. vegetative index) available with a high spatial resolution (Wulfsohn et al. 2012; Meyers et al. 2011, Carrillo et al., 2016). More recently, for grape maturation, Meyers et al. (2020) proposed a refinement of such approaches by considering an empirical reasoning to limit the distance between sampling points based on successive values of vegetative index in a same row. Although interesting, the approach proposed by Meyers et al (2020) does not consider any optimisation procedure and can't guarantee that the proposed solution is among the best possible.

As a result, existing studies rarely take into account the two contradictory components leading to an optimal sampling: the optimisation of the sampling effort (time which includes the measurement time and the travel time/distance) and the minimisation of estimation error. For the wine industry, these two components are very important to produce the best possible estimation in the shortest possible time. Without reliable references, the sampling protocols used are often based on rules of thumb and the same protocol is always applied whatever the field, *i.e.* two representative rows are chosen

corresponding to one round trip within the field and observations are more or less randomly carried out along these rows (Rousseau Jacques, pers. Communication) or sometimes according to a grid previously defined by an expert. Note that the use of high resolution spatial data like remote sensing or soil mapping to consider more sophisticated sampling process remains rare, at least in France since less than 2% of the vineyard area is benefiting from this type of service in France (Lachia et al., 2019). Whatever the protocol, the underlying rules of thumb used by the wine industry to perform yield sampling remain difficult to justify rigorously.

A recent paper (Oger et al., 2019) proposed a new sampling approach which combines stochastic methods with constraint optimization to produce optimal sampling routes in viticulture when vines are planted in rows. The Oger et al. (2019) approach is interesting because it simultaneously takes into account the stochastic nature of the spatial variable (erratic and spatial auto-correlation components) and the sampling time by integrating constraints which aim at minimizing the practitioner's travel time. The approach developed by Oger et al. (2019) thus makes it possible to select the best possible observation sites by simultaneously ensuring the representativeness of the measurement in the attribute space (values) by considering the distribution of a high spatial resolution auxiliary data and the optimality of the path to go from one observation site to another. Assuming that the yield is fully known and a value is available for each within-field site, this method is interesting because it allows to reverse its application in order to verify what the optimal sampling route according to the characteristics of the field would be. The application of such an approach to fields with known yield values is interesting because it allows: (i) to check whether empirical sampling protocols like the two-row round trip are relevant and if they apply properly to all the fields, (ii) to explore whether original and non-trivial routes can emerge from the application of this algorithm, and (iii) to see whether specific field characteristics promote particular sampling procedures. The objective of this study is therefore to apply the methodology of Oger et al. (2019) on theoretical yield data with known features in order to study how factors like the spatial organisation of the yield, the length of the rows, the erratic variance, etc. may affect the optimal sampling route and the error of the yield estimation. This work remains theoretical in the sense that it requires prior knowledge of the yield at any point in the field (which is not the case in reality). However, the aim of this study is to verify whether the same optimal sampling route patterns apply for all the fields or, on the contrary, whether specific sampling routes tailored to each field (or type of field) should be considered. The paper will briefly present the approach as well as the theoretical yield data and their characteristics. It will then present the results with a discussion that will focus on practical issues.

7.3 Materials and Methods

7.3.1 Data

Theoretical yield data were generated through a simulation process. This simulation process, described in Oger et al (2020), generates spatialized data by summing Gaussian fields to non-spatialized residual noise (erratic data). By setting the semi-variogram of Gaussian field and the noise proportion it was thus possible to control parameters of the resulting theoretical data. This paper focused on the influence of three parameters on the optimal sampling: i) The range, for the Gaussian field semi-variogram. This corresponds to the autocorrelation distance of the theoretical yield data. The range defines the minimum distance (in meters) that must separate two sites for them to be considered as spatially independent. It defines the average size of the yield spatial patterns within a field, ii) The nugget effect, which is the proportion of erratic (non-spatialized) variance of the theoretical yield data.

This measurement is expressed as a percentage of the total variance. iii) The row length in relation to field width.

For the simulation processes, the magnitude of variation of values for the range and the nugget effect were determined from within field yield observations obtained from yield monitoring systems in precision viticulture (Taylor et al 2005; Bramley et al. 2019). For row length, simple rectangular structures with areas of 1 hectare were tested. Tableau 7.1 summarizes the values of the parameters tested in this article.

Theoretical fields were generated by varying only one parameter at a time, with the other two parameters taking their default values. The initial resolution is 1 pixel/m². Yield values are then extracted on the rows assuming a trellised structure with a 2.5 m distance between rows and 1 m between vines on the row (4000 vine plants/ha). Simulations were run with a Gaussian yield distribution with an average yield around 1000 g/vine and a coefficient of variation at 30% For each combination, 10 different fields were simulated. The final theoretical dataset consists of 60 (6×10) simulated fields.

Tableau 7.1 : Parameter values for the generation of theoretical fields (default values in bold font).

Theoretical field yield parameters	**Values**		
Range (m)	25	**50**	75
Nugget effect (%)	**20**		50
Row length (m) × field width (m)	50 × 200	**100 × 100**	200 × 50

7.3.2 Sampling Route Optimisation

For each field, the optimal sampling route was obtained by applying the approach described in Oger *et al.* (2019). This approach uses constraint programming principles and stochastic methods to find the best sampling route according to defined constraints. Without going into too much computational detail, a first constraint ensures that the n selected measurement sites are separated by a minimum Euclidean distance in order to avoid autocorrelation and to make sure observed yield values are independent. This minimum Euclidean distance is defined by half the range of yield data (refer to previous section). For each field, the second constraint aims at ensuring that the N measurement sites are representative of the yield value distribution of the field, one measurement site is selected in each of the intervals defined by the n yield quantiles as proposed by Carrillo et al., (2016). Finally, the selected sampling route must be optimal in terms of walking distance. Measurement sites are selected to fit the two first constraints while their position and the order in which they are visited must minimise the walking distance. Optimisation was performed using a solver, a software program which considered possible combinations to select the best one. While exploring possible combinations, the approach seeks to find a better solution than the best one found so far. For simple cases with small values of N, the real optimum is found in a short time. In the most complex cases, computations are stopped after ten hours to ensure the solution is close enough to the real optimum. This time is generally sufficient to find a value close enough to the optimum or the optimum itself but without being able to demonstrate it. The core of the sampling approach was written in Java using the Choco solver (Prud'homme et al. 2016). Computation were made on a Linux server with Intel(R) Xeon(R) CPU X5690 3.47GHz.

Distances were expressed as walking time (min.). Walking times do not consider additional constraints specific to a given field that could alter the walking speed (grass, slope, soil surface conditions, etc.). They only take into account vineyard specificities associated to the trellised structure. It is not possible to move between two rows while being in the field. Going from one inter-row to another implies having

to reach one of the field edges. Each measurement site can be accessed from two different inter-rows. This distance also takes into account a starting point where the sampling route must begin and end. It is positioned in the southwest corner of each field. The distance optimized by the solver corresponds to this walking distance that passes through each measurement site and returns to the starting point. This promotes the choice of measurement sites close to the starting point. Common starting points enable simple comparison of the sampling routes obtained.

7.3.3 Sampling route characterisation

In order to clarify the presentation of the results, two types of sampling routes were considered. The first one corresponded to what is assumed to be most commonly performed by practitioners; this consists in an empirical sampling protocol where measurements are carried out following one round trip within the fields across two, or more, representative rows. Rows are therefore walked from one end to the other, forming a sampling route joining the two sides of the field. This type of route is called thereafter *row-based sampling route* (RBSR). The second type of sampling route never reaches both sides of the field i.e. no row is walked entirely. This type of sampling route corresponds in reality to a large diversity of cases, but a common feature is to focus on the field edge close to the starting point. All the sampling routes presenting these features were considered as *edge-based sampling rout*e (EBSR).

Sampling routes obtained with the solver were characterised using three criteria: i) The type of sampling route: RBSR or EBSR. ii) The walking time required to get from one observation site to another, regardless of the protocol chosen to carry out the measurements and the time associated with these measurements. The time required to make observations (number of clusters, average cluster weight, etc.) at a sampling site may vary depending on the protocol used. However, it was assumed in this work that, for a given situation, the measurement protocol was the same for each sampling site. As a result, for the same number of observation sites, the sampling time was only influenced by the travel distance between the observation sites. The walking time therefore only depends on the distance to be covered and the walking speed of the practitioner, which is assumed here to be constant at 0.9 m/s. iii) The estimation error, corresponds to the difference between the value predicted from measurement sites along the sampling route and the actual average yield of the field. The predicted value ($\hat{\bar{Y}}$) is constructed as the average of the n yield observations made during sampling. The actual yield value ($\bar{Y}$) corresponds to the average of all the simulated yield values of the field. The calculation of the estimation error, expressed as a percentage, is defined by Eq. 7.1.

$$Estimation\ Error\ (\%) = \left| \frac{\hat{\bar{Y}} - \bar{Y}}{\bar{Y}} \right| \times 100 \qquad Eq.7.1$$

7.4 Results

Figure 7.1.A, 1.B and 1.C shows results of optimal sampling routes expressed as estimation errors and walking times for the different field characteristics. Figure 7.1.D shows results obtained with a simple random sampling. All the curves share the same logical trends; the estimation errors decrease with an increase in the number N of samples. However, improving the quality of the estimation has a cost since the sampling effort estimated by the 'walking time' increases with the number of measurements. A comparison between Figure 7.1.A and Figure 7.1.D shows the value of the optimal sampling approach as proposed in this study compared to a simple random sampling approach. Sampling optimisation simultaneously improves the estimation error by 5 % to 9% and reduces the running time by half for

the examples considered. Only results obtained for simulated fields with different range were presented (Figure 7.1.D) for random sampling, but very similar results (results not shown) were obtained for the other simulated fields (row length, nugget effect).

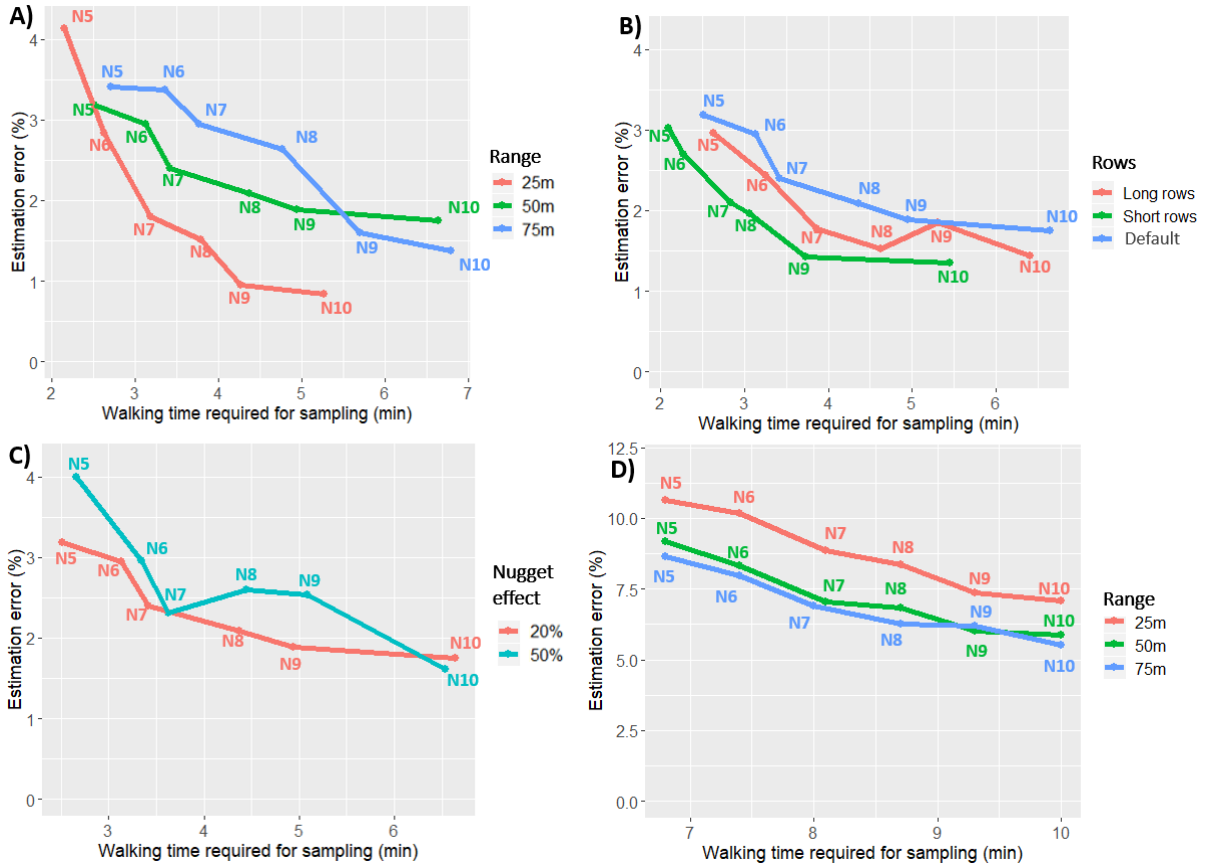

Figure 7.1 : Average estimation error and walking times depending on field characteristics: a) the field range, b) row length, c) percentage of random variability (nugget effect). Results are the mean value over ten simulations. Each curve is made of 6 points corresponding to sampling routes with $n = \{5,6,7,8,9,10\}$ sampling sites. d) gives the same result as a) but for random sampling and results are the mean value over ten simulations and 100 repetitions per simulation.

Figure 7.1 shows that the characteristics of the fields do not affect the optimal sampling route in the same way. The range (Figure 7.1.A) and, to a lesser extent the row length (Figure 7.1.B), significantly affect the optimal sampling route, while the proportion of erratic variance in the total yield variance of the field (nugget effect) has a small effect on the optimal sampling route (Figure 7.1.C). For clarity, Figure 7.1 does not show the variability resulting from the ten simulations, in average standard deviation of the results is of 1.7 % and 0.6 minutes respectively for the error and the walking time.

Regarding the range (Figure 7.1.A.), fields with lower ranges (25m) show lower estimation errors for a given 'walking time'. On average, for a range of 25m, it is possible to achieve an estimation error less than 2 % with an optimal three minutes sampling route with 7 sampling sites, whereas it takes more than 5 minutes (with 9 sampling sites) to achieve the same estimation error for fields with larger ranges (50 m and 75 m). In general, the lower the range, the shorter the sampling route and the walking time. Focusing on, the length of the rows (Figure 7.1.B), it is also a factor which affects an optimal sampling route. For short rows, lower estimation errors are achieved with less sampling effort.

Effect of nugget effect (Figure 7.1.C) is less obvious although, larger nugget effects (50%) are associated with slightly larger estimation errors compared to fields with a low nugget effect (20%).

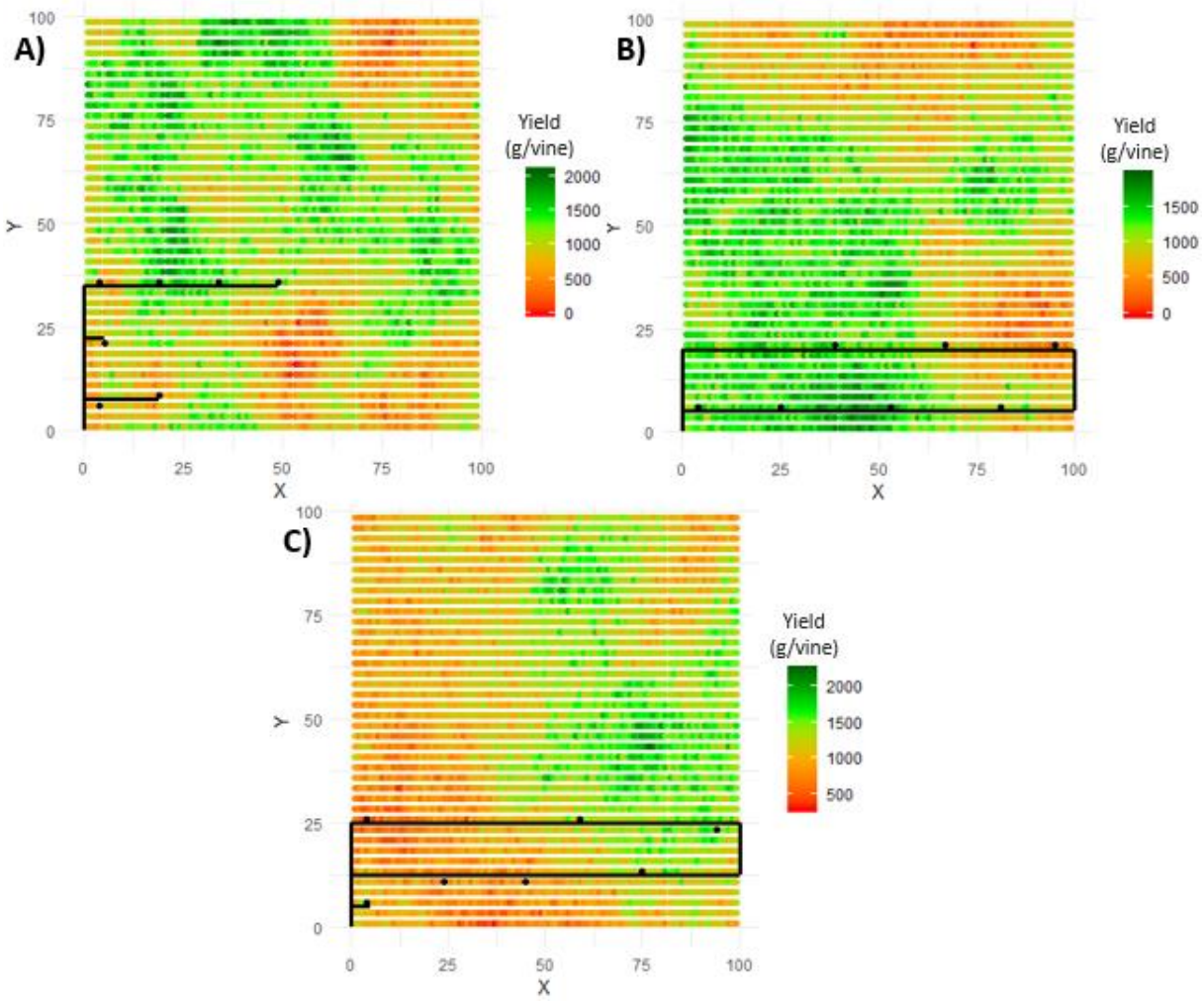

Figure 7.2 : Illustration of sampling routes for three typical fields with different range with $n = 7$
Field A: Range = 25m, Nugget effect = 20%, Row length = 100m
Field B: Range = 50m, Nugget effect = 20%, Row length = 100m
Field C: Range = 75m, Nugget effect = 20%, Row length = 100m

Focusing on the range effect, Figure 7.2 shows examples of sampling routes for three fields with different range values. The three examples share some common features; sampling routes are optimized from the starting point located in the southwest corner of the field (coordinates X= 0, Y=0). It is clear that the sampling points (and the resulting sampling route) intend to minimize the distance to this starting point for each field. Figure 7.2 also shows the two types of sampling routes described previously (EBSR or RBSR). The field with the shorter range is associated with an edge-based sampling route (EBSR), while the fields with longer ranges (50m and 75m) are associated with a row-based sampling route (RBSR). In this example, for the same number of sampling sites, the optimal route changes with the range.

