Modélisation spatio-temporelle pour l'esca de la vigne à l'échelle de la parcelle

Shuxian Li

Thèse

pour l'obtention du grade de

Docteur de l'Université de Bordeaux

École Doctorale Sciences et Environnement

Shuxian LI

Modélisation spatio-temporelle pour l'ESCA de la vigne à l'échelle de la parcelle

---

sous la direction de : **Lucia Guérin-Dubrana et Anne Gégout-Petit**

Soutenue le 16 Decembre 2015

Devant la commission d'examen :

M. Benoit Marçais	Directeur de Recherche	INRA Nancy	Rapporteur
Mme. Liliane Bel	Professeur	AgroParisTech	Rapporteur
Mme. Isabelle Cousin	Directrice de Recherche	INRA Orleans	Examinateur
M. Christian Cilas	Directeur de Recherche	Cirad Montpellier	Examinateur
Mme. Marie Chavent	Maître de Conférences	Université de Bordeaux	Directrice de thèse
Mme. Anne Gégout-Petit	Professeur	Université de Lorraine	Co-directrice de thèse
Mme. Lucia Guérin-Dubrana	Maître de Conférences	Bordeaux Sciences Agro	Co-directrice de thèse

Remerciements

Mes premiers remerciements vont à Anne Gégout-Petit et Lucia Guérin-Dubrana tout d'abord pour avoir accepté de diriger mes travaux de recherche puis pour m'avoir accordée leur confiance et avoir pris le temps de me former pendant ces trois années de thèse. Je les remercie de m'avoir soutenue, pour la patience dont elles ont fait preuve lorsque je leur présentais mes avancées et leurs précieux conseils pour mieux communiquer mes résultats. Plus particulièrement je remercie Anne qui, malgré la distance, a su m'accompagner tout au long de ce travail de recherche et Lucia pour m'avoir fait découvrir la biologie. Merci à Marie Chavent pour avoir accepté de diriger ma thèse et avoir participé à mon comité de pilotage.

Je souhaite également remercier Joël Chadoeuf et Florent Bonneu pour leurs nombreux conseils sur les méthodes de tests Joint Count, j'ai eu le plaisir de collaborer avec eux lors de la rédaction de l'article sur le test Joint Count. Un grand merci également à Xiangming Xu pour sa révision de cet article.

Je tiens à remercier Pierre Curmi pour notre collaboration pour l'article sur le sol et Philippe Chéry pour ses explications des méthodes géostatistiques. Je remercie mon Comité de pilotage, Avner Bar-Hen, Jean-Noël Aubertot, Christian Gary, ainsi que Florent et Pierre pour m'avoir guidée lors des grandes étapes de ma thèse.

Merci à Liliane Bel et Benoit Marçais d'avoir accepté d'être les rapporteurs de ma thèse ainsi que tous les membre du jury, Isabelle Cousin, Christian Cilas.

Pour la base de données sur laquelle repose mes travaux, je souhaite remercier les ingénieurs Sylvie Bastien, Pauline Souquet, David Morais. Pour leurs travaux préliminaires, je remercie les stagiaires David, Damien, Louis et Chloé.

Je tiens à remercier Frédéric Fabre pour son aide sur la modélisation ainsi que ses conseils bibliographiques, notamment pour le livre qu'il m'a prêté

Je souhaite également remercier le directeur du laboratoire Denis Thierry, ainsi que toute l'équipe UMR SAVE de l'INRA avec laquelle j'ai passé ces trois années, pour l'ambiance qui règne dans le laboratoire et pour m'avoir fait découvrir la vie quotidienne dans un laboratoire de biologie. En particulier, je souhaite remercier Amira et Rana, qui ont commencé et termineront leur thèse en même tant que moi.

Même si je n'ai pas passé beaucoup de temps à l'Inria, je souhaite également remercier tous les doctorants de l'équipe CQFD pour les nombreuses discussions et leur aide précieuse en statistique. Je souhaite remercier Ray Godfrey pour la qualité de ses corrections de l'anglais dans mes articles ainsi que pour m'avoir accueillie à son cours de discussion en anglais.