Figure 7.3 shows however that for large ranges, EBSR may also be promoted as an optimal sampling route. In this case, a large range (in comparison to the dimension of the field) affects the spatial variability of yield which tends to follow a trend (gradient). In practice, this type of spatial distribution may be observed when the yield is driven by an isotropic factor such as the slope, soil depth gradient, water access, etc. The optimal sampling route is in this case dependent on the relative direction of the rows with the yield gradient. When the yield gradient and the rows present more or less the same direction (Figure 7.3A), RBSR is promoted. Conversely, when the gradient is perpendicular to the row direction (Figure 7.3B), EBSR is promoted by the algorithm.

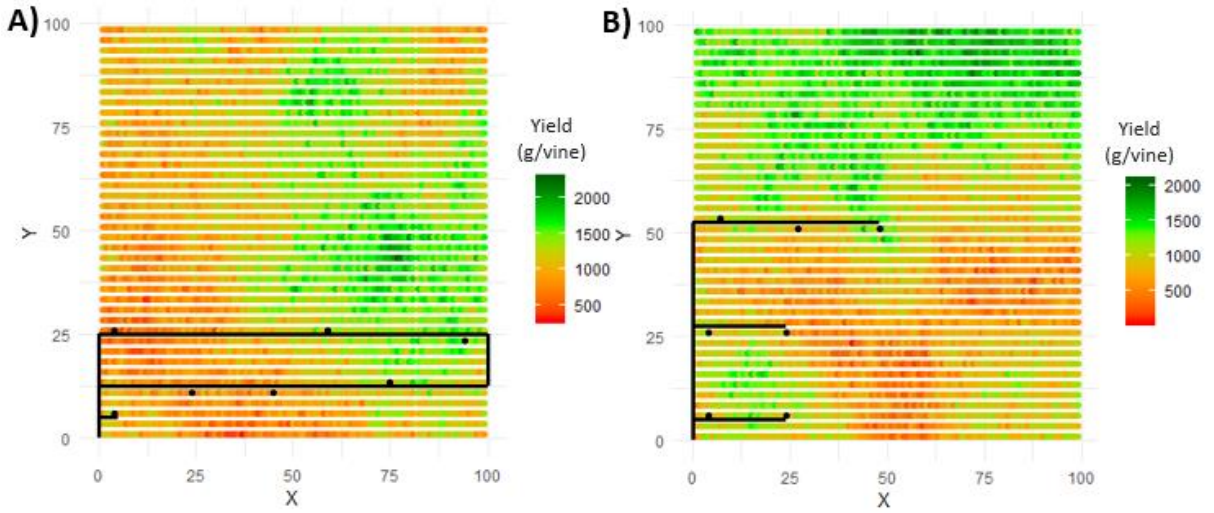

Figure 7.3 : Illustration of sampling routes for two high range fields with opposite gradient orientation and $n = 7$
Field A & B: Range = 75m, Nugget effect = 20%, Row length = 100m

Figure 7.4 shows the effect of row length on optimal sampling routes across three examples. Fields with short rows are associated with RBSR even with a limited number of sampling sites ($n = 5$) (Figure 4.A). Conversely, long rows promote EBSR where entire rows are never explored (Figure 7.4.B and C).

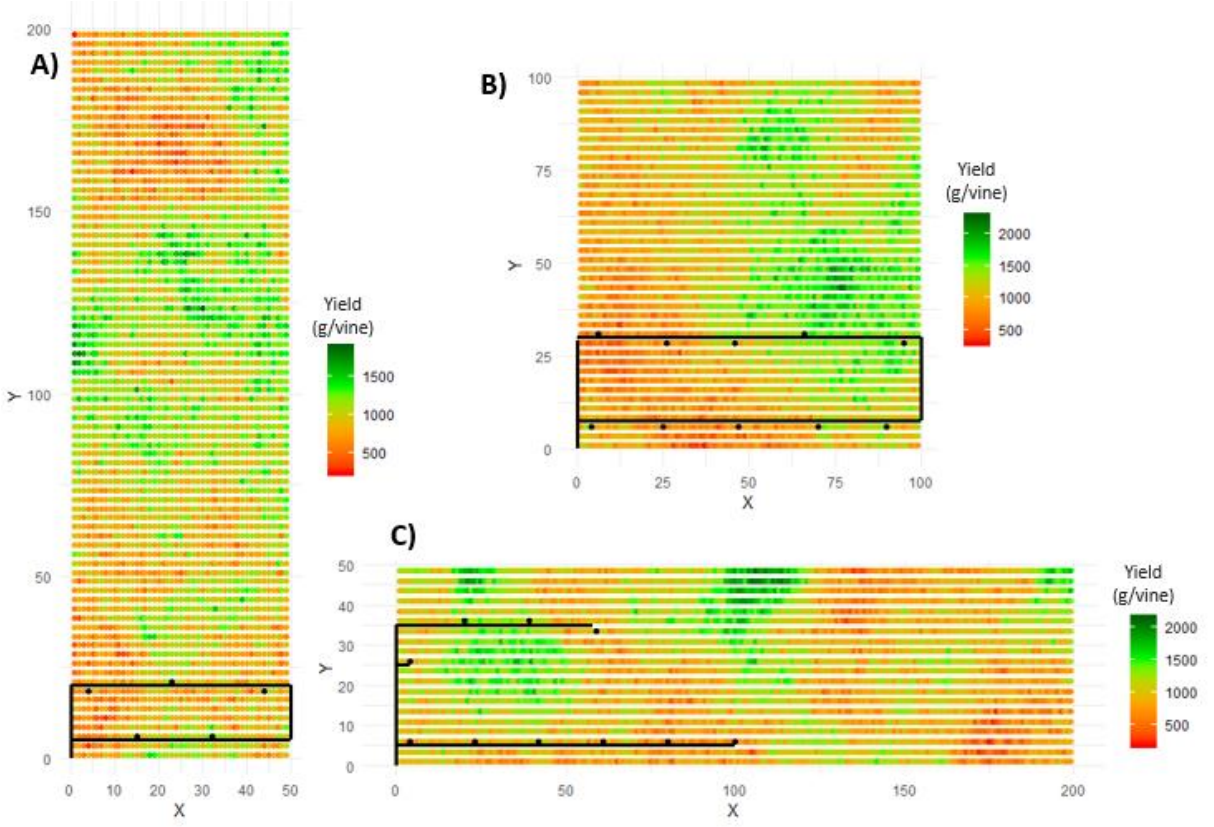

Figure 7.4 : Illustration of sampling routes for three different row length
Field A: Range = 50m, Nugget effect = 20%, Row length = 50m, $n = 5$
Field B: Range = 50m, Nugget effect = 20%, Row length = 100m, $n = 10$
Field C: Range = 50m, Nugget effect = 20%, Row length = 200m, $n = 10$

For different field parameters, Figure 7.5 gives the proportion of sampling strategies corresponding to RBSR against EBSR in function of the number of sampling sites. Each point of the figure corresponds to the average results over ten simulated fields.

Figure 7.5.A shows clearly that for fields with short ranges, the optimal sampling route is an EBSR in a large majority. As already seen before, the range has a significant effect on the choice of the optimal sampling strategy and this result is confirmed here over several fields. However, for fields with ranges of 50 m and 75 m, the effect is lessened and the proportion of full row sampling routes (RBSR) reaches a limit; the proportion curve associated with high ranges (75m) never reaches 100 % of RBSR. This result is explained by simulated fields whose yield gradient is more or less perpendicular to the row direction which promotes EBSR over RBSR (Figure 7.3).

Figure 7.5.B also shows clearly the incidence of the length of the row on the best possible sampling route. Long rows always promote EBSR while short row fields always promote EBSR. This result verifies that of Figure 7.4: when the rows get longer, the optimal sampling strategy always avoids going all along the rows. Exactly the opposite is true for short rows, which is why EBSR is systematically proposed in this case.

Finally, Figure 7.5.C shows that a higher proportion of erratic variance (nugget effect) tends to promote EBSR.

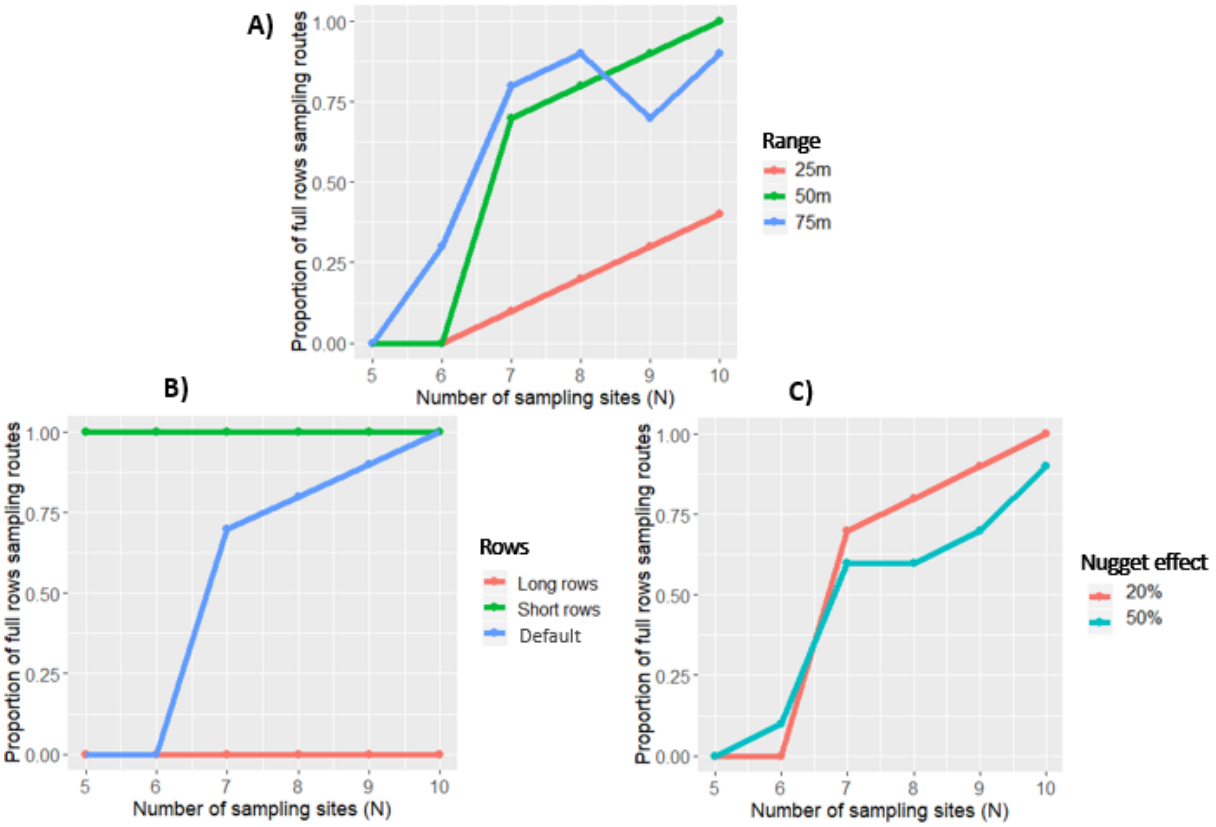

Figure 7.5 : Proportion of RBSR sampling strategies sampling route depending on the number of sampling sites (n) and field type (range, nugget effect and row length of the field).

7.5 Discussion

In the wine industry, a tendency to adopt the same sampling route for all fields is commonly encountered. However, based on a posteriori knowledge of yield distribution, results exposed in this paper show that the optimal strategy to design a sampling route for grape yield estimation may vary from one field to another in function of field characteristics. The optimal route sampling seeks to minimize the effort to find sites that are representative of the distribution of yield values. Logically, lower range of yield reduces the minimum distance to be covered to find two spatially independent sites. Therefore, low ranges make it possible to find a higher variability of yield values in the direct

vicinity of the starting point which explains why EBSR is promoted in this case. This also explains why the travel distance decreases with the yield range (Figure 7.1), EBSR being generally shorter as it does not require to travel twice the length of the rows to find relevant observation sites. The extreme case would be a field with no spatial autocorrelation of yield values (*i.e.*, yield values are perfectly random with no range), in which choosing n independent sampling sites might result in selecting n contiguous vines on the same row. For large range yield, when the yield gradient and the rows present more or less the same direction (Figure 7.3.A), RBSR is logically promoted. Conversely, when the gradient is perpendicular to the row direction (Figure 7.3.B), EBSR is promoted by the algorithm. This is consistent considering that RBSR allows a larger diversity of yield values to be explored more quickly when the variability is organised along the rows, whereas EBSR is more efficient to explore the diversity of yield values by travelling through different rows when yield gradient is perpendicular to the rows. Similarly, short rows always provide more flexibility to find short sampling routes. Indeed, they bring the possibility to access a large diversity of yield values quickly in a limited time (Figure 7.4.A) which always promote RBSR. For long rows, RBSR becomes time-consuming with no added information in exploring entire rows (Figure 7.4.C) which justify EBSR in this case. For the same reason, although this conclusion is not that consistent with the results, an increasing nugget effect may result in more heterogeneous yield values in the surroundings of the starting point, which logically promotes EBSR.

Note that operational hybrid sampling routes do exist for fields corresponding to more complex configuration. In this case, sampling route is largely based on RBSR which consists in a one-way round trip across two rows with one or more measurement sites coming from a third incompletely covered row added (Figure 7.3.A).

It should be kept in mind that the result of this study are based on simulated data which represent a simplified version of reality. The errors of estimation exposed here are not indicative of what can be found in practice, the context here is a purely theoretical framework where the spatial distribution of the yield is fully known. For example, it was assumed that for each measurement site, the yield was fully known, as if all bunches of the plant had been weighed. Such a destructive approach is not realistic in a commercial situation because of measurement time and yield loss. In practice, the estimation of the yield on a site is itself the result of a sampling of one or two bunches chosen and weighed by the operator. The result of this process is an error in estimating the yield at each site and a resulting error in estimating the average yield of the field which is necessarily higher than that reported in this work. Uncertainty in the representativeness of the sampling sites and measurement errors are therefore neglected.

Considerations discussed in this study are based on simple field characteristics. This simplified framework enables to identify the impacts of different parameters affecting sampling route. However, the characteristics of a real field are often more complex. For example, rows can have irregular length, fields can have irregular shapes, different sizes etc. Other elements can also affect travel time such as slopes or the presence of discontinuity in the row structure allowing the practitioner to pass from one row to the other without having to walk all along it. Logistical issues may also count in the sampling route design. The intention of this paper is therefore not to give settled values to be respected but rather guidelines to consider in order to optimize yield sampling at a lower cost, and effort when information about the yield spatial structure is available. It is thus to be noted that simple and quick field observations such as the row length can be used to instruct the choice towards an EBSR or RBSR strategy. The row length is simple and available information that can be considered without additional cost. This can moreover be achieved without interfering with the decision on the trade-off between estimation error and sampling time, which is left to the practitioner's discretion. The starting point correspond here to the fixed entry point for a given field. Its position has an influence on the distance

to be covered to reach certain sites. When possible, adjusting its position could reduce the total sampling time. Note that the total sampling time also depends on the measurement time, which is not discussed here as this work focuses on minimising walking time. A proper sampling strategy should consider both walking time minimisation and suited measurement protocol.

Thus, based on the study of yield spatial structure, results shed light on some generic considerations when sampling for grape yield estimation. However, yield spatial structure is generally not known before sampling. Ancillary data i.e. data that are correlated to yield can then be used for this purpose. These data are often chosen because they are readily available, at higher resolution and at lower cost than yield data. Vegetation indices such as NDVI (Carillo et al. 2016) measured by satellite, UAV or aerial imagery and historical yield data (Araya-Alman et al. 2017) are examples of auxiliary data which are already being considered for grape yield estimation. However, as the correlations between yield and ancillary data are specific to each field (Carillo et al. 2016), the use of these data must be made on the basis of field knowledge and local calibration as far as possible. Temporality must be considered as it might affect the correlation between variables. When fully known, an ancillary can then be used to directly drive measurement site selection according to the same considerations exposed in this paper and thus help in yield estimation before harvest.

The results obtained in this study are dependent on the sampling strategy used. In this case, this later aimed at selecting measurement sites that are representative of the yield distribution. The choice of this approach may explain why the proposed optimal routes are strongly influenced by the spatial structure of the yield and its organisation with respect to row orientation. However, most targeted sampling methods aim to consider the distribution of the variable to be estimated and may well lead to similar results. This study does not allow us to demonstrate this, however the proposed methodology may well be used to evaluate sampling methods by simultaneously taking into account: the quality of the estimate made and the sampling effort.

Finally, these considerations on optimal routes for yield sampling may be applied to other variables of interest such as fruit maturation (Meyers et al. 2020), Brix degree (Kasimatis et al. 1985) or water status (Herrero-Langreo et al. 2017). This study could also be extended to other crops associated with distance constrained by a trellised structure.

7.6 Conclusion

This work shows that to be optimal, a sampling route must be tailored to the characteristics of the field. The row length as well as the spatial organization of the within-field yield variability are factors that determine the optimality of a sampling route. This work opens up interesting perspectives. Indeed, the approach could be used to identify whether other factors affect the optimal definition of a sampling route (e.g. by taking into account the slope and the resulting effort). Thus, this work could well be applied to real cases and propose optimal sampling routes by taking into account the actual length of the rows, the actual starting points corresponding to field access, the expected spatial organisation of the yield data based on previous knowledge (yield maps from previous years, multispectral images, soil electrical resistivity maps) etc. Beyond these practical aspects, this work also highlights the interest of spatial simulation in association with constraint optimisation, to provide insights for optimising the logistics of viticulture operations. Constraint models could be adapted to fit real case studies. In particular, similar approaches could well be used to propose the optimization of machine routes in order to respond to economical as well as environmental issues such as the reduction of fossil fuel consumption.

Chapitre 8 : Détection de valeurs aberrantes dans le cadre d'un échantillonnage en production végétale.

Ce chapitre 8 propose une première approche, mise en place au cours de la thèse pour la détection des valeurs aberrantes pour les données auxiliaires. La plupart des éléments présentés ici n'ont pas été mobilisés pour la mise en place des résultats des précédents chapitres. Toutefois, ces travaux s'inscrivent dans un processus d'amélioration des méthodes proposées en vue d'établir une approche fonctionnelle de détection des points ou sites de mesure aberrants.

8.1 Les valeurs aberrantes pour l'échantillonnage en production végétale

8.1.1 L'enjeux des valeurs aberrantes et de leur détection

Le chapitre 5 fait apparaitre l'importance de privilégier des échantillons présentant une forte variance lorsque l'inférence se base sur l'étalonnage d'un modèle linéaire. Le critère de variance, proposé pour raisonner le choix des sites de mesure, favorise en effet les sites de mesure présentant des valeurs extrêmes, dans la mesure où celles-ci contribuent à minimiser le ratio $\frac{1}{\sum_{i\in N}(X_i-\overline{X_N})^2}$. Sélectionner de tels sites de mesure représente cependant un risque, ces valeurs extrêmes sont susceptibles de ne pas être représentatives de la relation entre variable d'intérêt et donnée auxiliaire. Ce risque se retrouve pour d'autres méthodes de sélection d'un échantillon (Herrero- et al., 2018) telles que la méthode de Kennard and Stone (Kennard & Stone, 1969).

Face à ce problème deux approches sont envisageables. La première consiste à adapter l'approche de sélection des échantillons, par exemple en interdisant le choix de sites pour lesquelles les valeurs de la donnée auxiliaire sont trop extrêmes. L'objectif évident est ici d'améliorer la robustesse de l'estimation en écartant délibérément certaines valeurs au risque de conduire à l'utilisation d'une approche qui ne soit pas optimale. La deuxième approche repose sur un filtrage préalable de possibles aberrants présents dans les données auxiliaires. L'avantage de cette deuxième option réside dans la possibilité d'être adaptée spécifiquement à chaque parcelle puis d'appliquer une approche optimale pour la sélection des sites de mesure quelle que soit la parcelle considérée. Cette deuxième option est celle qui a été choisie dans le cadre de ce travail. Cette option requiert tout d'abord de bien définir ce qu'est une valeur aberrante en considérant les spécificités liées au caractère spatialisé des données considérées.

On qualifie d'aberrant (« outlier » en anglais) toute valeur qui s'écarte tellement du reste des observations qu'elle éveille le soupçon d'avoir été générée par un mécanisme différent (Hawkins 1980). Il s'agit d'un concept général existant au-delà des problèmes d'échantillonnage (Chu Su 2011, Sharma et al. 2017) et partagé en dehors de la discipline de l'agronomie (Wang et al. 2019) comme pour la détection de fraude (Paula et al. 2016, Porwal & S. Mukund 2018), la cybersécurité (Alrawashdeh & Purdy 2016) ou la santé (Gebremeskel et al. 2016). Etre capable de détecter ces valeurs aberrantes est donc un enjeu essentiel dans de nombreux domaines et de multiples approches ont été proposées dans ce sens. Dans le contexte de l'estimation par échantillonnage, cela permet donc de s'assurer que l'estimation ne sera pas biaisée par la présence d'aberrants ou que, dans le cas du *model sampling*, les données sélectionnées soient pertinentes pour étalonner un modèle.

Il existe une grande diversité d'aberrants. Leurs origines sont multiples et ils peuvent être classés selon deux catégories. Prenons l'exemple d'un indice de végétation pour lequel un site de mesure apparait aberrant et présente une valeur anormalement forte. Il peut tout d'abord s'agir d'un évènement qui

traduit une réalité sur le terrain. Ce site de mesure correspond bien, sur le terrain, à une ou plusieurs plantes présentant une vigueur anormalement forte et correctement représentée par l'information disponible. On parle alors d'évènement rare. Mais il peut également s'agir d'une erreur ou d'une imprécision de la mesure. L'acquisition de la donnée aura pu être faussée par l'outil de mesure ou le contexte de la mesure. La valeur de vigueur anormalement forte peut par exemple provenir d'une zone d'enherbement locale, l'activité photosynthétique mesurée par le capteur ne représente alors pas la réalité des propriétés de la culture sur le site de mesure. La distinction entre ces deux types d'évènements est un enjeu essentiel pour les domaines qui étudient l'occurrence d'évènements rares. Dans le cadre d'un échantillonnage en production végétale, ni les évènements rares ni les erreurs ne sont souhaitables pour l'échantillon final. L'enjeu est donc de détecter l'ensemble de ces occurrences.

8.1.2 Diversité des valeurs aberrantes

Il est possible de distinguer plusieurs types de valeurs aberrantes en agronomie, indépendamment de leurs origines, ceux-ci sont décrits dans les sous-parties suivantes.

8.1.2.1 *Valeurs aberrantes globales*

Certains sites ou ensembles de sites de mesure peuvent prendre des valeurs remarquables au regard de la distribution des valeurs présentes sur la parcelle. Ils correspondent typiquement à des valeurs anormalement fortes ou faibles. En viticulture, ils peuvent notamment correspondre à un ensemble regroupé de ceps peu vigoureux ou manquants (Figure 8.1). L'ensemble des valeurs aberrantes globales n'est pas disjoint de l'ensemble des valeurs aberrantes locales, un unique cep manquant pourra alors constituer un aberrant local et global. Ce type de valeurs aberrantes n'est pas spécifique aux propriétés spatiales des données utilisées et se retrouve dans une grande diversité de domaines.

8.1.2.2 *Valeurs aberrantes locales*

La particularité des données associées aux problématiques d'échantillonnage en production végétale se trouve dans le fait que ces données sont généralement spatialisées. Chaque valeur est associée à des coordonnées spatiales. Il est donc possible de définir une distance entre chaque couple de valeur. L'existence d'une structure spatiale dans ces données permet d'affirmer que plus la distance séparant deux observations est faible, plus on peut s'attendre à ce que les valeurs observées soient proches. L'étude de cette propriété permet d'identifier un premier ensemble de sites qui semble déroger à cette règle. Ces sites sont définis comme des aberrants spatiaux ou aberrants locaux. Ils sont habituellement caractérisés par leur différence par rapport à un ensemble de sites proches. Les travaux de Chu Su (2011) notamment, présentent un ensemble d'outils pour la détection de valeurs aberrantes locales dans le contexte de l'agriculture de précision.

8.1.2.3 *Les bords de rang*

Une spécificité de l'étude des données parcellaires concerne la prise en compte des effets de bord. Il s'agit d'un phénomène bien connu en agriculture (Brown & Weibel, 1957 ; Wang et al. 2017 ; Austin & Blackwell., 1980). L'étude de l'hétérogénéité des parcelles révèle que les plantes en périphérie des cultures présentent des conditions de croissance souvent bien différentes du reste de la parcelle. Il s'agit de zones particulièrement exposées à certains ravageurs (Pavan et al. 2012). La diversité du voisinage en bord de parcelle peut aussi affecter le microclimat (vent, luminosité, température etc.) ou les propriétés du sol.

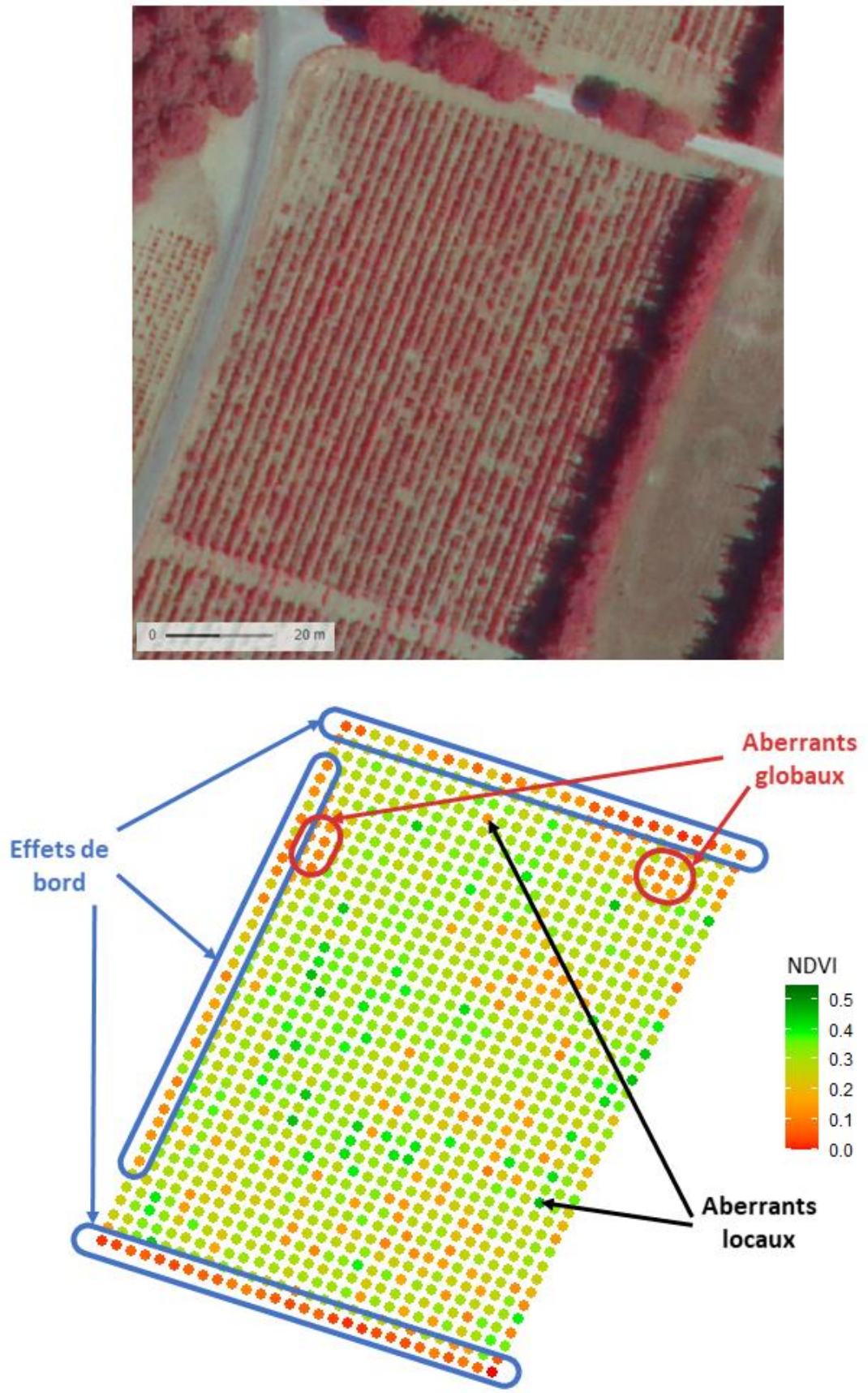

Figure 8.1 : Illustration des effets de bord et valeurs aberrantes locales et globales. En haut, l'image obtenue par imagerie aérienne (résolution 1 pixel = 0,25 m²) pour les longueurs d'onde permettant de calculer le NDVI. En bas, les valeurs de NDVI extraites pour chaque pied de vigne de la parcelle.

En résulte généralement un développement moindre des plantes exposées à ces conditions et des rendements moindres (Sozzi et al. 2020), l'inverse peut aussi se produire ; ces zones peuvent générer des conditions plus favorables à la croissance végétative car la concurrence inter-plante y est plus faible.

L'existence de telles zones pose question pour l'échantillonnage. Les superficies qu'elles représentent peuvent varier en fonction de la complexité du contour des parcelles et de leur surface, et le choix de l'intégration de mesures en bord de parcelle dans l'échantillon final présente un enjeu. Ce choix doit être fait au regard de la représentativité de l'échantillon final, en considérant la part de superficie que constituent ces zones. Dans le cas pratique de l'estimation du rendement en viticulture, les échantillonneurs évitent généralement la réalisation de mesures sur les trois ceps aux extrémités des rangs et sur les deux rangs les plus en périphérie de chaque côté de la parcelle. La mise à l'écart des bordures de parcelles est récurrente en agriculture de précision (Devaux et al. 2019, Cogato et al. 2019).

La Figure 8.1 reprend les concepts de valeurs aberrantes locales, globales et d'effet de bord sur des données de NDVI pour une parcelle viticole. La figure a été élaborée à partir d'une image acquise sur une parcelle de l'INRAE Pech Rouge (Narbonne, France), les classes de valeurs de NDVI y ont été représentées pour chaque site de mesure potentiel.

8.1.2.1 *Valeurs aberrantes temporelles*

Un dernier type d'aberrants apparaît lorsque l'échantillonnage est réalisé selon une composante temporelle (Gupta et al., 2014). Plusieurs mesures sur une même parcelle ou sur un même site à des dates différentes rendent possible l'apparition de valeurs aberrantes. De manière analogue aux données spatiales, deux données temporelles tendent à prendre des valeurs proches lorsqu'elles sont associées à des moments de mesure proche. Une mesure différente d'autres mesures temporellement proches pourra donc apparaitre comme étant aberrante. Le cadre de la thèse étant centré sur l'échantillonnage à un instant déterminé, ces aberrants ne sont pas considérés dans les travaux présentés ci-après.

8.2 Description d'une méthode pour la détection des valeurs aberrantes en production végétale

8.2.1 Approches basées sur la distribution pour la détection de valeurs aberrantes globales.

8.2.1.1 *Principales approches existantes*

La plupart des approches existantes pour la détection de valeurs aberrantes globales se basent sur des hypothèses sur la distribution des données, il s'agit de méthodes paramétriques. Ces méthodes doivent être mobilisées en prêtant attention à la distribution mais permettent facilement d'identifier des grandeurs improbables. Un des principaux avantages de ces méthodes réside dans leur interprétabilité. La méthode dite du $Zscore$ constitue l'une des plus communes. Elle se base sur l'hypothèse d'une loi normale de moyenne μ et d'écart type σ et vise à caractériser l'écart entre une valeur et la moyenne de la distribution, exprimée en nombre d'écarts-types (Eq. 8.1). Cette approche, déjà mobilisée en agronomie (Torres et al. 2017, Mandić-Rajčević & Colosio 2019), classe les différentes valeurs selon leur $Zscore$ et écarte les individus présentant les scores extrêmes en considérant un seuil préalablement défini.

$$Zscore = \frac{value - \mu}{\sigma} \qquad Eq.8.1$$

D'autres approches basées sur des méthodes de classification telles que la méthodologie DB-Scan (Ester et al. 1996) ou les algorithmes de forêt d'isolement (Liu et al., 2008) permettent également de détecter les aberrants globaux et sont déjà utilisés en agronomie (Leroux et al. 2018, Jung et al. 2020). Comme les précédentes, ces méthodes nécessitent cependant un arbitrage dans la définition de valeurs seuils.

8.2.1.2 *Distribution des données auxiliaires*

A la manière du $Zscore$, l'approche retenue ici se base sur la distribution des données. Les indices de végétation utilisés comme données auxiliaires pour l'estimation du rendement tendent à suivre une distribution gaussienne. Une dissymétrie apparaît cependant du fait de la présence des valeurs aberrantes. La Figure 8.2 illustre ce résultat par un histogramme de distribution des valeurs pour la parcelle présentée en Figure 8.1 et les trois parcelles du domaine de l'institut Agro de la section 6.4.

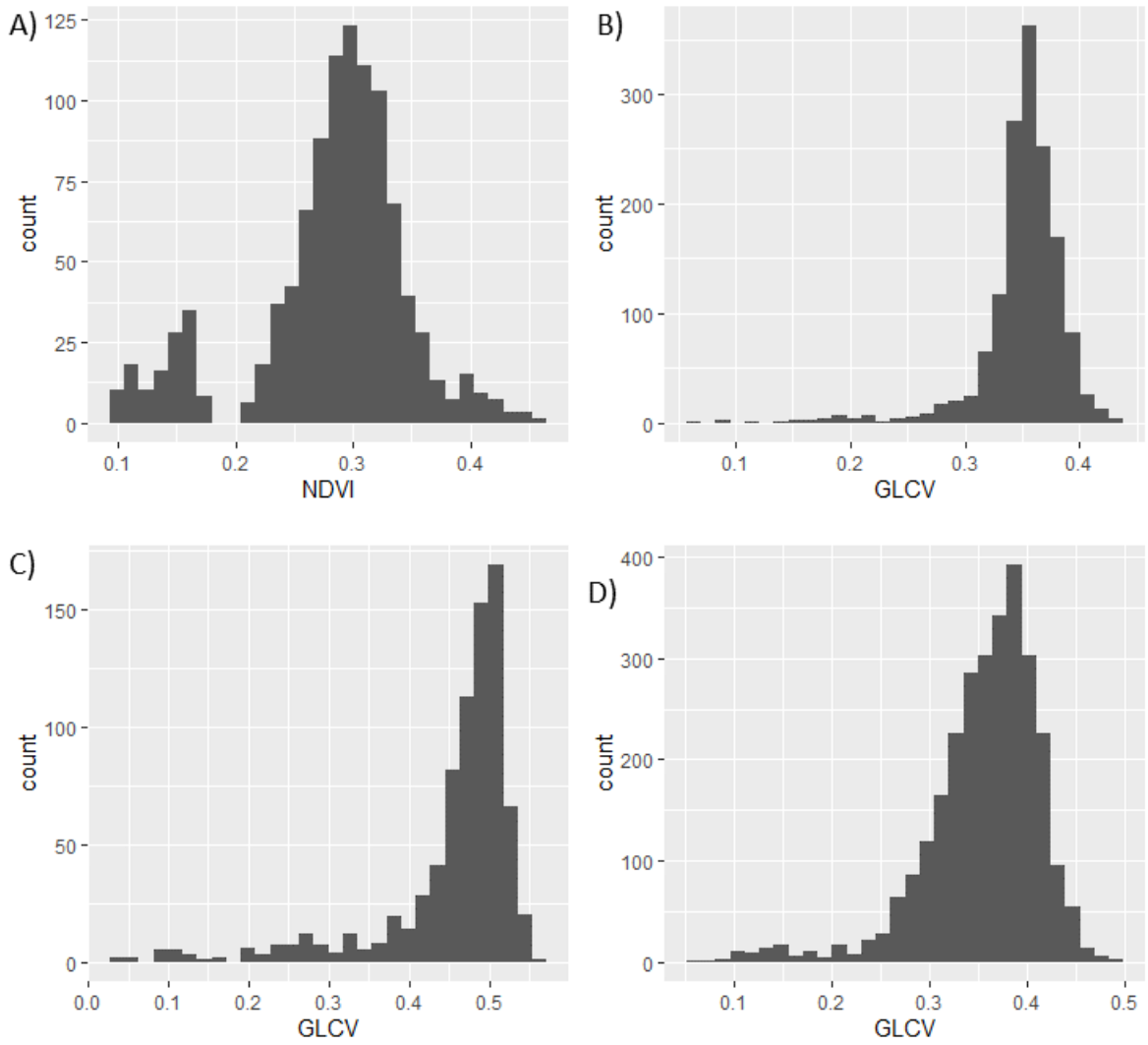

Figure 8.2 : Représentation des données d'indice de biomasse sous forme d'histogramme de 30 catégories. Pour 4 parcelles
A : parcelle figure 1 NDVI sans effets de bord
B, C & D : GLCV lissés pour trois parcelles du domaine du Chapitre (Villeneuve-lès-Maguelone, France)

Ce résultat est cohérent avec la littérature existante. Plusieurs auteurs se basent en effet sur l'hypothèse d'une loi normale pour décrire la distribution d'indices de végétation au niveau d'une parcelle (Matese et al., 2016 ; Li et al. 2018).

Afin de pouvoir utiliser les propriétés de la distribution, il convient d'en estimer les paramètres. La méthode doit être robuste à la présence de potentiels valeurs aberrantes dans la distribution.

La moyenne μ est ainsi estimée à partir de la médiane qui constitue un estimateur plus robuste aux valeurs aberrantes. Bien que moins sensible que la moyenne, cet estimateur est néanmoins affecté par un déséquilibre entre les proportions de valeurs aberrantes fortes et faibles.