Un grand merci à mes parents, sans lesquels je ne serai jamais arrivée jusque-là, pour leur soutien depuis toujours et pour m'avoir encouragée à tenter l'aventure en France. Malgré la distance et le décalage horaire, ils ont toujours su m'apporter réconfort lors des moments difficiles. Merci enfin à Thomas, pour son soutien tout

au long de ces années de travail assidu et pour m'avoir supportée lors des périodes les plus difficiles.

Résumé

Modélisation spatio-temporelle pour l'esca de la vigne à l'échelle de la parcelle

L'esca de la vigne fait partie des maladies de dépérissement incurables dont l'étiologie n'est pas complément élucidée. Elle représente un des problèmes majeurs en viticulture. L'objectif général de cette thèse est d'améliorer la compréhension des processus épidémiques et des facteurs de risque. Pour ce faire, nous avons mené une étude quantitative du développement spatio-temporel de l'esca à l'échelle de la parcelle.

Dans un premier temps, pour détecter d'éventuelles corrélations spatiales entre les cas de maladie, des tests statistiques non paramétriques sont appliqués aux données spatio-temporelles d'expression foliaires de l'esca pour 15 parcelles du bordelais. Une diversité de profils spatiaux, allant d'une distribution aléatoire à fortement structurée est trouvée. Dans le cas de structures très agrégées, les tests n'ont pas montré d'augmentation significative de la taille des foyers, ni de propagation secondaire locale à partir de ceps symptomatiques, suggérant un effet de l'environnement dans l'explication de cette agrégation. Dans le but de modéliser l'occurrence des symptômes foliaires, nous avons développé des modèles logistiques hiérarchiques intégrant à la fois des covariables exogènes liées à l'environnement et des covariables de voisinage de ceps déjà malades mais aussi un processus latent pour l'auto-corrélation spatio-temporelle. Les inférences bayésiennes sont réalisées en utilisant la méthode INLA (Inverse Nested Laplace Approximation). Les résultats permettent de conforter l'hypothèse du rôle significatif des facteurs environnementaux dans l'augmentation du risque d'occurrence des symptômes. L'effet de propagation de l'esca à petite échelle à partir de ceps déjà atteints situés sur le rang ou hors rang n'est pas montré. Un modèle autologistique de régression, deux fois centré, qui prend en compte de façon plus explicite la structure spatio-temporelle de voisinage, est également développé. Enfin, une méthode géostatistique d'interpolation de données de nature anisotropique atypique est proposée. Elle permet d'interpoler la variable auxiliaire de résistivité électrique du sol pour estimer à l'échelle de chaque plante de la parcelle, la réserve en eau du sol disponible pour la vigne.

Les méthodes géostatistique et spatio-temporelles développées dans cette thèse ouvrent des perspectives pour identifier les facteurs de risques et prédire le développement de l'esca de la vigne dans des contextes agronomiques variés.

Mots clés : Vigne; maladie du bois; join count test; modèle Bayésien hiérarchique; modèle auto-logistique

Abstract

Spatio-temporal modelling of esca grapevine disease at vineyard scale

Esca grapevine disease is one of the incurable dieback disease with the etiology not completely elucidated. It represents one of the major threats for viticulture around the world. To better understand the underlying process of esca spread and the risk factors of this disease, we carried out quantitative analyses of the spatio-temporal development of esca at vineyard scale. In order to detect the spatial correlation among the diseased vines, the non-parametric statistical tests were applied to the spatio-temporal data of esca foliar symptom expression for 15 vineyards in Bordeaux region. Among vineyards, a large range of spatial patterns, from random to strongly structured, were found. In the vineyards with strongly aggregated patterns, no significant increase in the size of cluster and local spread from symptomatic vines was shown, suggesting an effect of the environment in the explanation of this aggregation. To model the foliar symptom occurrence, we developed hierarchical logistic regression models by integrating exogenous covariates, covariates of neighboring symptomatic vines already diseased, and also a latent process with spatio-temporal auto-correlation. The Bayesian inferences of these models were performed by INLA (Inverse Nested Laplace Approximation) approach. The results confirmed the effect of environmental factors on the occurrence risk of esca symptom. The secondary locally spread of esca from symptomatic vines located on the same row or out of row was not shown. A two-step centered auto-logistic regression model, which explicitly integrated the spatio-temporal neighboring structure, was also developed. At last, a geostatistical method was proposed to interpolate data with a particular anisotropic structure. It allowed interpolating the ancillary variable, electrical resistivity of soil, which were used to estimate the available soil water content at vine-scale. These geostatistical methods and spatio-temporal statistical methods developed in this thesis offered outlook to identify risk factors, and thereafter to predict the development of esca grapevine disease in different agronomical contexts.