Pour l'écart-type σ, particulièrement sensible aux valeurs très éloignées de la distribution, on souhaite estimer ce paramètre uniquement au regard des mesures les plus proches de la médiane. Pour cela, l'approche se base sur une utilisation inversée de la règle empirique des trois sigmas, ou règle des 68–95–99.7 (Walck, 2007). Cette règle indique que pour une distribution suivant une loi normale de paramètre $N(\mu, \sigma^2)$, l'intervalle $[\mu - \sigma, \mu + \sigma]$ tend à contenir 68.26% des valeurs issues de la distribution. Formulé autrement, l'intervalle séparant le 15.87éme $(= \frac{100-68.26}{2})$ percentile et le 84.13éme $(= 100 - 15.87)$ percentile est de longueur 2σ. L'estimation de σ est effectué en prenant la moitié de la longueur effective des cet intervalle sur les données réelles. Cette approche permet une estimation de σ en s'affranchissant des valeurs les plus extrêmes de la distribution et donc de la présence d'éventuels aberrants que l'on cherche justement à identifier.

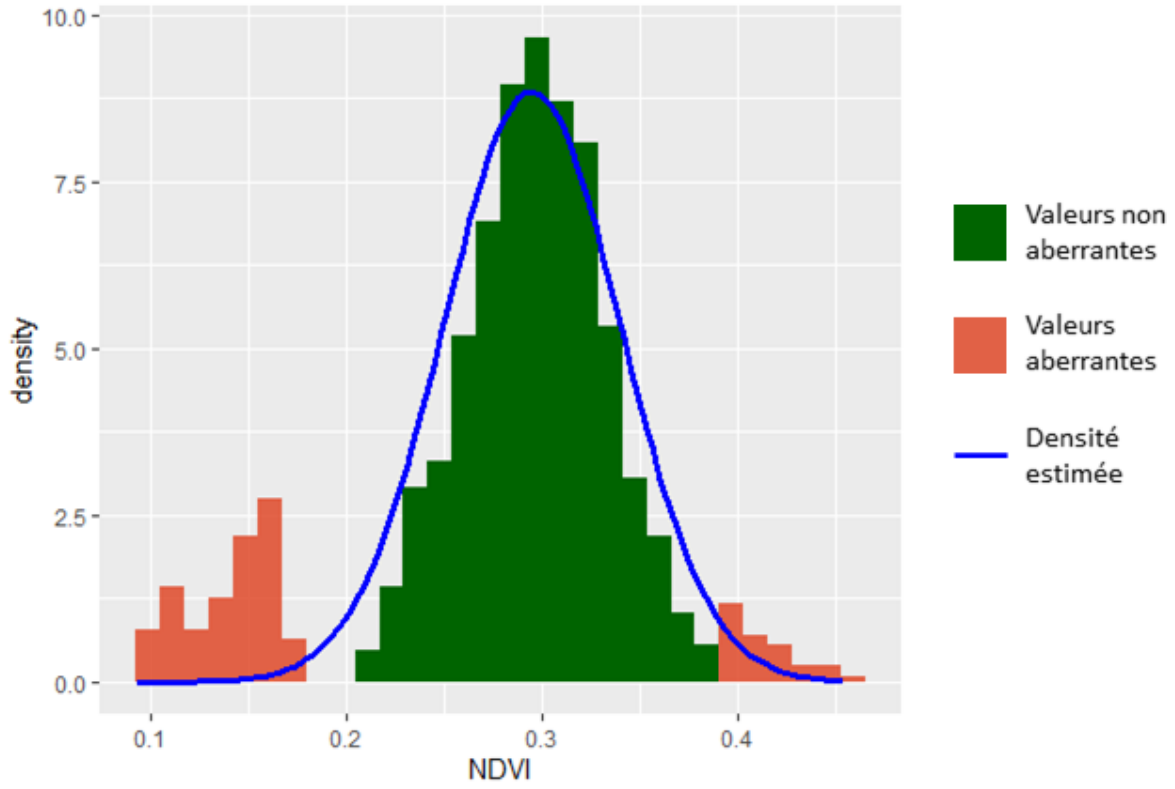

Figure 8.3 : Estimation de la distribution des données de la parcelle exemple de la figure 8.1 avec une loi normale. En rouge les valeurs associées à une probabilité <5% considérées comme aberrantes.

La Figure 8.3 illustre l'application de ces méthodes pour estimer la distribution du NDVI pour la parcelle présentée Figure 8.1 .Elle met clairement en évidence l'intérêt de l'approche pour identifier les paramètres d'une distribution en s'affranchissant des valeurs extrêmes.

8.2.1.1 *Identification de valeurs improbables*

Une fois estimés, les paramètres de la distribution des données permettent de définir la probabilité qu'une observation tombe dans un certain intervalle :

$$X \sim N(\mu, \sigma^2), \qquad P(X \in [a,b]) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} . dx \qquad Eq.8.2$$

Il devient alors possible d'associer une observation x_1 à un $Zscore$ ou à la probabilité d'obtenir la même valeur ou une valeur encore plus extrême que celle observée (équivalent à une p-value) :

$$X \sim N(\mu, \sigma^2), \qquad Eq.8.3$$
$$P(|X-\mu| > |x_1-\mu|) = \; P(X \in]-\infty, -|x_1-\mu|]) + P(X \in [|x_1-\mu|, +\infty[)$$

La dernière étape consiste donc à définir le seuil à partir duquel une observation peut être définie comme suffisamment improbable pour être écartée. Ce seuil peut être fixé de manière arbitraire. Pour une probabilité, le seuil fixé correspond aux chances de considérer à tort une valeur comme étant aberrante. De manière analogue aux tests statistiques, on peut par exemple choisir une valeur de 5%.

La Figure 8.3 présente également en rouge les valeurs considérées comme aberrantes et écartées. Le seuil y est fixé à 5% de probabilité d'obtenir la même valeur ou une valeur encore plus extrême.

8.2.1 Ecart au voisinage et valeurs aberrantes localement

8.2.1.1 *Le voisinage d'un site de mesure*

La détection des valeurs aberrantes locales ou spatialement aberrantes se fait sur la base de comparaison aux valeurs proches disponibles (Leroux et al. 2018). Dans ce but, il convient de définir l'ensemble des valeurs proches qui seront mobilisées pour qualifier une valeur étudiée. Cet ensemble est appelé le voisinage de la valeur étudiée. Le choix d'un voisinage se base généralement sur la définition d'une distance et d'une valeur limite au-delà de laquelle les observations ne sont pas considérées. La distance communément utilisée reste la distance euclidienne classique. Le choix de la distance seuil à partir de laquelle les valeurs sont incluses ou non dans le voisinage est plus variable et sujet à expertise. Cette distance peut être définie de manière à contenir l'ensemble des sites que l'on sait être auto-corrélés spatialement avec la valeur étudiée. Elle est alors basée sur la portée du semi-variogramme (Annexe : A propos du semi-variogramme) de la grandeur étudiée. Cette distance seuil peut également être définie à partir des contraintes opérationnelles qui s'appliquent à l'échantillonnage, comme la précision sur les coordonnées de la mesure, la précision du GPS de l'opérateur en charge de l'échantillonnage ou encore la dimension des sites de mesure. Le choix du voisinage doit être adapté à l'approche mobilisée pour détecter les sites aberrants (Vega et al. 2019 ; Gozdowski et al 2010).

La Figure 8.4 présente plusieurs exemples de voisinages. Par la suite on définit comme appartenant au voisinage toutes les valeurs associées à une distance de moins de 5m de la valeur étudiée. Cette distance correspond en pratique à la précision d'un récepteur GNSS (Global Navigation Satellite System) piéton équipé d'une correction gratuite EGNOS.

8.2.1.2 *Mesure de la proximité d'un site à son voisinage*

Une fois un voisinage défini, il convient de trouver une grandeur permettant de classer les observations et d'identifier les valeurs spatialement aberrantes. Plusieurs approches ont été proposées dans la littérature pour résoudre ce problème (Kou et al. 2006, Chen et al. 2008, Filzmoser et al. 2014, Harris et al. 2014, Leroux et al. 2018) et reprennent le formalisme de Lu et al. (2003). Soit $f(x_i)$ la valeur attributaire observée pour le site x_i. Ces approches reposent sur la définition d'une grandeur $g(x_i)$ quantifiant les valeurs attributaires observées dans le voisinage de x_i. Cette fonction g correspond généralement à une moyenne pondérée ou à une médiane des valeurs voisines. Une troisième grandeur h est définie afin de comparer l'écart entre f et g. Il peut s'agir d'un ratio $h(x_i) =$

$f(x_i)$ / $g(x_i)$ ou d'une différence $h(x_i) = f(x_i) - g(x_i)$. La grandeur $h(x_i)$ représente alors la propension de x_i à constituer un aberrant local.

La Figure 8.4 illustre cette méthode par une mise en œuvre sur la parcelle présentée dans la Figure 8.1. Pour trois types de voisinages différents, les valeurs de données auxiliaires sont comparées à leur prédiction par une moyenne du voisinage pondéré. La pondération est mise en œuvre par une approche déterministe basée sur l'inverse de la distance.

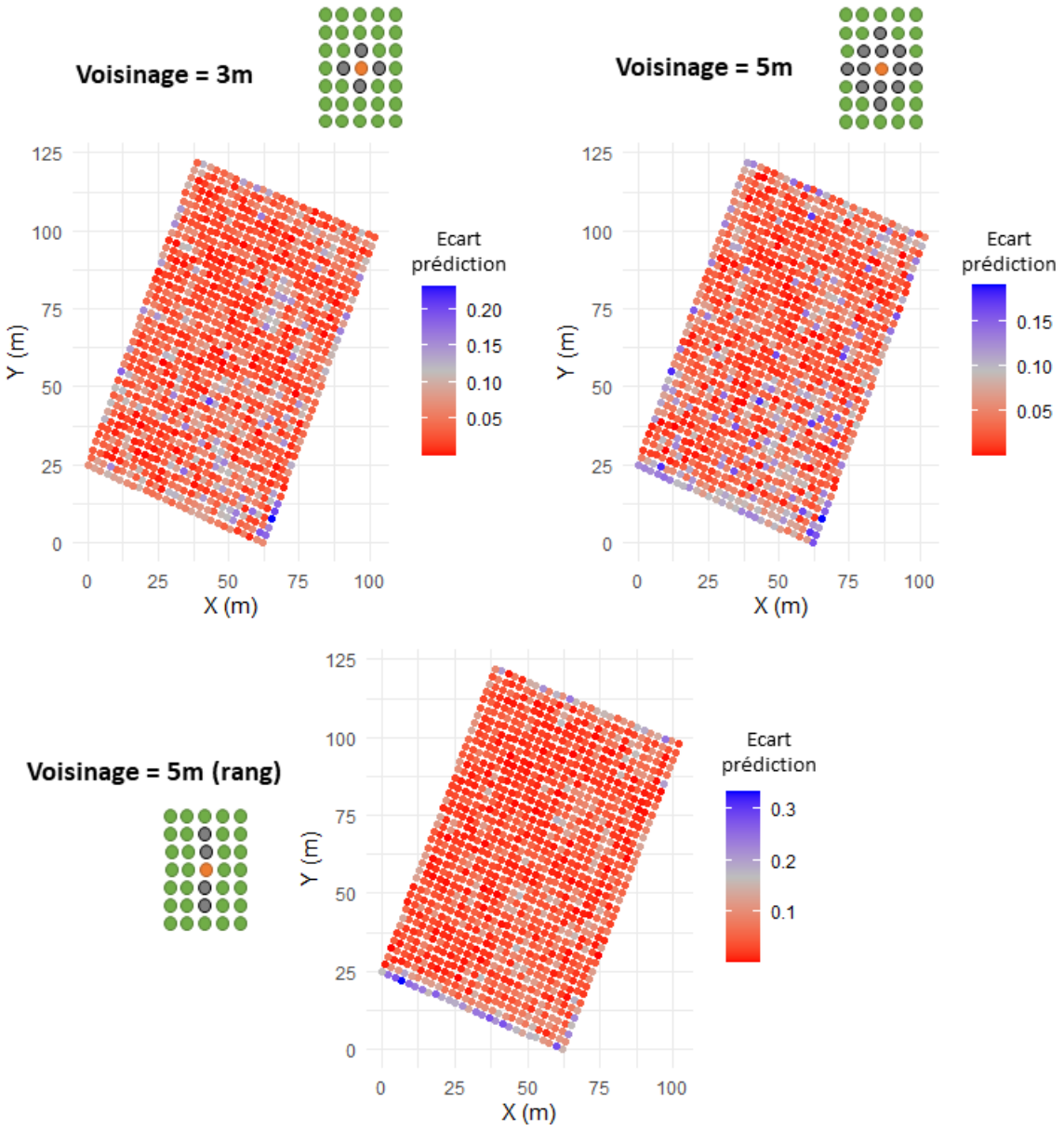

Figure 8.4 : Application des méthodes associées à la recherche d'aberrants locaux pour la parcelle présentée en figure 8.1. Illustration des approches par écart à la prédiction pour trois types de voisinage différents.

8.2.1.3 *Seuils de détection*

Il convient de définir une valeur seuil à partir de laquelle une donnée peut être considérée comme aberrante. Il n'existe pas de règle universelle permettant de définir ces valeurs qui sont propres à chaque parcelle, de la grandeur mesurée et du voisinage considéré. Ces seuils peuvent être définis de

plusieurs manière sur la base d'une expertise ou d'approches probabilistes à partir des distributions des données (Lu et al. 2003).

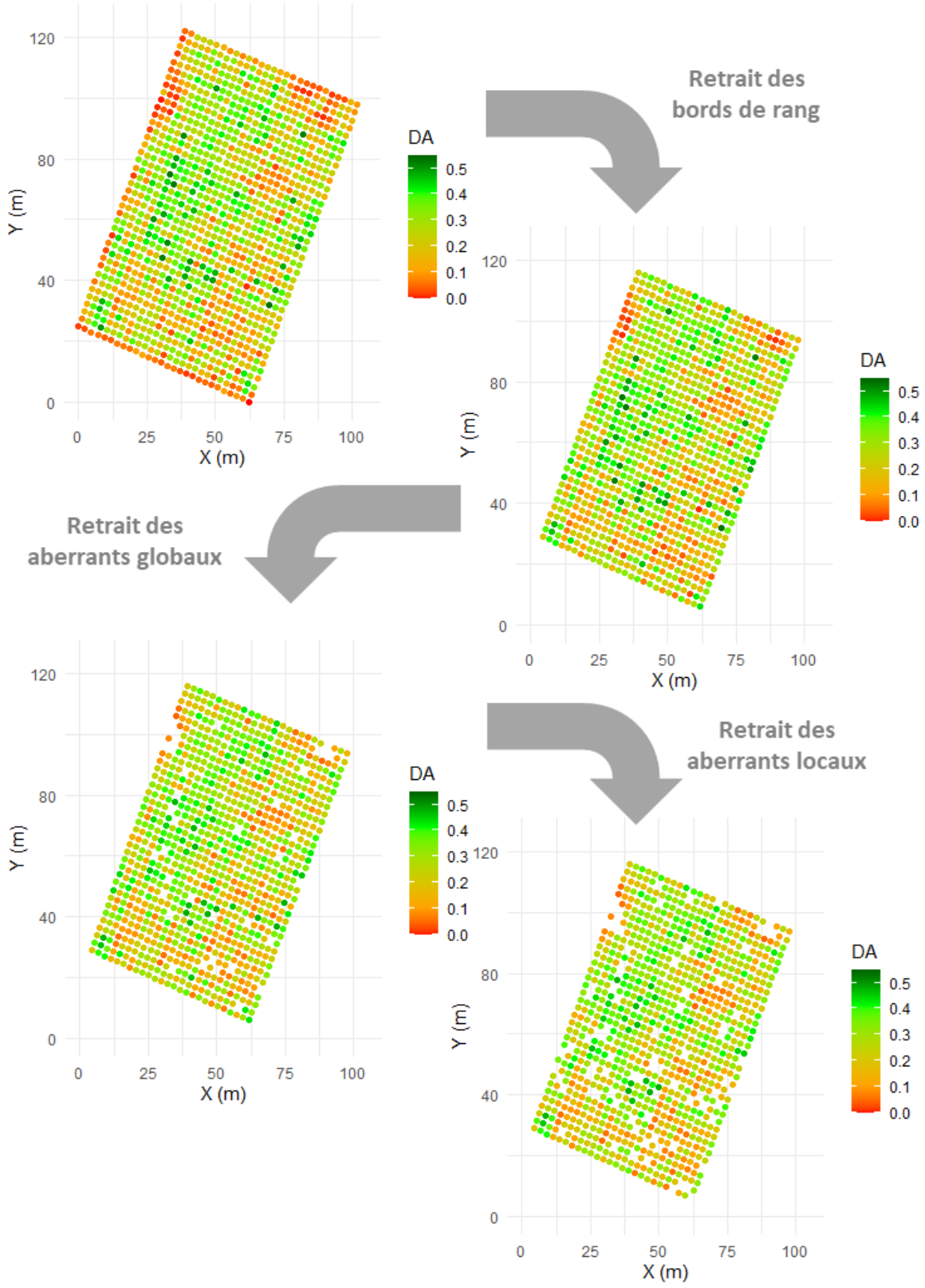

Figure 8.5 : Les différentes étapes dans la détection des valeurs aberrantes pour une parcelle type. Dans l'ordre de haut en bas : données initiales complètes ; retrait des effets de bord ; retrait des valeurs aberrantes globales et locales.

Lorsque l'on dispose d'un jeu de données pour lequel les valeurs aberrantes sont identifiées, il est possible de comparer les performances obtenues pour différentes valeurs de seuil ou différents voisinages afin d'identifier les meilleurs paramètres. Cela permet également de caractériser les risques d'obtention de faux positifs et faux négatifs de la méthode.

8.3 Mise en place de l'approche

La Figure 8.5 résume le prétraitement proposé sur une la même parcelle que celle présentée à la Figure 8.1. Les effets de bord sont écartés en délaissant le rang le plus extrême de chaque côté de la parcelle et les deux premières mesures de chaque côté des rangs restants. Dans un second temps, les valeurs associées à une probabilité inférieure à 0.05 d'existence au regard de la distribution sont également retirées pour écarter les aberrants globaux. Enfin, pour les aberrant locaux, les 5% de sites présentant les plus forts écarts à la prédiction avec un voisinage pris à 5m sont écartés.

Cette même méthode est testée sur les données simulées présentées dans le chapitre 3. Pour ces simulations, la portée est fixée à 50m, le ratio à 0.33, la corrélation à 0.6 et le nombre de zones aberrantes à 6. 300 simulations sont réalisées. Pour chaque simulation, deux échantillons de 5 sites de mesure sont sélectionnés respectivement avant et après filtrage. Ceux-ci sont réalisés selon la méthode de Kennard and Stone (1969). Cette méthode, sensible à la présence d'aberrants, impose de prendre selon la donnée auxiliaire :

- Le maximum (max)
- Le minimum (min)
- Les trois valeurs respectivement les plus proches de : $min + (max - min) \times \frac{i}{4}$, $i \in \{1,2,3\}$

A chaque fois l'erreur d'estimation est calculée sur la base d'une inférence reposant sur un modèle linéaire. La Figure 8.6 donne les résultats sous forme d'une boîte à moustache.

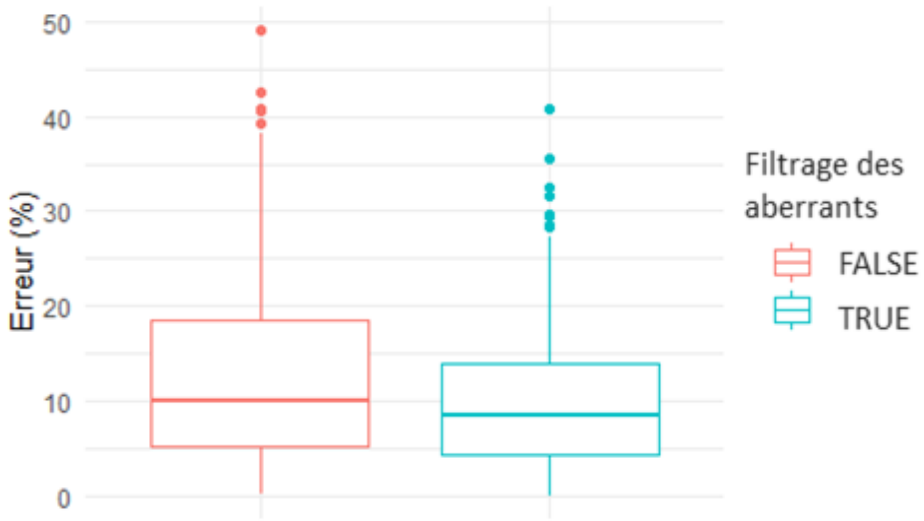

Figure 8.6 : Comparaison des erreurs d'estimation avec (en bleu) et sans (en rouge) filtrage préalable des valeurs aberrantes sous forme de boîte à moustache. Les résultats proviennent de 300 échantillons obtenus suivant la méthode de Kennard & Stone.

Cette expérimentation fait apparaitre le gain potentiel apporté par la détection des valeurs aberrantes. La variance des erreurs apparaît sensiblement plus importante en l'absence de détection des valeurs aberrantes. Un test de comparaison de variance conclut que la variance des erreurs après le retrait des

aberrants est inférieure avec une p-value égale à $2{,}8 \times 10^{-5}$. La Figure 8.6 fait notamment apparaitre un nombre plus important d'estimations avec des erreurs très importantes (supérieures à 30%).

Les erreurs d'estimation moyennes restent relativement proches, le retrait des valeurs aberrantes permet de diminuer la moyenne des erreurs d'environ 2.4%. Un test de Welch rejette l'hypothèse selon laquelle les moyennes sont égales. L'hypothèse alternative, selon laquelle la moyenne des erreurs après le retrait des aberrants est inférieure, est acceptée avec une p-value égale à 0.0002.

8.4 Conclusion et perspectives pour la détection des valeurs aberrantes

Ce dernier chapitre présente une première approche pour l'identification des valeurs aberrantes dans le cadre d'un échantillonnage. Celle-ci, testée sur des données simulées, repose sur l'élimination successive :

- Des effets de bord, les ceps en périphérie de la parcelle étant généralement associés à des conditions de croissance différentes ;
- Des aberrants globaux, qui se distinguent du reste de la distribution ;
- Des aberrants locaux, qui se distinguent des autres valeurs présentes dans un voisinage proche.

Pour une même variable, la diversité des parcelles limite cependant l'automatisation des approches et la définition de valeurs seuils pour le voisinage et les critères choisis. Cela, couplé à la diversité des grandeurs échantillonnées en production végétale et des données auxiliaires envisageables rend difficile l'établissement d'une méthodologie applicable de manière systématique. Eliminer 5% des sites comme étant des aberrants locaux apparaît comme une solution opérationnelle compte tenu du volume de données disponibles. En effet, 95 % de sites restants constituent un ensemble de points suffisant pour s'assurer de pouvoir trouver un échantillon solution pertinent, que ce soit pour l'étalonnage d'un modèle ou pour l'optimisation de la distance de parcours. Cette valeur ne serait cependant pas généralisable à des parcelles présentant une très faible ou très forte proportion d'aberrants.

Cette variabilité des parcelles et des données a également des conséquences sur les distributions que peuvent suivre les valeurs de données auxiliaires. Si la majorité des parcelles semblent favorables à la représentation de la distribution par une loi normale. Il peut exister des exceptions présentant des distributions plus complexes telles que des lois de Weibull (Le Maire et al., 2006). Peuvent exister également des modèles de mélange de plusieurs lois normales (Skakun et al., 2017), ce phénomène peut apparaitre par exemple pour des parcelles composées de plusieurs zones avec des conditions de développement différentes.

Il reste donc difficile d'établir une méthode universelle pour la détection de ces valeurs. L'automatisation de la détection des aberrants dans un contexte d'échantillonnage représente un axe de recherche important qui nécessiterait du temps et des données réelles supplémentaires pour l'obtention d'une méthode opérationnelle. Toutefois, d'un point de vue opérationnel, l'approche proposée montre, dans le cas du *model sampling*, qu'il est possible de réduire le risque d'étalonner de « mauvais modèles » en évitant la sélection de valeurs trop extrêmes. Cette étude a permis de mettre en évidence la nécessité d'introduire une approche de détection des aberrants. Elle montre également qu'il s'agit d'une question complexe nécessitant une base de cas plus conséquente que celle qui était à disposition dans le cadre de ce travail pour proposer une approche robuste et généralisable.

Conclusions et perspectives :

Cette thèse s'est intéressée à la problématique de l'échantillonnage spatial des parcelles pour l'estimation de grandeurs caractéristiques. Les méthodes développées sont potentiellement applicables à tous les systèmes de production végétale. Toutefois, la thèse s'est largement focalisée sur l'estimation du rendement en viticulture. D'un point de vue opérationnel, notre démarche s'est inscrite dans le cadre de l'agriculture de précision (viticulture de prédiction) et fait l'hypothèse de la disponibilité d'une ou plusieurs observations à haute résolution spatiale (cartographie des sols, télédétection, etc.) permettant de renseigner sur la nature de la variabilité de la parcelle. Ces observations sont appelées données auxiliaires et sont utilisées pour orienter le choix des sites de mesure afin d'estimer au mieux une variable d'intérêt. L'objectif applicatif de la thèse a été de proposer des méthodes d'amélioration des pratiques d'échantillonnage pour mieux répondre aux enjeux du coût et d'imprécision de l'estimation. Pour répondre à cette problématique, nous avons proposé de répondre à ces deux enjeux simultanément au moment du choix des sites de mesure.

Dans la littérature scientifique, le choix des sites est souvent uniquement raisonné vis-à-vis de l'erreur d'estimation. La question du coût de l'estimation, liée aux déplacements dans la parcelle, y est souvent écartée. En pratique, elle est pourtant intégrée de manière empirique sur le terrain. Nous avons montré que la prise en compte de cette contrainte au moment du choix des sites de mesure permet une réduction importante des distances qu'il est nécessaire de parcourir. Les expérimentations menées dans le chapitre 2 et 3 montrent que ces réductions sont bien supérieures à celles apportées par une simple optimisation de la distance a posteriori (i.e. après que les sites de mesure aient été choisis). La thèse s'est donc intéressée à la mise en place d'une approche visant simultanément à considérer le coût et la minimisation de l'erreur d'estimation dans le choix des sites de mesure.

La recherche d'un échantillon qui soit optimal au regard de la longueur de son parcours et de ses propriétés statistiques représente un problème original. Il requière en effet de combiner ensemble des approches stochastiques et d'optimisation appartenant à des domaines scientifiques différents. Il s'agit également d'un problème combinatoire dont la résolution nécessite l'utilisation de méthodes adaptées. Dans ces travaux, ce problème d'optimisation a dans un premier temps été adressé à l'aide de la programmation par contraintes. Pour de tels problèmes, ces méthodes permettent en effet de s'attaquer à le recherche d'une solution optimale. Bien qu'efficaces sur les instances les plus petites, elles requièrent cependant des temps de calculs qui augmentent de manière exponentielle avec la complexité du problème. Une seconde approche complémentaire, basée sur les méthodes de recherche opérationnelle, a donc été mise en place dans l'objectif de trouver des solutions intéressantes dans un temps plus court. Celle-ci est basée sur l'utilisation d'algorithmes de recuit simulé et de colonies de fourmis. Elle propose l'amélioration itérative d'un échantillon afin de rapidement faire converger celui-ci vers une solution.

La réponse à la problématique a ainsi débouché sur la mise en place d'une nouvelle approche pour l'échantillonnage. Cette approche, nommée *constrained sampling,* complète les approches de *model sampling* en intégrant l'optimisation de la longueur d'échantillonnage dès la sélection des sites. Plusieurs variantes de cette méthode ont vu le jour au cours de cette thèse dans le but d'en améliorer les performances globales, qu'il s'agisse de la distance du parcours ou des erreurs d'estimation.

En s'appuyant sur un formalisme statistique, nous avons confirmé les résultats expérimentaux de Carillo et al. (2016) montrant que le recours à un modèle pour l'estimation d'une espérance permet

d'obtenir des estimations de meilleure qualité. En dérivant l'expression de la variance d'un tel estimateur, il apparaît que le choix des sites de mesure a des conséquences directes sur l'imprécision de l'estimation. Nous avons ainsi pu proposer un critère permettant de guider le choix des sites de mesure en fonction des valeurs de la donnée auxiliaire. Selon ce critère, il est préférable de sélectionner des sites de mesure dont la moyenne est proche de l'ensemble de la parcelle au sens des données auxiliaires. Il fait ainsi apparaître la notion de représentativité des échantillons. Dans le même temps, ce critère favorise la dispersion des valeurs de l'échantillon autour de leur propre moyenne. L'augmentation de la variance de l'échantillon réduit l'imprécision dans l'estimation des paramètres du modèle et contribue ainsi à améliorer la qualité de l'estimation finale. Le problème de représentativité n'est d'ailleurs pas exclusif à l'utilisation d'un modèle mais apparaît aussi dans le biais d'une estimation basée sur la moyenne.

Un travail original de la thèse a consisté à « inverser » l'utilisation de l'approche afin d'étudier les solutions optimales d'échantillonnage proposées par notre algorithme pour différentes parcelles. Le cas d'étude s'est focalisé sur l'estimation du rendement en viticulture. La structure en rang des parcelles y contraint en effet fortement le déplacement des opérateurs. L'intérêt a été de pouvoir comparer les solutions optimales trouvées par rapport aux pratiques d'échantillonnage couramment utilisées par les professionnels. Ce travail a permis de mettre en œuvre des parcours d'échantillonnage optimaux originaux par rapport aux pratiques courantes. Il a également permis de montrer que la stratégie optimale d'échantillonnage variait selon certaines caractéristiques de parcelles et que les parcours d'échantillonnage optimaux dépendaient de leurs caractéristiques respectives (longueur des rangs, forme et organisation de la variabilité intra-parcellaire).

Le cas d'étude de l'estimation du rendement en viticulture a permis de tester et de valider ces approches sur des données réelles et simulées. Bien que complexe, ce cas ne permet certainement pas de prendre en compte l'ensemble des problématiques d'échantillonnage en production végétale. En particulier, il est probable que d'autres contraintes puissent être déterminantes dans le choix des sites de mesures intra-parcellaire pour d'autres cultures. On pourra citer par exemple la banane qui est une production asynchrone et dont la phénologie des plantes pourrait constituer une contrainte à prendre en compte pour définir un parcours d'échantillonnage optimal. Toutefois, ce travail propose un cadre général susceptible d'être adapté aux spécificités des cultures par l'introduction de contraintes nouvelles, par l'adaptation à de nouvelles données auxiliaires, ou par la définition d'un site de mesure adapté à la densité de la culture considérée. Dans tous les cas, de nouvelles validations, seraient nécessaires afin de mieux identifier l'apport et les limites des approches de *constrained sampling* pour chaque situation.

Le cas d'étude choisi a conditionné le nombre de sites de mesure constituant un échantillon, fixé ici entre 5 et 10. Ce paramètre influence fortement le choix des outils mis en place pour la résolution du problème. Si les critères identifiés au cours de la thèse permettent de raisonner un nombre plus important de sites de mesure, la complexité du problème augmente néanmoins avec ce nombre. Une amélioration des méthodes d'optimisation et des heuristiques proposées pourrait s'avérer nécessaire pour permettre une résolution du problème d'échantillonnage avec un plus grand nombre de sites de mesure.

De manière plus générale, les méthodes de résolution mises en place dans le cadre de cette thèse démontrent leur faisabilité et leur pertinence. L'obtention d'une approche qui soit applicable dans un contexte de production nécessiterait des travaux supplémentaires afin d'optimiser les temps de calculs et la gestion de la mémoire.

Plusieurs autres pistes pourraient être étudiées afin de généraliser certains verrous identifiés :

- La prise en compte de plusieurs données auxiliaires, ce qui reviendrait à guider le choix des sites de mesure selon un espace multivarié. L'avantage opérationnel serait de mobiliser une plus grande part d'information lorsque celle-ci est disponible. Cela contribuerait par exemple à valoriser des données historiques de rendement sur plusieurs années, des indices de végétations disponibles pour différents stades phénologiques ou à croiser les données provenant de différentes sources (statut hydrique, propriétés des sols, exposition à certains ravageurs etc.). Intégrer cet aspect multivarié suppose naturellement de revoir un certain nombre de développements proposés dans le cadre de la thèse et en premier lieu l'adaptation du critère proposé.
- L'utilisation d'autres types de modèles : le modèle linéaire constitue un point d'entrée intéressant de par sa robustesse, sa simplicité et son applicabilité à un grand nombre de situations. Sur la base des connaissances disponibles, l'utilisation de modèles plus spécifiques aux relations pouvant exister entre certaines variables d'intérêt et données auxiliaires constituerait une évolution intéressante des approches de *model sampling* en général. Cette recherche est toutefois conditionnée au développement et à l'adoption de données à haute résolution en agriculture afin de bénéficier d'une base de données plus large que celle disponible aujourd'hui pour proposer et valider ces approches.
- Le filtrage de valeurs considérées comme aberrantes. Cet aspect, amorcé dans le dernier chapitre, augmenterait la confiance qu'il est possible de placer dans l'estimation en s'assurant de la qualité des sites choisis pour l'étalonnage d'un modèle.

Une autre perspective serait de rendre adaptables les approches d'échantillonnage. La mise en place d'une méthode souple capable de prendre en compte les premières observations afin d'ajuster en temps réel la stratégie d'échantillonnage représente une perspective intéressante pour le développement opérationnel de l'approche d'échantillonnage. En effet, en supposant que la méthode soit embarquée sur un terminal mobile géo-référencé, elle permettrait de produire des résultats permettant à l'opérateur d'adapter l'échantillonnage en temps réel en fonction des premières observations effectuées sur la parcelle. Cette question initialement inclue dans le contour de la thèse n'a pas été développée car il est apparu très vite qu'elle constituait un verrou scientifique fort. En effet, elle suppose de mettre en œuvre des statistiques robustes avec un très faible nombre de données (les trois ou quatre premières observations effectuées sur la parcelle). Pour lever ce verrou, l'utilisation de méthodes statistiques bayésiennes intégrant une expertise sur les lois pourrait *a priori* constituer une piste de recherche intéressante.

Annexes :

A propos du NDVI :

Le NDVI (ou *normalized difference vegetation index*) appartient à la catégorie des indices de végétation. Ces derniers sont obtenus à partir d'images multispectrales prises par satellite, avion ou drone qui mesurent les réflectances d'objets ou de parcelles sur différentes longueurs d'onde. En utilisant les propriétés optiques de la végétation, principalement dans le rouge et le proche infrarouge, il est alors possible de caractériser la présence de végétation. En effet les réflectances dans le rouge diminuent avec la présence de la végétation car elles correspondent au pic d'absorbance de la chlorophylle tandis que les réflectances dans le proche infrarouge augmentent. Il existe une grande diversité d'indices (Bannari et al. 1995) obtenus sur la base d'opération entre les mesures faites sur différentes bandes spectrales. Ceux-ci très utilisés en agriculture de précision car facilement accessibles à grande échelle. Ils permettent suivre la dynamique de la végétation d'une parcelle et d'estimer d'autres paramètres biophysiques caractéristiques des couverts végétaux, comme la biomasse, le rendement ou l'indice de surface foliaire (Rousseau et al., 2008). Le NDVI (Rouse & Haas 1973, Tucker 1979) est le plus connu et le plus utilisé d'entre tous, il se calcule à partir des mesures de réflectance dans le rouge (ρ_R) et le proche infrarouge (ρ_{PIR}) :

$$NDVI = \frac{\rho_{PIR} - \rho_R}{\rho_{PIR} + \rho_R}$$

Les valeurs du NDVI sont comprises entre -1 et +1 et augmentent avec la présence de végétation. Les valeurs négatives étant généralement associées à l'absence de couvert végétal.

A propos du semi-variogramme :

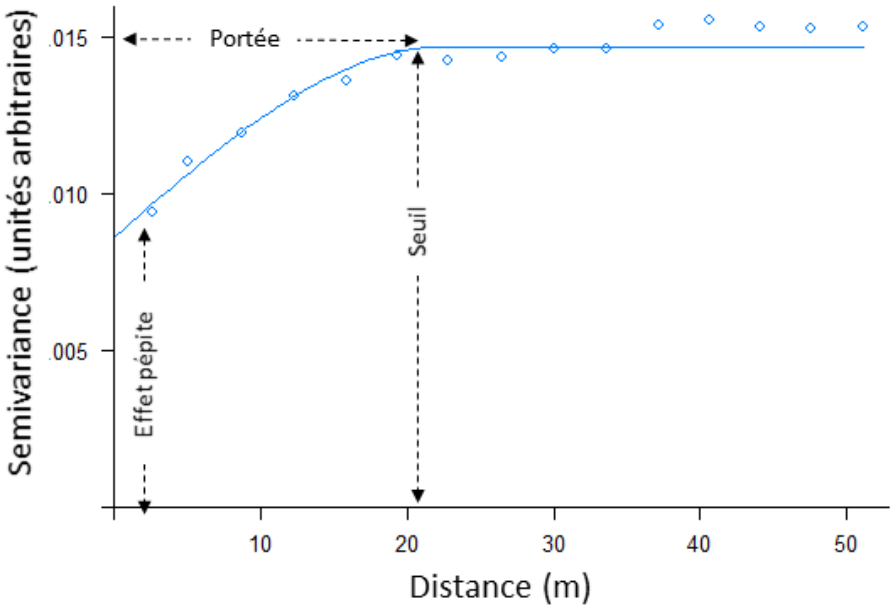

Figure A.1 : Semi-variogramme (courbe bleue) ajustée sur des données expérimentales (points bleus). Le modèle est décrit par trois principaux paramètres : la portée, le seuil et l'effet pépite

En agriculture de précision l'étude de la variabilité spatiale est presque systématiquement associée au semi-variogramme ou variogramme (Bachmaier et Backes, 2008). Le semi-variogramme est un modèle décrivant la variance entre deux observations en fonction de la distance qui les sépare. Ce variogramme peut prendre plusieurs formes selon les données qu'il représente (exponentielle, gaussien, sphérique…). Lorsqu'il existe une structure spatiale, il est représenté par une courbe pour laquelle se distinguent deux parties sous l'hypothèse de stationnarité : tout d'abord une évolution rapide de la variance qui augmente avec la distance séparant les observations, puis un palier, au-delà d'une certaine distance la variance n'évolue plus (Figure A.1). Trois grandeurs sont généralement utilisées pour décrire cette courbe : l'effet pépite (ou *nugget effect*) qui représente l'ordonnée à l'origine et la part de variabilité non structurée spatialement ou erratique ; le seuil (ou *sill*) qui représente la variance atteinte par le plateau et qui correspond à la variance totale de la parcelle ; la portée (ou *range*) qui correspond à la distance à partir de laquelle deux observations sont considérées comme indépendantes.