Additional keywords: Grapevine; wood trunk disease; Join count test; Bayesian hierarchical model; auto-logistic model

Contents

Chapter 1

Contexte et objectifs scientifiques

1.1 Introduction

En épidémiologie végétale, pour mieux comprendre les processus qui gouvernent la dynamique d'une maladie, il est nécessaire de coupler l'échelle spatiale et temporelle (Gibson and Austin, 1996). Pour la dimension spatiale, plusieurs éléments sont à considérer : l'échelle ou les échelles spatiale(s) choisie(s) (i.e plante, population, paysage), l'hétérogénéité à cette échelle, et aussi la structure spatiale du peuplement hôte (Shaw, 1994, Turechek and McRoberts, 2013, Gilligan and van den Bosch, 2008). L'étude de la dynamique spatio-temporelle des maladies peut se faire par des approches déterministes, stochastiques ou couplant le deux. Dans le cas de maladie complexe non élucidée, ou partiellement élucidée, les modèles stochastiques sont particulièrement intéressants. Ils permettent de tenir compte des effets latents, et/ou aléatoires, de mettre en évidence des processus écologiques sous-jacents, par exemple en établissant les relations statistiques entre la maladie et des facteurs d'intérêt (Shaw, 1994, Keeling and Ross, 2008). De nombreux progrès ont été réalisés dans le développement d'outils de modélisation spatio-temporelle et d'inférence pour l'écologie et très utiles dans les études épidémiologiques de maladie (Wikle et al., 1998, Turechek and McRoberts, 2013). Le travail présenté dans ce mémoire s'inscrit dans ce cadre et a pour but d'améliorer les connaissances épidémiologiques de l'esca de la vigne par une étude de la dynamique spatio-temporelle de la maladie à l'échelle de la parcelle.

L'esca est l'une des maladies du bois de la vigne, associée à des champignons pathogènes qui se développent dans les tissus ligneux de la charpente. Cette maladie conduit à un dépérissement progressif de la plante. Elle est présente dans toutes les régions viticoles des deux hémisphères. En France, avec l'eutypiose et le dépérissement à Botryosphaeria (appelé black dead arm), l'esca est aussi très présente dans tous les vignobles et peut être considérée comme une maladie endémique. Les chiffres de l'Observatoire des Maladies du Bois montrent que 66 à 100 % des vignobles d'une région étaient atteints par l'esca et le black dead arm (BDA) en 2013 (Bruez et al., 2013). En moyenne, sur le territoire français, environ 12 % de ceps sont qualifiés d'improductifs (perte de production, pieds manquants ou en cours de remplacement) à cause de cette maladie. Les coûts économiques engendrés sont liés au renouvellement des ceps morts et à la perte de rendement. La qualité des raisins prélevés sur les ceps atteints d'esca est diminuée (Lorrain et al., 2012). Les pertes économiques

causées par les maladies du bois dont l'esca sont considérables pour la filière « vigne » dont le poids socio-économique est majeur pour plusieurs régions françaises. L'esca est une maladie d'autant plus pré occupantes que la seule lutte actuellement repose sur des mesures de prophylaxie.

En Europe, plusieurs observations indiquent une recrudescence des maladies du bois dont l'esca de la vigne depuis la fin du XXème siècle et la première décennie des années 2000 (Larignon et al., 2009, Bertsch et al., 2013). Des hypothèses sont données comme le retrait de l'arsénite de sodium (seule matière active homologuée pour lutter contre l'esca de la vigne jusque 2001 en France), le changement climatique et les modifications des pratiques culturales (Surico et al., 2006).