Article de conférence : ECPA 2019

Combining target sampling with route-optimization to optimise yield estimation in viticulture

B. Oger[1-2], P. Vismara[2-3] and B. Tisseyre[1]
[1] ITAP, Univ. Montpellier, Montpellier SupAgro, Irstea, France
[2] MISTEA, Univ. Montpellier, Montpellier SupAgro, INRA, France
[3] LIRMM, Univ. Montpellier, Montpellier SupAgro, CNRS, France
baptiste.oger@supagro.fr

Abstract

This paper describes a new approach for yield sampling in viticulture. It combines approaches based on auxiliary information and path optimization to offer more consistent sampling strategies, integrating statistical approaches with computer methods. To achieve this, groups of potential sampling points, comparable according to their auxiliary data values are created. Then, an optimal path connecting several points, one from each group of potential sampling points and minimizing the route distance is constituted. This part is performed using constraint programming, a programming paradigm offering tools to deal efficiently with combinatorial problems. The paper presents the formalization of the problem, as well as the tests performed on real fields. Results show that combining target sampling and path optimization can reduce by 45% the average sampling circuit length compared to previous methods based on auxiliary data while being almost equivalent in yield prediction error.

Keywords: sampling optimization, yield estimation, model sampling, NDVI, constraint programming.

Introduction

In order to optimize harvest organization, prepare the winemaking process and establish commercial strategies, the wine industry needs to know the yield of each vine field. Ideally, yield has to be estimated a few days before harvest with a relative error of less than 10 %. Although models have been developed to forecast the yield at the regional level (Cristofolini and Gottardini 2000), their results were not precise enough to manage logistic issues linked to harvest operations at the farm or at the winery level. Therefore, precise estimation of vine field yield always requires fruit sampling and counting. This estimation must be carried out quickly (few minutes per field) at a time when the workload at harvest or for the preparation of the harvest is critical. Practical constraints, like the time available to visit all the fields before harvest, limit the number of sampled sites per field. Therefore, yield estimation is based on a low number of sampling sites (4/5 per field) where yield components (number of clusters, number of berries per cluster, mean berry weight) are manually measured by a practitioner. Due to these practical constraints and the high within-field variability of grape yield usually observed, the small number of observation results in high errors in yield estimation (generally around 20 to 30%).

Recent works (Carillo et al. 2016) have shown the interest of integrating auxiliary data to improve sampling strategies and yield estimation for perennial crops. Among possible auxiliary information, vegetation indices derived from multispectral airborne images is of great interest since they can be used to characterise the spatial variability of several fields; in one acquisition, with a high spatial resolution (< 1 m.) and at an optimal date. In viticulture, Carrillo et al., (2016) showed the potential of

normalised difference vegetation index (NDVI) to drive target sampling of the main grape yield components (e.g. bunch number, berry weight) to improve yield estimation. They demonstrated the value of using NDVI information to determine relevant within-field sampling sites selection based on the distribution of NDVI values.

Although interesting, the methodology proposed by Carrillo et al. (2016) presents a significant drawback. Indeed, it does do not take into consideration the relative position of the sites to be sampled, and the fact that vine fields are structured in rows. This peculiarity implies that rows cannot be crossed, leading to sampling plans optimized in terms of prediction but potentially unrealistic in terms of sampling routes and resulting travelled distance (and time) for the operator.

This paper proposes a new approach to optimally design within-field sampling routes which take into account the spatial organisation of the crop (rows) and spatial location of sampling sites. The originality of the approach, called constraint sampling, is to combine statistical and computer methods. It can be decomposed into two steps. In the first step, potential sampling sites are sorted into different groups according to their auxiliary data value in a similar way to traditional targeted sampling. The second step finds an optimal route that passes through one sampling site from each group. A Constraint Programming solver is used to build an optimal route in terms of travelled distance. This kind of solver has already been used in precision viticulture to solve the differential harvest problem (Briot et al. 2016).

Materials and methods

Sampling sites and selection principles

The purpose of *constraint sampling* is to select N sampling sites constituting a sampling route in the vine field. Accounting for classical sampling practices in viticulture, N will vary between 5 and 10. It is assumed that there is a finite number of sites on the plot where sampling can be carried out, these sites are called potential sampling sites. For instance, in the data presented in this article, a potential sampling site is defined every 15m. The sampling sites are then chosen from the list of potential sampling sites. In order to be able to apply selection methods based on auxiliary data, an NDVI value must be associated with each potential sampling point. For each potential site, the method assumes that: i) the co-ordinate of the potential site, ii) the row that the potential site belongs to and iii) the corresponding NDVI values are known.

The method requires a distance matrix to be computed. This matrix gives the distance between each couple of potential sampling sites. This distance must take into account the structure of the vineyard. It corresponds to the shortest walking distance between two sites. If the points are in the same row, it corresponds to classical Euclidian distance. If they belong to different rows, the distance is computed considering that the practitioner has to leave the row, reach the desired rows passing by all extremities of intermediate rows and finally reach the targeted sampling site. As rows have two extremities, two different distances can be computed and the shortest one is kept.

To consider auxiliary data efficiently, the idea of *model sampling* proposed by Carillo et al. (2016) was considered. The *model sampling* approach aims at calibrating a linear regression which relates the yield

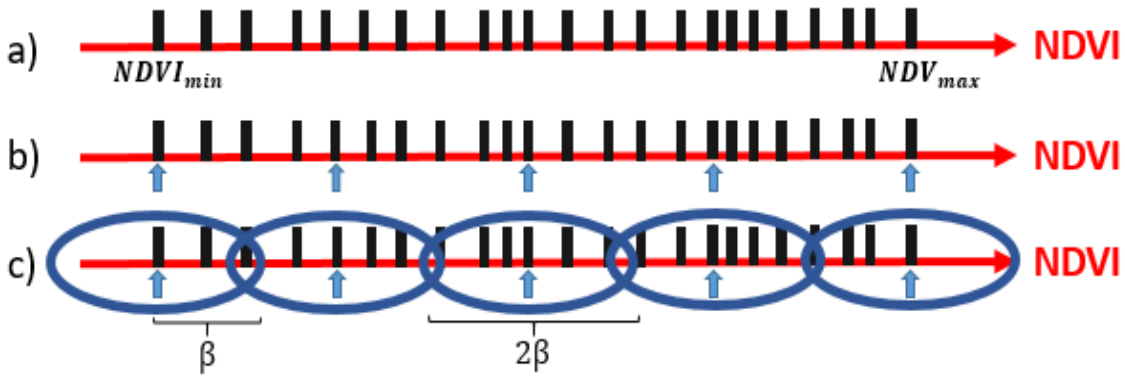

Figure A.2 : choice of potential sampling sites (for N = 5 sampling sites) based on NDVI values with method (iii) adapted from Kennard & Stone (1969); a) Distribution of observed NDVI values (each black dash represents one NDVI value). b) Selection of values corresponding to a potential sampling (arrows) and c) Groups of potential sampling sites built to account for the distribution of NDVI values, the width of groups is controlled by β

to auxiliary data (NDVI). This author also shown that yield components, especially berry weight, were linearly related to NDVI. Therefore, sampling sites can be selected representatively to build this linear model. This model is then used to estimate yield using all available high-resolution NDVI data.

The approach proposed in this paper relies on the following principle. Potential sampling sites are split into N homogeneous groups. Once groups are formed, one element from each group is selected in order to optimize the length of the route connecting all these points. It will ensure a good repartition of selected points by accounting for auxiliary data distribution, which is a key element to build a linear model with very few points (Kennard and Stone 1969). Three different ways to create these groups were tested:

The first method relies on quantiles. If K is the number of potential sampling points, then each group has $\frac{K}{N}$ elements. A first group will contain the $\frac{K}{N}$ potential sampling points with the lowest NDVI values. A second one the $\frac{K}{N}$ potential sampling sites with NDVI values just above those of the first group and the last group contains the $\frac{K}{N}$ highest NDVI values.

The second method uses the K-means algorithm. This clustering algorithm is efficient for partitioning K elements into N groups. The main principle of K-means is to minimize the difference between points in the same groups and thus maximise the difference between a group's mean.

The last method is derived from the Kennard & Stone (1969) approach. As described by these authors, this approach selects elements by iteratively choosing a new site that is the furthest from the sites already selected in terms of auxiliary data (Figure A.2.b). The approach was adapted to create groups, one centred around each value selected by Kennard & Stone approach. A parameter called β, expressed as a percentage of the NDVI range, set the width of the groups (Figure A.2.c). Using this method, groups may overlap with each other or, on the contrary, some sites may not belong to any group. This depends on the number of groups (N) and their length (β). Figure A.2 illustrates the method with potential sampling sites projected on an NDVI axis.

The second step of the approach consists in selecting one sampling site per group. These N sampling sites must be all different and have to form the shortest possible sampling circuit. There are numerous possible choices to select these sampling sites and many ways to order them to form a circuit. It is therefore a highly combinatorial optimization problem. Constraint Programming is one of the programming paradigms able to deal with such problems. It aims at solving a problem expressed as a set of variables and a set of constraints on these variables. Such a problem is called a Constraint

Satisfaction Problem (CSP). A Constraint Solver is used to find a solution to the problem that satisfies all the constraints. The efficiency of these solvers relies on the implementation of many methods such as filtering, which allows quick detection of combinations of values that do not lead to an optimal solution. The interest of constraint solvers lies in their ability to address many types of constraints.

Without going into small detail, let $S = \{1, \dots, K\}$ be the set of potential sampling sites and $\{G_i\}_{i\in(1,\dots,N)}$ the set of groups covering S, formed in the previous step. For decomposition (i) and (ii) all groups are disjoint and $\{G_i\}_{i\in(1,\dots,N)}$is a partition of S. P_i is defined as the selected site for group G_i. The first constraint imposes that all P_i must be different (this constraint is immediately satisfied in the case of methods (i) and (ii)). P_0 represents the point of departure and arrival; it is a fixed parameter representing the initial position of the practitioner. The length of the optimum route passing through all the $P_{i\in(0,\dots,N)}$ must be a minimum. This is a particular case of the vehicle routing problem (VRP) where the goal is not to find a Hamiltonian tour (visiting once every site) but a tour covering only a subset of sites. Recent work about the *WeightedSubCircuit* constraint (Vismara et al. 2018) has proposed a filtering algorithm that is well adapted to address this type of situation. All these constraints and variables constitute the constraint satisfaction problem. An instance of this problem is built from each dataset and solved with the solver in order to get an ordered set of sites that form a sampling circuit. The program returns the list of sampling sites, the order in which they are visited and the associated distance.

Yield estimation

The aim is to estimate Y, the average grape weight (GW) per vine. For each site selected by the sampling method ($s \in selected$), $GW(s)$, the observed grape weight per vine value, is available. A linear model linking the NDVI to GW is built from these sites (Eq. 1):

$$\widehat{GW}(s) = a \times NDVI(s) + b \qquad (1)$$

For a given site s, $\widehat{GW(s)}$ represents an estimate of $GW(s)$. The parameters a and b are obtained from a linear regression on the N sites selected by the sampling method.

With $S = \{1, \dots, K\}$ being the full set of potential sampling sites available, $\widehat{\overline{GW}}$, the estimate of $\overline{GW}$, can be computed from the model using all these potential sites (Eq. 2):

$$\hat{Y} = mean_{s\in S}(\widehat{GW}(s)) \qquad (2)$$

Estimation error

The estimation error is a deviation from the actual yield value (Y), expressed as a percentage (Eq. 3).

$$Error\ (\%) = \frac{\left|Y - \hat{Y}\right|}{Y} \qquad (3)$$

Reference methods

The method is compared to two references:

A conventional *random sampling* where the N sampling sites are randomly selected. $\widehat{\overline{GW}}$ is directly estimated from the mean of observed GW values. Here, $selected$ represent the N sampling sites chosen randomly (Eq. 4). *Random sampling* represents what is generally done in practice in terms of yield estimation

$$\hat{Y} = mean_{s\ \in\ selected}(GW(s)) \qquad (4)$$

A *model sampling* whose method principles have been described by Carillo et al. (2016). Sampling sites are chosen according to NDVI values. One site is randomly selected for each of the N NDVI quantiles. *Model sampling* uses a model based on the NDVI/yield relationship, as described in Eq. 1. The main difference is that *model sampling* does not consider the spatial position of the selected sampling sites.

To compare the length of the routes between the different methods, the optimal route between the selected sites must be computed for the reference methods. This is done with the Concorde TSP solver. As for *constraint sampling*, the route includes the starting point of the practitioner (P_0).

Experimental data

The data used to test the method came from INRA Pech-Rouge (Narbonne, France). The experiment and the database were detailed by Carrillo et al., (2016). It is briefly summarised hereafter. NDVI values from 9 different vine fields were considered. All of them are non-irrigated and exposed to Mediterranean climate with precipitation occurring during spring and with hot and dry summers. The characteristics of each plot are shown in Table 1.

Tableau A.1 : Description of the 9 fields.

Field (Id)	Area (ha)	Variety	Row Spacing (m)	Vine Spacing (m)	Potential Sampling Sites
P22	1.72	Syrah	2.5	1	45
P63	1.33	Syrah	2.5	1	42
P65	0.69	Syrah	2.5	1	33
P76	1.14	Carignan	2.25	1.5	37
P77	1.24	Syrah	2.5	1	19
P80	0.54	Syrah	2.5	1	40
P82	1.15	Syrah	2.5	1	53
P88	0.85	Syrah	2.25	1.5	21
P104	0.81	Carignan	2.25	1.5	23

NDVI values were derived from a 1 m. resolution multi-spectral image taken the 31th of August 2008 by Avion Jaune (Montpellier, Hérault, France). The spectral regions captured in the images were: blue (445–520 nm), green (510–600 nm), red (632–695 nm) and near-infrared (757–853 nm). From these, 1 m square image pixels, aggregation method described in Carillo et al. (2016) was used to obtain 9m square image pixels reducing the effect of canopy discontinuity and bare soil on measured values. NDVI was finally computed from processed images. Mechanical or chemical weeding was performed over the inter-row spacing; therefore, row cover crop did not affect NDVI values.

Sampling sites were selected regularly over the fields with measurement made on each node of a 15m² width sampling grid. At each node, yield components [bunch number per vine (BuN) and bunch weight (BuW)] were measured in 2009. Each site was considered as 5 consecutive vines in the row. BuW was

estimated at harvest by weighing 10 bunches (2 bunches per vine) also randomly taken from the same 5 consecutive vines. BuN was determined by counting all bunches of the 5 consecutive vines of each sampling point. Grape weight per vine (GW) was then calculated from BuW and BuN. The distance between vines along the row was 1m or 1.5m. Data were associated with the spatial co-ordinates of the central vine. The final data base was a set of 313 sites over the 9 different fields. The number of sites per field varied from 19 to 45 sampling sites. Each site was then characterized by GW as field parameter and NDVI values. NDVI value was assigned to each site as the mean of 4 pixels corresponding to a square of 36m².

Actual yield values measured at harvest are not available. For each field, the average of all available measured GW values is then used as the reference yield value (Y) when computing estimation error (Eq. 3).

Implementation

The core of the approach was written in java and used the Choco solver (Prud'homme et al. 2016). The calculations to obtain the distance matrix, groups of individuals classified according to their NDVI value, estimation errors and route distance were made with R.

As explained in the description of the constraints, the approach presented here takes into account the starting point of the practitioner (P_0) which is included in the sampling circuit. Varying the starting point thus changes the sampling route. In order to increase the number of situations tested, this starting point was positioned on different ends of row across the vineyards. The approach was then applied to 86 situations instead of 9.

Results and discussion

Evaluation of sampling strategies

Figure 2a and 2b show the results of estimation errors and sampling route distance observed for the different sampling methods. Remember that "i-quantile", "ii-kmeans", "iii-kennard β=10" and "iii-kennard β=15" refer to the *constraint sampling* methods (i.e. methods that account simultaneously for auxiliary data distribution and distance between sampling points) while *model sampling* and *random sampling* refers to methods that account only on auxiliary data distribution. "iii-kennard β=10" and "iii-kennard β=15" are based on the same approach with different group widths (β=10% & β=15%). Results for the different starting points are averaged for each field and then all together to give the same weight to each field.

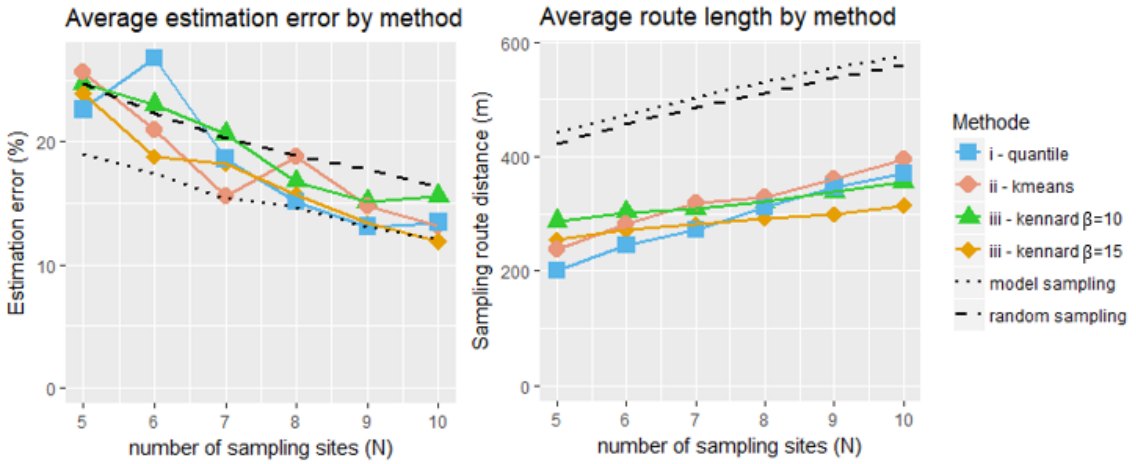

Figure A.3.a: (Estimation error %) and A.3.b. (Sampling route distance): Results and comparison to reference samplings

Figure A.3.a shows that all the methods follow the same trend with a decreasing error as the number of sampled sites increases. This result is logical, and consistent with the literature.

As Carrillo et al (2016) have already shown, taking into account auxiliary data (*model sampling*) slightly improves the quality of yield estimation compared to a *random sampling*. Despite higher variability in the observed error, the integration of constraints does not increase estimation errors, the methods (i), (ii), (iii) allow, in most cases, to maintain lower errors than *random sampling*. Kennard and Stone decomposition with β=15 may be the best option when creating the N groups, the results could match those of *model sampling* on most of the cases. Note however that observed errors with constraint methods are higher than for *model sampling* (i.e. without constraints). This result may be logical considering that the addition of the constraints may lead to the choice of less optimal sites within the groups of potential sites. Also, the irregularity of the curves associated with *constraint sampling* can be explained by a smaller number of experiments. These curves are based on 86 results compared to the 1,000 repetitions considered for reference methods, resulting in a higher variability.

Figure A.3.b clearly illustrates the gains brought by *constraint sampling* in terms of travel distance across the vineyard. Logically, the travelled distance within the plot increases linearly with N, the number of sampling sites visited. The four curves representing *constraint sampling* are at the same level, with a reduced distance of about 45% compared to *model sampling* and *random sampling*. Overall, this method offers a good compromise between the quality of the estimate and the travel constraint on the plot.

Applying the approach to new data could consolidate the results presented here. It would also be interesting to test the method with plots of vines having different characteristics (shape, size), under different cultivation practices (weed management between rows) or with different auxiliary data available (e.g.: historical yield).

This is a first model that can still be improved. Increasing the number of usable auxiliary variables or allowing the method to adjust directly as the first sites are selected for instance, could improve the accuracy and quality of the results. From an efficiency point of view, improvements in the Constraint Programming model could reduce computation times.

<u>Computation times</u>

Computation times increase with the number of possible combinations. The higher K and N (the number of potential sampling sites and the number of sites to be selected respectively), the longer it

will take. The way groups are created also affects the computational time. For instance, when using the Kennard and Stone approach to build a group, an increase of the β parameter (group width) can consistently increase computation times. In general, for plots with N<8, the computation times are in the order of a second. It took about a few hours in the most complex cases.

Conclusions

The methodology presented in this paper described a new approach for yield sampling in viticulture. The originality of the approach comes from the association of a previously published method based on auxiliary data and optimisation algorithms to propose relevant sampling routes in term of estimation error and travelled distance. While the *model sampling* principle guides sampling choice considering auxiliary information, optimisation through constraint programming ensures the relevancy of the chosen route in term of walking distance for the practitioner. Results presented here are of course preliminary results.

As available time is often the principal constraint for growers, they tend to rely on random samplings limited to a small part of the vineyard. Integrating spatial aspect accounting for travelling constraints is a key element to propose new methods that are relevant for field application. Further tests should be considered to confirm these first results and identify the limitations of the approach.

Acknowledgements

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (#Digitag).

Liste des figures :

Références :

Aarts, E. H. L., & Laarhoven, V. (1985). Statistical cooling: A general approach to combinatorial optimization problems, *Philips J. Res., 40(4), 193-226*

Abdelghafour, F., Keresztes, B., Germain, C., & Da Costa, J. P. (2017). Potential of on-board colour imaging for in-field detection and counting of grape bunches at early fruiting stages. *Advances in Animal Biosciences, 8(02), 505–509*. doi:10.1017/s2040470017001030

Adamchuk, V. I., Viscarra Rossel, R. A., Marx, D. B., & Samal, A. K. (2011). Using targeted sampling to process multivariate soil sensing data. Geoderma, 163(1-2), 63–73. doi:10.1016/j.geoderma.2011.04.004

Akers, N. P., & Vilar, E. (1986). RF sampling gates: a brief review. IEE Proceedings A Physical Science, Measurement and Instrumentation, *Management and Education, Reviews, 133(1), 45*. doi:10.1049/ip-a-1.1986.0006

Alrawashdeh, K., & Purdy, C. (2016). Toward an Online Anomaly Intrusion Detection System Based on Deep Learning*. 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).* doi:10.1109/icmla.2016.0040

Anderson, N. T., Underwood, J. P., Rahman, M. M., Robson, A., & Walsh, K. B. (2018). Estimation of fruit load in mango orchards: tree sampling considerations and use of machine vision and satellite imagery. *Precision Agriculture volume 20, p 823–839*. doi:10.1007/s11119-018-9614-1

Aquino, A., Barrio, I., Diago, M.-P., Millan, B., & Tardaguila, J. (2018). vitisBerry: An Android-smartphone application to early evaluate the number of grapevine berries by means of image analysis. *Computers and Electronics in Agriculture, 148, 19–28*. doi:10.1016/j.compag.2018.02.021

Araya-Alman, M., Acevedo-Opazo, C., Guillaume, S., Valdés-Gómez, H., Verdugo-Vásquez, N., Moreno, Y., & Tisseyre, B. (2017). Using ancillary yield data to improve sampling and grape yield estimation of the current season. *Advances in Animal Biosciences, 8(02), 515–519*. doi:10.1017/s2040470017000656

Araya-Alman, M., Leroux, C., Acevedo-Opazo, C., Guillaume, S., Valdés-Gómez, H., Verdugo-Vásquez, N., Pañitrur-De la Fuente, C., Tisseyre, B. (2019). A new localized sampling method to improve grape yield estimation of the current season using yield historical data. *Precision Agriculture*. doi:10.1007/s11119-019-09644-y

Austin, R. B., & Blackwell, R. D. (1980). Edge and neighbour effects in cereal yield trials. *The Journal of Agricultural Science, 94(03), 731*. doi:10.1017/s0021859600028720

Bachmaier, M., & Backes, M. (2008). Variogram or semivariogram? Understanding the variances in a variogram*. Precision Agriculture, 9(3), 173–175.* doi:10.1007/s11119-008-9056-2

Bannari, A., Morin, D., Bonn, F., & Huete, A. R. (1995). A review of vegetation indices. *Remote Sensing Reviews, 13(1-2), 95–120.* doi:10.1080/02757259509532298

Barnes, E. M. & Baker., M. G. (2000). Multispectral data for mapping soil texture: Possibilities and limitations. *Applied Engineering in Agriculture, 16(6), 731–741*. doi:10.13031/2013.5370

Basso, B., Ritchie, J. T., Pierce, F. J., Braga, R. P., & Jones, J. W. (2001). Spatial validation of crop models for precision agriculture. *Agricultural Systems, 68(2), 97–112*. doi:10.1016/s0308-521x(00)00063-9

Binns, M. R., Nyrop, J. P., & van der Werf, W. (2000). Sampling and monitoring in crop protection: The theoretical basis for developing practical decision guides. *CABI Publishing.*

Bramley, R.G.V. and Hamilton, R.P. (2004). Understanding variability in winegrape production systems. 1. Within vineyard variation in yield over several vintages. *Australian Journal of Grape and Wine Research, 10, 32-45.*

Bramley, R. G. V., Ouzman, J., & Boss, P. K. (2011). Variation in vine vigour, grape yield and vineyard soils and topography as indicators of variation in the chemical composition of grapes, wine and wine sensory attributes. *Australian Journal of Grape and Wine Research, 17(2), 217–229.* doi:10.1111/j.1755-0238.2011.00136.x

Bramley, R. G. V., Ouzman, J., Trought, M. C. T., Neal, S. M., & Bennett, J. S. (2019). Spatio-temporal variability in vine vigour and yield in a Marlborough Sauvignon Blanc vineyard. *Australian Journal of Grape and Wine Research, 25(4*), 430-438.

Briot, N., Bessiere, C., Tisseyre, B. & Vismara, P. (2015). Integration of Operational Constraints to Optimize Differential Harvest in Viticulture. *Proceed. 10th European Conference on Precision Agriculture (ECPA 2015), 487–494.*

Brown, C. M., & Weibe, R. O. (1957). Border Effects in Winter Wheat and Spring Oat Tests1. *Agronomy Journal, 49(7), 382.* doi:10.2134/agronj1957.00021962004900070001

Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma, 80(1-2), 1–44*. doi:10.1016/s0016-7061(97)00072-4

Carmona, M.J., Chaïb, J., Martínez-Zapater, J.M., Thomas, M.R. (2008). A molecular genetic perspective of reproductive development in grapevine. *J Exp Bot 59, 2579–2596*. doi:10.1093/jxb/ern160

Carrillo, E., Matese, A., Rousseau, J., & Tisseyre, B. (2016). Use of multi-spectral airborne imagery to improve yield sampling in viticulture. *Precision Agriculture, 17(1*), 74-92.

Chang, J., Clay, D. E., Carlson, C. G., Clay, S. A., Malo, D. D., Berg, R., … Wiebold, W. (2003). Different Techniques to Identify Management Zones Impact Nitrogen and Phosphorus Sampling Variability. *Agronomy Journal, 95(6), 1550*. doi:10.2134/agronj2003.1550

Chen, D., Lu, C.-T., Kou, Y., & Chen, F. (2007). On Detecting Spatial Outliers. *GeoInformatica, 12(4), 455–475*. doi:10.1007/s10707-007-0038-8

Chen, Z., Bai, Z. D., Sinha, B. K. (2004). Ranked Set Sampling: Theory and Applications. *Springer*

Chloupek, O., Hrstkova, P., & Schweigert, P. (2004). Yield and its stability, crop diversity, adaptability and response to climate change, weather and fertilisation over 75 years in the Czech Republic in comparison to some European countries. *Field Crops Research, 85(2-3), 167–190.* doi:10.1016/s0378-4290(03)00162-x

Clingeleffer, P., Dunn, G.M., Krstic, M., Martin, S. (2001). Crop Development, Crop Estimation and Crop Control to Secure Quality and Production of Major Wine Grape Varieties: A National Approach. *Grape and Wine Resarch & Development Corporation.*

Cogato, A., Pagay, V., Marinello, F., Meggio, F., Grace, P., & De Antoni Migliorati, M. (2019). Assessing the Feasibility of Using Sentinel-2 Imagery to Quantify the Impact of Heatwaves on Irrigated Vineyards. *Remote Sensing, 11(23), 2869*. doi:10.3390/rs11232869

Cristofolini, F. & Gottardini, E. (2000). Concentration of airborne pollen of Vitisvinifera L. and yield forecast: a case study at S.Michele all'Adige, Trento, Italy. *Aerobiologia, 16(1), 125–129.*

Dell, T. R., & Clutter, J. L. (1972). Ranked Set Sampling Theory with Order Statistics Background. *Biometrics, 28(2), 545.* doi:10.2307/2556166

Devaux, n., Crestey, t., Leroux, C., Tisseyre B. (2019). Potential of Sentinel-2 satellite images to monitor vine fields grown at a territorial scale. *Vol. 53 No. 1 (2019): OENO One, 53, 1*

Diago, M.-P., Correa, C., Millán, B., Barreiro, P., Valero, C., & Tardaguila, J. (2012). Grapevine Yield and Leaf Area Estimation Using Supervised Classification Methodology on RGB Images Taken under Field Conditions. *Sensors, 12(12), 16988–17006*. doi:10.3390/s121216988

Dorigo, M. and Gambardella, L. M. (1997). Ant colonies for the travelling salesman problem. Biosystems, Volume 43, Issue 2, 1997, 73-81, doi:10.1016/S0303-2647(97)01708-5.

Dry, P.R. (2000). Canopy management for fruitfulness. *Australian Journal of Grape and Wine Research 109–115.*

Dunn, G., & Martin, S. (2004). Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest. *Australian Journal of Grape and Wine Research, 10(3), 196–198*. doi:10.1111/j.1755-0238.2004.tb00022.x

Elfil, M., Negida, A. (2017). Sampling methods in Clinical Research; an Educational Review. *Emergency. 2017; 5 (1): e52*

FAO (2010) Handbook on Master Sampling Frames for Agricultural Statistics Frame Development, Sample Design and Estimation. *p 63-70*

Filzmoser, P., Ruiz-Gazen, A., & Thomas-Agnan, C. (2013). Identification of local multivariate outliers. *Statistical Papers, 55(1), 29–47*. doi:10.1007/s00362-013-0524-z

Fleischer, S. J., Blom, P. E., & Weisz, R. (1999). Sampling in Precision IPM: When the Objective Is a Map. *Phytopathology, 89(11), 1112–1118*. doi:10.1094/phyto.1999.89.11.1112

Fortes, R., Millán, S., Prieto, M. H., & Campillo, C. (2015). A methodology based on apparent electrical conductivity and guided soil samples to improve irrigation zoning. *Precision Agriculture, 16(4), 441–454*. doi:10.1007/s11119-015-9388-7

Gebremeskel, G. B., Yi, C., He, Z., & Haile, D. (2016). Combined data mining techniques based patient data outlier detection for healthcare safety. *International Journal of Intelligent Computing and Cybernetics, 9(1), 42–68*. doi:10.1108/ijicc-07-2015-0024

Gozdowski, D., Samborski, S., & Dobers, E. S. (2010). Evaluation of methods for the detection of spatial outliers in the yield data of winter wheat. *Colloquium Biometricum, 40, 41–51.*

González-Fernández, A. B., Rodríguez-Pérez, J. R., Sanz-Ablanedo, E., Valenciano, J. B., & Marcelo, V. (2019). Delineating vineyard zones by fuzzy K-means algorithm based on grape sampling variables. *Scientia Horticulturae, 243, 559–566*. doi:10.1016/j.scienta.2018.09.012

Grimm, J., Herzog, K., Rist, F., Kicherer, A., Töpfer, R., & Steinhage, V. (2019). An adaptable approach to automated visual detection of plant organs with applications in grapevine breeding. *Biosystems Engineering, 183, 170–183*. doi:10.1016/j.biosystemseng.2019.04.018

Guilpart, N., Roux, S., Gary, C., & Metay, A. (2017). The trade-off between grape yield and grapevine susceptibility to powdery mildew and grey mould depends on inter-annual variations in water stress. *Agricultural and Forest Meteorology, 234-235, 203–211*. doi:10.1016/j.agrformet.2016.12.023

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 26(9), 2250–2267.* doi:10.1109/tkde.2013.184

Hall, D. G., Childers, C. C., & Eger, J. E. (2007). Binomial Sampling to Estimate Rust Mite (Acari: Eriophyidae) Densities on Orange Fruit. *Journal of Economic Entomology, 100(1), 233–240.* doi:10.1093/jee/100.1.233

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., & Clarke, A. (2013). Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods. *Mathematical Geosciences, 46(1), 1–31*. doi:10.1007/s11004-013-9491-0

Hahsler M, Hornik K (2007). TSP – Infrastructure for the Traveling Salesperson Problem. Journal of *Statistical Software, 23(2), 1–21.*

Herrero-Langreo, A., Tisseyre, B., Roger, J. M., & Scholasch, T. (2017). Test of sampling methods to optimize the calibration of vine water status spatial models. *Precision Agriculture, 19(2), 365–378*. doi:10.1007/s11119-017-9523-8

Hodgson, E. W., Burkness, E. C., Hutchison, W. D., & Ragsdale, D. W. (2004). Enumerative and Binomial Sequential Sampling Plans for Soybean Aphid (Homoptera: Aphididae) in Soybean. *Journal of Economic Entomology, 97(6), 2127–2136*. doi:10.1093/jee/97.6.2127

Howell, G.S. (2001). Sustainable grape productivity and the growth-yield relationship: A review. *American Journal of Enology and Viticulture 52, 165–174.*

Huddleston, H. F. (1978). Sampling techniques for measuring and forecasting crop yields. *Economics, Statistics and Cooperatives Service, u.s. Department of Agriculture. ESCS No. 09*

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies Volume 19, Issue 4, December 2000, Pages 577-613.* doi:10.1016/s0261-3794(99)00028-1

Jordan, C., Shi, Z., Bailey, J. S., & Higgins, A. J. (2003). Sampling Strategies for Mapping 'Within-field' Variability in the Dry Matter Yield and Mineral Nutrient Status of Forage Grass Crops in Cool Temperate Climes. *Precision Agriculture, 4(1), 69-86*. doi:10.1023/a:1021815122216

Jung, C., Lee, Y., Lee, J., & Kim, S. (2020). Performance Evaluation of the Multiple Quantile Regression Model for Estimating Spatial Soil Moisture after Filtering Soil Moisture Outliers. *Remote Sensing, 12(10), 1678*. doi:10.3390/rs12101678

Justeau-Allaire, D., Vismara, P., Birnbaum, P., & Lorca, X. (2019). Systematic Conservation Planning for Sustainable Land-use Policies: A Constrained Partitioning Approach to Reserve Selection and Design. Proceedings of the Twenty-Eighth International Joint Conference on Artificial

Intelligence. *AI for Improving Human Well-being. Pages 5902-5908.* doi:10.24963/ijcai.2019/818

Kagan, A. (2001). Another Look at the Cramér-Rao Inequality. *The American Statistician, 55(3), 211–212.* doi:10.1198/000313001317098194

Kasimatis, A. N., Vilas, E. P. (1985). Sampling for Degrees Brix in Vineyard Plots. A*m J Enol Vitic. January 1985 36: 207-213*

Kennard, R.W. and Stone, L.A. (1969) Computer-aided Design of Experiments. *Technometrics, 11, 137-148*. doi:10.1080/00401706.1969.10490666

Kerry, R., Oliver, M. A., & Frogbrook, Z. L. (2010). Sampling in Precision Agriculture. *Geostatistical Applications for Precision Agriculture, 35–63*. doi:10.1007/978-90-481-9133-8_2

Kerry, R., & Oliver, M. A. (2003). Variograms of Ancillary Data to Aid Sampling for Soil Surveys*. Precision Agriculture, 4(3), 261–278*. doi:10.1023/a:1024952406744

Kou, Y., Lu, C.-T., & Chen, D. (2006). Spatial Weighted Outlier Detection. *Proceedings of the 2006 SIAM International Conference on Data Mining, 614–618*. doi:10.1137/1.9781611972764.71

Krstic, M.P., Welsh, M.A. and Clingeleffe, P.R. 1998. Variation in Chardonnay yield components between vineyards in a warm irrigated region. *In: R.J. Blair, A.N. Sas, P.F. Hayes, and P.B. Hoj (eds). AWRI, Urrbrae, SA, Sydney, Australia, 269-270*

Kruskal, W., & Mosteller, F. (1979). Representative Sampling, II: Scientific Literature, Excluding Statistics*. International Statistical Review / Revue Internationale de Statistique, 47(2), 111.* doi:10.2307/1402564

Lachia, N., Pichon, L., Tisseyre, B., (2019). A collective framework to assess the adoption of precision agriculture in France: description and preliminary results after two years. *In: Precision agriculture'19., ED. John v. Stafford, Ampthill, UK, Wageningen Academic Publishers. 851-857.*

Lameck, W. U. (2013). Sampling Design, Validity and Reliability in General Social Survey. *International Journal of Academic Research in Business and Social Sciences, 212-218.* doi:10.6007/IJARBSS/v3-i7/27

Laplace, P. S. (1809). Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres, et sur leur application aux probabilités », *Mémoires de la Classe des sciences mathématiques et physiques de l'Institut de France,1809, p. 353-415*

Lawler, E. L., Shmoys, D. B., Kan, A. H. G. Rinnooy, Lenstra, J. K. (1985). The Traveling Salesman Problem. *John Wiley & Sons, Incorporated.*

Le Maire, G. ,François, C., Soudani, K., Davi, H., Le Dantec, V., Saugier, B. (2006). Forest leaf area index determination : a multiyear satellite independent method based on within-stand normalized difference vegetation index spatial variability. *Journal of Geophysical Research, American Geophysical Union, 2006, 111 (G2),*

Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., & Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture, 19(5), 789–808.* doi:10.1007/s11119-017-9555-0

Leroux, C., Jones, H., Clenet, A., & Tisseyre, B. (2018). Knowledge discovery and unsupervised detection of within-field yield defective observations. *Computers and Electronics in Agriculture, 156, 645–659.* doi:10.1016/j.compag.2018.12.024

Lessio, F., Tota, F., & Alma, A. (2014). Tracking the dispersion of Scaphoideus titanus Ball (Hemiptera: Cicadellidae) from wild to cultivated grapevine: use of a novel mark–capture technique. *Bulletin of Entomological Research, 104(04), 432–443*. doi:10.1017/s0007485314000030

Li, L., Mu, X., Macfarlane, C., Song, W., Chen, J., Yan, K., & Yan, G. (2018). A half-Gaussian fitting method for estimating fractional vegetation cover of corn crops using unmanned aerial vehicle images. *Agricultural and Forest Meteorology, 262, 379–390*. doi:10.1016/j.agrformet.2018.07.028

Li, T., Hao, X., Kang, S., & Leng, D. (2017). Spatial Variation of Winegrape Yield and Berry Composition and their Relationships to Spatiotemporal Distribution of Soil Water Content. *American Journal of Enology and Viticulture, 68(3), 369–377.*

Liaghat, S. & Balasundram, S.K. (2010). A Review: The Role of Remote Sensing in Precision Agriculture. *American Journal of Agricultural and Biological Sciences, 5(1), 50–55.* doi:10.3844/ajabssp.2010.50.55

Liu, F. T. , Ting K. M., & Zhou, Z. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining, Pisa, 2008, pp. 413-422*, doi: 10.1109/ICDM.2008.17.