L'esca de la vigne est connue depuis de nombreux siècles avec sa description retrouvée dans des ouvrages anciens (Mugnai et al., 1999). Cependant, peu de travaux scientifiques sur cette maladie ont été menés jusque la fin du XXème siècle.

Le travail sur la description des symptômes et sur l'identification des champignons pathogènes isolés des nécroses réalisé par Larignon en 1991 a été suivi par de nombreux travaux sur le rôle des différents champignons pathogènes ainsi que leur interaction avec la plante (Mugnai et al., 1999, Bertsch et al., 2013). Cependant, les études décrivant le développement de l'esca au vignoble et les facteurs impliqués sont moins nombreuses. Dans une revue bibliographique sur les maladies du bois, Larignon et al. (2009) souligne le peu de d'informations disponibles sur les facteurs environnementaux impliqués dans le développement de la maladie et son expression. Pourtant ces informations sont indispensables pour aider à la gestion de cette maladie.

Ce premier chapitre présente de façon plus approfondie l'objet de l'étude, l'esca de la vigne et l'état des connaissances utiles pour situer et motiver les objectifs scientifiques qui seront décrits en fin de ce chapitre.

1.2 L'esca de la vigne : une pathologie complexe non élucidée

L'esca de la vigne peut être défini comme un syndrome, c'est à dire une maladie qui présente plusieurs types de symptômes : des symptômes internes, dits primaires, sous forme de nécroses internes des tissus ligneux causées par les agents fongiques pathogènes et des symptômes secondaires, foliaires, dont l'extériorisation n'est pas toujours réalisée. Ainsi le diagnostic de l'esca n'est pas facile au vignoble. Selon l'âge de la vigne et le type de champignon associé, plusieurs syndromes de l'esca ou apparentés à l'esca ont été décrits (Surico et al., 2008) : maladie de Petri, jeune esca ou GLSD (Grapevine leaf stripe disease), esca et esca "proper". Dans cette partie seront détaillés la symptomatologie de l'esca (dit esca proper), les agents pathogènes associés et les relations entre les différents types de symptômes, les éléments d'épidémiologie et les facteurs favorisant le développement de l'esca.

1.2.1 Différents type de symptômes

1.2.1.1 Les symptômes foliaires

Deux formes de la maladie sont classiquement décrites à partir de l'expression des symptômes foliaires : une forme lente ou chronique, caractérisée par des colorations du limbe des feuilles : des digitations jaunes pour les cépages blanc ou rouges, bordées de jaune pour les cépages noirs, situées entre les nervures vertes (Dubos, 1999), ce qui donne un aspect tigré aux feuilles Figure 1.1. La deuxième forme foliaire, appelée forme apoplectique ou apoplexie est caractérisée par un dessèchement rapide et généralisé sur l'ensemble du cep, conduisant à la mort de la plante. Des études récentes montrent qu'entre ces deux formes, il existe de nombreux états intermédiaires allant de quelques feuilles symptomatiques à un cep entièrement apoplectique (Lecomte et al., 2012). Ces auteurs proposent une grille de sévérité pour prendre en compte ces états intermédiaires de sévérité. Sous les latitudes du territoire français, les symptômes foliaires de l'esca apparaissent au vignoble courant juin, mais l'expression peut s'étaler jusque fin août. L'expression foliaire de la maladie est erratique, c'est-à-dire qu'un cep exprimant les symptômes une année donnée peut ou pas exprimer les symptômes l'année suivante (Surico et al., 2000a).