Liu, S., Li, X., Wu, H., Xin, B., Tang, J., Petrie, P. R., & Whitty, M. (2018). A robust automated flower estimation system for grape vines. *Biosystems Engineering, 172, 110–123.* doi:10.1016/j.biosystemseng.2018.05.009

Lu, C.-T., Chen, D., & Kou, Y. (n.d.). Algorithms for spatial outlier detection. *Third IEEE International Conference on Data Mining*. doi:10.1109/icdm.2003.1250986

Lutton, J.L., & Bonomi, E.; « Le recuit simulé », *Pour la science, no 129, juillet 1988, p. 68- 77.*

Machado, R. C., Andrade, D. F., Babos, D. V., Castro, J. P., Costa, V., Sperança, M., … Pereira-Filho, E. R. (2019). Solid sampling: advantages and challenges in atomic spectrometry — a critical review. *Journal of Analytical Atomic Spectrometry*. doi:10.1039/c9ja00306a

Magnani, R., Sabin, K., Saidel, T., & Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS, 19(Supplement 2), S67–S72.* doi:10.1097/01.aids.0000172879.20628.e1

Mandić-Rajčević, & Colosio. (2019). Methods for the Identification of Outliers and Their Influence on Exposure Assessment in Agricultural Pesticide Applicators: A Proposed Approach and Validation Using Biological Monitoring. *Toxics, 7(3), 37.* doi:10.3390/toxics7030037

Mann, J. (1999). Behavioral Sampling Methods for Cetaceans: A Review and Critique. Marine Mammal *Science, 15(1), 102–122*. doi:10.1111/j.1748-7692.1999.tb00784.x

Matese, A., Di Gennaro, S. F., & Berton, A. (2016). Assessment of a canopy height model (CHM) in a vineyard using UAV-based multispectral imaging. *International Journal of Remote Sensing, 38(8-10), 2150–2160.* doi:10.1080/01431161.2016.1226002

Meyers, J.M., Sacks, G.L., van Es, H.M., & Vanden Heuvel, J.E. (2011). Improving vineyard sampling efficiency via dynamic spatially explicit optimisation. *Australian Journal of Grape and Wine Research, 17(3), 306–315.* doi:10.1111/j.1755-0238.2011.00152.x

Meyers, J. M., & Vanden Heuvel, J. E. (2014). Use of Normalized Difference Vegetation Index Images to Optimize Vineyard Sampling Protocols. *American Journal of Enology and Viticulture, 65(2), 250–253.* doi:10.5344/ajev.2014.13103

Meyers, J. M., Dokoozlian, N., Ryan, C., Bioni, C., & Vanden Heuvel, J. E. (2020). A New, Satellite NDVI-Based Sampling Protocol for Grape Maturation Monitoring. *Remote Sensing, 12(7), 1159*. doi:10.3390/rs12071159

Miranda, C., Santesteban, L., Urrestarazu, J., Loidi, M., & Royo, J. (2018). Sampling Stratification Using Aerial Imagery to Estimate Fruit Load in Peach Tree Orchards. *Agriculture, 8(6), 78.* doi:10.3390/agriculture8060078

Murthy, C. S., Thiruvengadachari, S., Raju, P. V., & Jonna, S. (1996). Improved ground sampling and crop yield estimation using satellite data. *International Journal of Remote Sensing, 17(5), 945–956.* doi:10.1080/01431169608949057

Nuske, S., Achar, S., Bates, T., Narasimhan, S., & Singh, S. (2011). Yield estimation in vineyards by visual grape detection*. 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems.* doi:10.1109/iros.2011.6095069

Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Narasimhan, S., & Singh, S. (2014). Automated Visual Yield Estimation in Vineyards. *Journal of Field Robotics, 31(5), 837–860*. doi:10.1002/rob.21541

Oger, B., Vismara, P. & Tisseyre, B. (2018). Combining target sampling with route-optimization to optimise yield estimation in viticulture*. Proceed. 12th European Conference on Precision Agriculture (ECPA 2019)*, 487–494.

Papadimitriou, C. H. (1977). The Euclidean travelling salesman problem is NP-complete », *Theoretical Computer Science, vol. 4, no 3, 1977, p. 237–244.*

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016). Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).* doi:10.1109/icmla.2016.0172

Pierce, F. J., & Nowak, P. (1999). Aspects of Precision Agriculture. *Advances in Agronomy Volume 67, 1–85*. doi:10.1016/s0065-2113(08)60513-1

Porwal, U., Mukund., S., (2018).Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach. *Mathematics, Computer Science. ArXiv*

Pothen, Z., & Nuske, S. (2016). Automated Assessment and Mapping of Grape Quality through Image-based Color Analysis*. IFAC-PapersOnLine, 49(16), 72–78*. doi:10.1016/j.ifacol.2016.10.014

Prud'homme C., Fages J.G. & Lorca X. (2016). Choco Documentation. *TASC, INRIA Rennes, LINA CNRS UMR 6241, COSLING S.A.S*, http://www.choco-solver.org

Rahmani, A. R., Leili, M., Azarian, G., & Poormohammadi, A. (2020). Sampling and detection of corona viruses in air: A mini review. *Science of The Total Environment, 740, 140207.* doi:10.1016/j.scitotenv.2020.140207

Ramsey, C. A., & Hewitt, A. D. (2005). A Methodology for Assessing Sample Representativeness. *Environmental Forensics, 6(1), 71–75*. doi:10.1080/15275920590913877

Rehman, A. U., Abbasi, A. Z., Islam, N., & Shaikh, Z. A. (2014). A review of wireless sensors and networks' applications in agriculture. *Computer Standards & Interfaces, 36(2), 263–270.* doi:10.1016/j.csi.2011.03.004

Reis, M. J. C. S., Morais, R., Peres, E., Pereira, C., Contente, O., Soares, S., … Bulas Cruz, J. (2012). Automatic detection of bunches of grapes in natural environment from color images. *Journal of Applied Logic, 10(4), 285–290.* doi:10.1016/j.jal.2012.07.004

Roscher, R., Herzog, K., Kunkel, A., Kicherer, A., Töpfer, R., & Förstner, W. (2014). Automated image analysis framework for high-throughput determination of grapevine berry sizes using conditional random fields. *Computers and Electronics in Agriculture, 100, 148–158.* doi:10.1016/j.compag.2013.11.008

Rossi, F. (2006). Handbook of Constraint Programming 1st Edition. *Elsevier Science.*

Rouse, J. W. Jr., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the great plains with ERTS. In S. C. Freden, E. P. Mercanti, & M. A. Becker (Eds.), *Proceedings of the Third ERTS Symposium, NASA SP-351 1, 309–317*

Sampford, M. R. (1962). An introduction to sampling theory with applications to agriculture. *292 pp. Oliver and Boyd, Ltd., Edinburgh and London 1962.*

Särndal, C. E., Swensson, B., Wretman, J. (1992). Model Assisted Survey Sampling. *Springer Series in Statistics. p 225-230*

Serrano, E., Roussel, S., Gontier, L. & Dufourcq, T. (2005). Estimation précoce du rendement de la vigne : corrélation entre le volume de la grappe de vitis vinifera en cours de croissance et son poids à la récolte (Early grape yield estimation: correlation between the volume of the cluster of vitis vinifera during growth and harvest weight), *In: H. Schultz (Ed.) Proceeding of the Groupe Européen d'Etude des Systémes de Conduite de la Vigne, give publ location: publisher of proceedings (pp. 311-318).*

Sheaffer, C. C., Moncada, K. M. (2012). Introduction to Agronomy: Food, Crops, and Environment, second edition. *Cengage Learning.*

Schauss P., Deville Y., Dupont P., Régin J.-C. (2007). The deviation constraint. *In: CPAIOR'07, pp 260–274*

Skakun, S., Franch, B., Vermote, E., Roger, J.-C., Becker-Reshef, I., Justice, C., & Kussul, N. (2017). Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment, 195, 244–258.* doi:10.1016/j.rse.2017.04.026

Sommer, K. J., Islam, M., & Clingeleffer, P. R. (2001). Sultana fruitfulness and yield as influenced by season, rootstock and trellis type. *Australian Journal of Grape and Wine Research, 7(1), 19–26.* doi:10.1111/j.1755-0238.2001.tb00189.x

Sozzi, M., Kayad, A., Marinello, F., Taylor, J., Tisseyre B. (2020). Comparing vineyard imagery acquired from Sentinel-2 and Unmanned Aerial Vehicle (UAV) platform. *Vol. 54 No. 2 (2020): OENO one*

Spezia, G. R., Souza, E. G. I; Nóbrega, L. H. P., Miguel A. Uribe-Opazo, M. A., Milan, M.,Bazzi, C. L. (2012). Model to estimate the sampling density for establishment of yield mapping. *Rev. bras. eng. agríc. ambient. vol.16 no.4 Campina Grande Apr. 2012.* doi:10.1590/S1415-43662012000400016

Stutzle, T. and Hoos, H. (1997). MAX-MIN Ant System and local search for the traveling salesman problem, *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC '97)*, Indianapolis, IN, USA, 1997, pp. 309-314, doi:10.1109/ICEC.1997.592327.

Taylor J., Tisseyre B., Bramley R., Reid A., (2005). A comparison of the spatial variability of vineyard yield in European and Australian production systems*. Proceed. 5th European Conference on Precision Agriculture (ECPA 2005), 907-914.*

Taylor, J. A., & Bates, T. R. (2013). Temporal and spatial relationships of vine pruning mass in Concord grapes. *Australian Journal of Grape and Wine Research*. doi:10.1111/ajgw.12035

Thompson, S. K. (2012). Sampling, Third Edition. *John Wiley & Sons, Inc. p 1-8* doi:10.1002/9781118162934

Tisseyre, B., Leroux, C., Pichon, L., Geraudie, V., & Sari, T. (2018). How to define the optimal grid size to map high resolution spatial data? *Precision agriculture, 19(5), 957-971.*

Torres, A. B. B., Filho, J. A., Rocha, A. R., Gondim, R. S., Souza, J. N. (2017). Outlier detection methods and sensor data fusion for precision agriculture. *ANAIS DO IX SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO UBÍQUA E PERVASIVA.*

Uribeetxebarria, A., Martínez-Casasnovas, J. A., Tisseyre, B., Guillaume, S., Escolà, A., Rosell-Polo, J. R., & Arnó, J. (2019). Assessing ranked set sampling and ancillary data to improve fruit load estimates in peach orchards*. Computers and Electronics in Agriculture, 164, 104931.* doi:10.1016/j.compag.2019.104931

Vasconcelos, M.C., Greven, M., Winefield, C.S., Trought, M.C., Raw, V., 2009. The flowering process of Vitis vinifera: a review. *American Journal of Enology and Viticulture 60, 411–434.*

Vega, A., Córdoba, M., Castro-Franco, M., & Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*. doi:10.1007/s11119-018-09632-8

Venkataratnam, L. (2001). Remote sensing and GIS inagricultural resources management. *Proceedings of the 1st National Conference on Agro-Informatics, June 3-4, Dharwad, India, pp: 20-29.*

Vismara P. & Briot N. (2018). A Circuit Constraint for Multiple Tours Problems. Proceed. 24th International Conference on Principles and Practice of Constraint Programming (CP 2018). *Lecture Notes in Computer Science. Vol.11008, 389–402.*

Walck, C. (2007). Hand-book on statistical distributions for experimentalists. *Stockholm: University of Stockholm.*

Wang, Z., Zhao, X., Wu, P., Gao, Y., Yang, Q., & Shen, Y. (2017). Border row effects on light interception in wheat/maize strip intercropping systems. *Field Crops Research, 214, 1–13.* doi:10.1016/j.fcr.2017.08.017

Wasserman, L. (2004). All of statistics: A concise course in statistical inference. *New York: Springer.*

Werra, D., Liebling, T. M., & Hêche J.F. (2003) Recherche opérationnelle pour ingénieurs. *Presses polytechniques et universitaires romandes.*

Whelan, B. M., & McBratney, A. B. (2000). The ‘‘Null Hypothesis’’ of Precision Agriculture Management. *Precision Agriculture, 2(3), 265–279.* doi:10.1023/a:1011838806489

Wolpert, J.A., & Vilas, E.P. (1992). Estimating vineyard yields: Introduction to a simple, two-step method. *American Journal of Enology and Viticulture, 43, 384-388*

Wulfsohn, D. (2010). Sampling Techniques for Plants and Soil. *Landbauforschung Völkenrode, Special Issue 340, 2010.*

Yemefack, M., Rossiter, D. G., & Njomgang, R. (2005). Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern Cameroon*. Geoderma, 125(1-2), 117–143.* doi:10.1016/j.geoderma.2004.07.007

Résumé :

En production végétale, les pratiques culturales (gestion du semis, des intrants, de la récolte etc.) sont raisonnées à partir de l'information dont disposent les agriculteurs. En ce sens, les méthodes d'estimation par échantillonnage constituent un ensemble d'outils essentiels dans l'acquisition d'informations à l'échelle de la parcelle, unité de gestion des cultures, afin de mieux raisonner les intrants. En se basant sur un échantillon d'observations réalisées sur quelques sites de mesures, ces méthodes permettent d'estimer par inférence les propriétés de la distribution des valeurs pour l'ensemble de la parcelle. Des travaux récents proposent de nouvelles méthodes d'échantillonnage, intégrant des données auxiliaires à haute résolution spatiale telles que les images de télédétection obtenues par drone, par avion ou par satellite. Ces données sont accessibles à moindre coût et fournissent une information exhaustive de la variabilité spatiale des parcelles susceptible de permettre d'améliorer la précision de l'estimation en améliorant le choix et le positionnement des sites d'observation. Différents travaux scientifiques ont proposé des approches d'échantillonnage orientées sur la base de données auxiliaires en agriculture. Bien que proposant des résultats intéressants, ces méthodes ne prennent cependant pas en compte l'effort d'échantillonnage (temps, distance) qu'il est nécessaire de fournir pour accéder aux sites de mesure et sont donc difficilement applicables dans un contexte de production. L'objectif de cette thèse est de proposer de nouveaux outils pour l'échantillonnage en production végétale basés sur l'utilisation de données auxiliaires, afin de répondre au double enjeu de la précision de l'estimation et de l'effort d'échantillonnage. Pour répondre à ces enjeux, plusieurs contraintes et critères, statistiques et agronomiques, sont identifiés. Les parcours d'échantillonnage sont ensuite optimisés et sélectionnés au regard des critères retenus. L'originalité de ces travaux se trouve dans la résolution d'un problème agronomique d'échantillonnage en combinant les méthodologies issues de deux domaines scientifiques très différents : des approches stochastiques visant à caractériser un grand nombre de candidats (les sites potentiels d'échantillonnage) ainsi que des approches d'optimisation informatique comme la programmation par contraintes ou les algorithmes de recherche opérationnelle pour identifier un échantillon solution parmi un grand nombre de possibilités. Afin de mieux comprendre et caractériser l'apport des approches stochastiques au regard de l'optimisation dans l'approche hybride proposée, la thèse a également étudié l'apport des données auxiliaires pour l'inférence de l'estimation. L'inférence basée sur l'étalonnage d'un modèle apparaît comme un des leviers important pour réduire la variabilité de l'estimation et l'imprécision qui en résulte. Les propriétés statistiques des estimateurs utilisés pour l'inférence peuvent alors être utilisées pour guider le choix des sites de mesure afin de proposer une approche cohérente. La validation des outils proposés est basée sur le cas d'étude de l'estimation du rendement en viticulture où la structure palissée (structurée en rangs infranchissables) contraint fortement les parcours d'échantillonnage. Les types de parcours obtenus via ces approches sont décrits et comparés aux pratiques de la profession viticole pour ouvrir sur quelques considérations faisant le lien entre stratégies d'échantillonnage et caractéristiques de parcelles (organisation spatiale de la variabilité, forme de la parcelle, orientation des rangs, etc.). A terme, l'approche proposée pourrait constituer un outil d'aide à la décision permettant d'adapter l'échantillonnage à chaque parcelle dans l'objectif d'améliorer simultanément la qualité de la prédiction avec le minimum d'effort d'échantillonnage.

Abstract:

In crop production, cultural practices (management of sowing, inputs, harvesting, etc.) are reasoned based on the information available to farmers. Accordingly, estimation methods by sampling constitute a set of essential tools in the acquisition of information at the scale of the plot, the crop management unit, in order to better reason inputs. Based on a sample of observations carried out at a few measurement sites, these methods allow inferential estimation of the properties of the distribution of values for the entire plot. Recent work proposes new sampling methods, integrating auxiliary data with high spatial resolution such as remote sensing images obtained by UAV, aircraft or satellite. These data are accessible at a lower cost and provide exhaustive information on the spatial variability of the plots that can improve the accuracy of the estimate by improving the choice and positioning of observation sites. Various scientific works have proposed approaches to sampling oriented on the auxiliary database in agriculture. Although providing interesting results, these methods do not take into account the sampling effort (time, distance) required to access the measurement sites and are therefore difficult to apply in a production context. The objective of this thesis is to propose new tools for sampling in crop production based on the use of auxiliary data, in order to address the double issue of the precision of the estimate and the sampling effort. To meet these challenges, several constraints and criteria, both statistical and agronomic, are identified. The sampling routes are then optimized and selected according to the chosen criteria. The originality of this work lies in the resolution of an agronomic sampling problem by combining methodologies from two very different scientific fields: stochastic approaches aimed at characterizing a large number of candidates (potential sampling sites) as well as computer optimization approaches such as constraint programming or operational research algorithms to identify a solution sample among a large number of possibilities. In order to better understand and characterize the contribution of stochastic approaches to optimization in the proposed hybrid approach, the thesis also studied the contribution of auxiliary data for the inference of the estimation. The inference based on the calibration of a model appears to be important leverage to reduce the variability of the estimate and the resulting imprecision. The statistical properties of the estimators used for inference can then be used to guide the choice of measurement sites in order to propose a consistent approach. The validation of the proposed methods is based on the case study of yield estimation in viticulture where the trellised structure (structured in uncrossable rows) strongly constrains the sampling routes. The types of routes obtained through these approaches are described and compared to the practices of the viticulture profession to open up a few considerations linking sampling strategies and plot characteristics (spatial organization of variability, plot shape, row orientation, etc.). Ultimately, the proposed approach could constitute a decision support tool allowing to adapt sampling to each plot with the objective of simultaneously improving the quality of prediction with the minimum sampling effort.