Figure 1.1: Symptômes foliaire d'esca de la vigne. Expression de la forme lente (photos de droite et milieu). Cep de vigne présentant la forme apoplectique (Photo de gauche) (Auteur : P. Lecomte)

1.2.1.2 Différentes nécroses internes associées à des champignons pathogènes

Dans la définition première de l'esca, les ceps atteints par l'esca sont caractérisés par la présence d'une pourriture blanche, appelée amadou, dans le tronc et les bras de vigne (Mugnai et al., 1999). Les descriptions plus récentes des nécroses montrent la complexité des nécroses internes (Larignon and Dubos, 1997, Mugnai et al., 1999, Maher et al., 2012). La Figure 1.2 montre des coupes transversales de tronc ou de bras de vigne atteintes d'esca et les différents types de nécroses. La caractérisation des nécroses internes et l'analyse de la microflore pathogène associée permettent de décrire deux processus de dégradation du bois lié à un développement successif de champignons (Larignon and Dubos, 1997). Le premier processus donne une nécrose claire et tendre en position centrale impliquant trois champignons : *Phaeomoniella chlamydospora*, isolée de nécrose en forme de ponctuation noire, *Phaeacremonium*

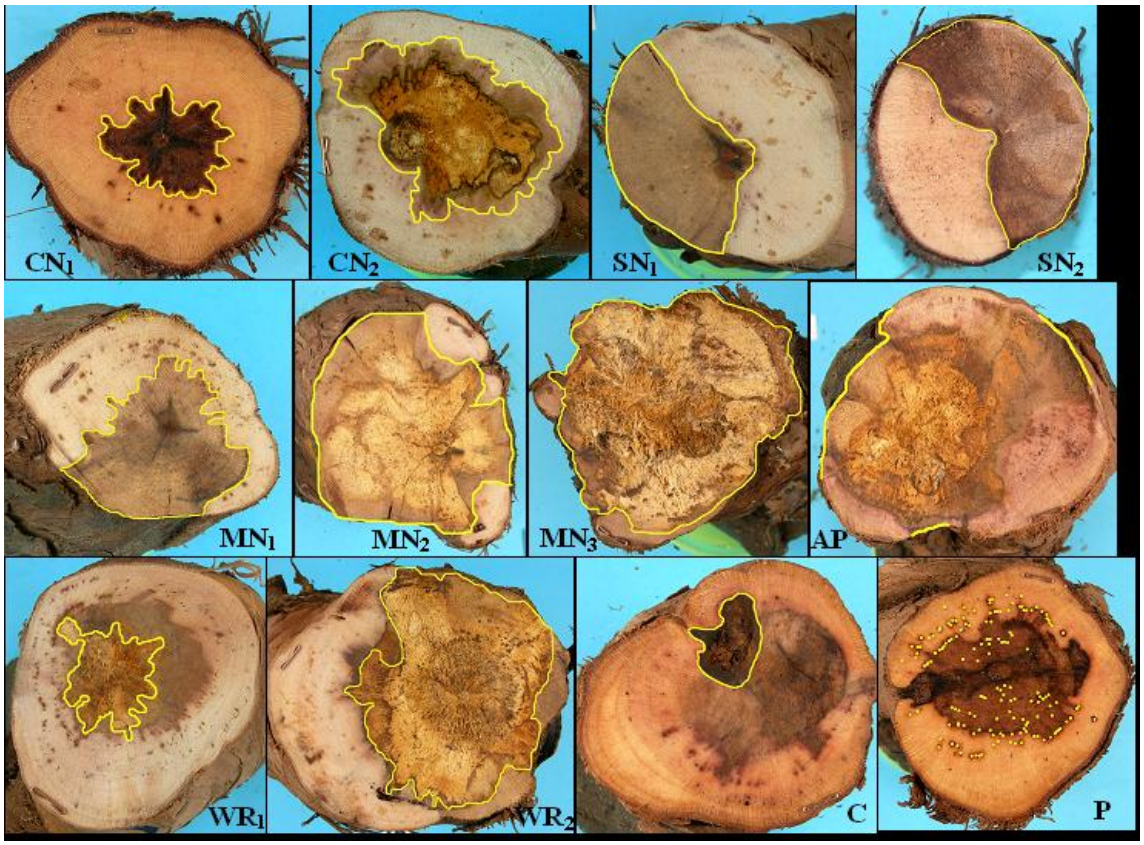


Figure 1.2: Sections transversales de tronc de vigne présentant différents types de nécroses internes (CN : nécrose centrale, SN : nécrose sectorielle, MN : nécrose mixte (centrale et sectorielle), WR : pourriture blanche, P : ponctuation noire, C : cône de cicatrisation)(Auteur : N.Maher)

aleophilum, isolé de la "prénécrose" de l'esca dans les tissus brun à rose, ceinturant la pourriture blanche, et *Fomitiporia mediterranea*, champignon basidiomycète, colonisateur secondaire qui produit cette pourriture du bois. Le deuxième processus à deux champignons conduit à une nécrose claire et tendre en position sectorielle qui fait intervenir l'agent de l'eutypiose, *Eutypa lata* puis *Fomitiporia mediterranea*. D'autres champignons pathogènes appartenant aux Botrysphaeriaceae sont aussi isolés de nécroses sectorielles de ceps exprimant l'esca (White et al., 2011, Maher et al., 2012). L'identification des champignons pathogènes des nécroses reste un sujet d'étude important. Grâce à la mise en oeuvre de méthodes moléculaires des communautés microbiennes, les études récentes montrent une diversité de la microflore colonisant les tissus nécrosés (Bruez et al., 2014). Ces mêmes champignons pathogènes sont aussi présents dans les tissus non nécrosés.

1.2.1.3 Lien entre symptômes foliaires et nécroses internes

Le développement de la maladie est difficile à étudier car les nécroses du bois ne sont pas visibles (caractère cryptique) et les symptômes foliaires ne s'expriment pas toujours. La cause du déclenchement des symptômes foliaires est recherchée en étudiant les toxines produites par les champignons pathogènes, au niveau des tissus ligneux et leur migration vers les organes herbacés. Des métabolites, pro-

duites par les deux champignons ascomycètes, Pch et Pal, ont été isolées (Evidente et al., 2000, Tabacchi et al., 2000, Abou-Mansour et al., 2004, Andolfi et al., 2011). Leur effet toxique sur des feuilles détachées a été montré (Bruno and Sparapano, 2006a,b, Bruno et al., 2007). Certains composés phytotoxiques ont été détectés dans la sève du xylème (Bruno and Sparapano, 2006b, Bruno et al., 2007) ou détectés dans les feuilles symptomatiques (Andolfi et al., 2009). Ces résultats suggèrent que les symptômes foliaires d'esca résultent de la migration vers les organes herbacés de substances phytotoxiques produits par les champignons ascomycètes impliqués. Aucun champignon pathogène n'est isolé des feuilles.
L'inoculation artificielle de chacun des champignons impliqués dans l'esca conduit à des symptômes dans le bois mais plus rarement à l'expression de symptômes foliaires (Laveau et al., 2009, Sparapano et al., 2001, Feliciano et al., 2004).
Les études des relations entre la sévérité des nécroses internes et la sévérité des symptômes foliaires montrent des résultats contradictoires. Calzarano and Di Marco (2007) ne montrent pas de corrélation significative entre le taux d'expression de l'esca au niveau foliaire et le taux de nécroses interne, pour deux variétés de vigne italiennes, Trebbiano d'Abruzzo et Sangiovese. En comparant la forme chronique et la forme sévère chez la variété Cabernet Sauvignon, Maher et al. (2012) montrent que la forme sévère de l'esca présente une forte dégradation des tissus ligneux périphériques comprenant la zone cambiale et la partie externe du xylème. Pour la forme chronique, les symptômes foliaires sont associés à des nécroses internes de formes différentes, de taille plus importante que dans des ceps asymptomatiques. La régression logistique montre que la pourriture blanche est le meilleur prédicteur de la forme chronique.

1.2.2 État des connaissances sur l'épidémiologie et les facteurs impliqués

Les études épidémiologiques visent à comprendre comment la maladie se développe et à identifier les sources d'inoculum primaire et secondaire et les facteurs qui gouvernent les maladies. Plusieurs types de facteurs sont différenciés : les facteurs liés à l'hôte, aux agents pathogènes, aux facteurs environnementaux et dans le cas d'un agroécosystème, les facteurs anthropiques. Dans cette partie, nous développerons rapidement la biologie des différents champignons associés à l'esca et développerons plus longuement l'état des connaissances des facteurs impliqués.

1.2.2.1 Biologie des principaux champignons pathogènes en cause

Phaeomoniella chlamydospora est un champignon ascomycète qui forme des structures de conservation appelées pycnides sur les plaies de taille âgées de cep de vigne au vignoble et dans les zones protégées de l'écorce (fente) ou sous l'écorce (Larignon et al., 2009). Les spores sont disséminées par l'eau, le vent ou les insectes arthropodes et pénètrent par les plaies de taille hivernales lors de périodes douces et pluvieuses (Larignon et al., 2009, Moyo et al., 2014). Ce champignon est présent dans les tissus vasculaires du bois. Il est retrouvé dans les bois d'un an en tant qu'endophyte (présent sans causer de symptôme). Les bois contaminés, qui servent à la fabrication des plants, sont des sources d'inoculum potentiel. Au cours

des étapes de fabrication des plants en pépinière, des contaminations sont possibles. Ce champignon est détecté aussi dans le sol mais la relation entre l'inoculum du sol et la contamination des plants n'est pas montrée. *Phaeoacremonium aleophilum* (forme teléomorphe *Togninia minima* (Tul. & C. Tul) présente un cycle biologique similaire. Il se dissémine par les voies aériennes, sous forme de spores sexuées ou asexuées et contamine les plaies de taille pendant la période végétative. Le champignon peut aussi être propagé par les bois contaminés (Larignon et al., 2009). Les études de diversité génétique de populations de *P. aleophilum* indiquent plusieurs sources d'inoculum primaire dans un même vignoble (Péros et al., 2000, Borie et al., 2002). *Eutypa lata*, champignon ascomycète et agent pathogène responsable d'une autre maladie du bois, l'eutypiose, est aussi associé à l'esca. La dissémination par les spores sexuées est aérienne et peut se produire toute l'année lors de précipitations (Dubos, 1999). En hiver, il infecte les plaies de taille. Il se conserve sous sa forme sexuée sur les vieux bois morts. Il ne se propage pas par le jeune bois ou les plants de vigne (Larignon et al., 2009). *Fomitiporia medeterranea* est un champignon basidiomycète aussi à dissémination aérienne. Il est présent au vignoble sur les écorces sous forme de carpophore libérant les basidiospores, spores infectieuses, libérées dans l'air. Comme *Eutypa lata*, ce champignon n'infecte pas les bois d'un an et les jeunes plants (Larignon et al., 2009). Ces différents champignons associés à l'esca peuvent infecter d'autres espèces végétales que la vigne. L'inoculum provenant de l'environnement proche des vignobles peut donc servir d'inoculum primaire (Larignon et al., 2009).

1.2.2.2 Facteurs liés à l'hôte

Les résultats d'observation du taux de maladie au vignoble ou ceux d'inoculation artificielle de variétés de vigne montrent des différences de sensibilité en fonction du cépage ou du clone d'un même cépage (Bruez et al., 2013, Murolo and Romanazzi, 2014). En France, les cépages les plus touchés au vignoble sont ceux du Jura (Trousseau, Poulsard, Savagnin). En région Aquitaine, les cépages les plus sensibles sont le Sauvignon et le Cabernet Sauvignon. Les vignes âgées entre 15 et 25 ans sont les plus affectées par la maladie mais l'esca s'exprime aussi sur des vignes jeunes. Dans un essai expérimental, Cordeau (1998) montre que les porte-greffes, conférant de la vigueur au greffon ou mal adaptés à un facteur limitant du sol, ont montré des taux de mortalité plus élevés. Plus récemment, Murolo and Romanazzi (2014) montrent un taux d'esca plus élevé lorsque les variétés Fiano et Sauvignon sont greffés sur le porte greffe SO4 en comparaison avec le porte greffe 1103P. Les auteurs suggèrent que cette différence peut être liée à une plus grande résistance à la sécheresse de 1103P par rapport à SO4. Ils montrent des différences de comportement entre clones seulement pour la variété Sauvignon.

1.2.2.3 Facteurs anthropiques

Le système de conduite et les types de taille de la vigne ont une influence sur le taux de maladie (Lecomte et al., 2011). Weber et al. (2007) montrent par exemple que le pré taillage de la vigne réduit le taux de maladies du bois. Geoffrion and Renaudin (2002) montrent que la taille Guyot-Poussart, avec les plaies de taille sur un même côté du rameau, est associé à moins de maladie que d'autres systèmes de taille.

En conditions contrôlées, Fischer and Kassemeyer (2015) montrent que des plantes stressées sont plus infectées par *P. chlamydospora* après 12 mois d'incubation. Au vignoble, les pratiques culturales qui conduisent à un déséquilibre physiologique de la plante trop important peuvent générer un taux de maladie élevé.

1.2.2.4 Facteurs pédo-climatiques

L'ensemble des caractères du climat local et du sol crée un environnement biophysique qui conditionne le développement de ce pathosytème complexe. Les facteurs pédo climatiques peuvent agir sur de nombreuses composantes du pathosystème : effet sur les agents pathogènes, la plante, les interactions plante-agents pathogènes. Des effets indirects sur l'environnement physique, biotique du pathosytème sont aussi à prendre en compte.

1.2.2.4.1 ***Les facteurs pédologiques*** Le sol est le support physique de la culture végétale, un lieu d'échange qui permet la nutrition de la plante. En particulier le sol va conditionner le régime hydrique et azoté de la plante. Des observations au vignoble de l'expression de la maladie, rapportées par Viala (1926) et Geoffrion (1971), indiquent que l'apoplexie se rencontrent plus fréquemment dans les parcelles à sol argileux et compact. Surico et al. (2000a) observent un taux d'esca plus élevé sur des sols lourds et humides. La comparaison de parcelles de vigne, dans la région bordelaise a montré que les taux de maladie les plus élevés sont observés dans les parcelles à forte réserve utile (alimentation en eau du sol non limitante) (Destrac-Irvine et al., 2007).

1.2.2.4.2 ***Les facteurs climatiques*** Le climat influence la production et la dissémination des spores des champignons impliqués dans l'esca, et les différentes phases de l'infection. Les précipitations sont reliées à des pics de capture des deux champignons *Phaeomoniella chlamydospora* et *Phaeoacremonium aleophilum* (Larignon and Dubos, 2000, Eskalen and Gubler, 2001). Dans les conditions agronomiques de la région de Bordeaux, Lecomte et al. (2012) montrent que l'expression des symptômes foliaires débute généralement fin juin avec une expression maximale fin juillet. En Italie, Surico et al. (2000a) mettent en évidence que des étés frais et pluvieux favorisent la forme lente de l'esca alors que des étés chauds et secs sont favorables à la forme apoplectique. Marchi et al. (2006) relie de façon inverse les précipitations en mai et juin ou seulement en juin avec l'esca "caché" (esca caché défini la non expression foliaire de la maladie une année alors qu'un cep a déjà exprimé la maladie auparavant). Cela signifie que ces conditions favorisent l'expression de la maladie.

1.2.3 Conclusion

Depuis longtemps, de nombreuses observations de l'expression de la maladie de l'esca ont montré qu'il existe différents facteurs qui peuvent avoir une influence sur le développement de la maladie. Cependant, ces études sont effectuées en comparant les comportements de la maladie dans des parcelles différentes. A notre connaissance, aucune modélisation n'a été effectuée pour identifier ces facteurs et quantifier ces relations avec la maladie et aucune étude à l'échelle intra-parcelle n'a été faite .

Comme le pathosystème peut être très différent entre les parcelles du fait d'interactions complexes et diverses, une étude à l'échelle intra-parcelle est nécessaire pour éviter la combinaison d'un trop grand nombre de facteurs.

L'étude de l'esca reste un problème difficile pour de multiples raisons. D'une part, les données épidémiologiques sont partielles. La maladie, même si elle apparaît précocement au sein des tissus du bois (cf section 1.2), n'est identifiable qu'à l'apparition des symptômes foliaires. Une fois "installée", elle est pérenne mais les symptômes foliaires peuvent être intermittents et les différents états du cycle biologique ne sont pas accessibles.

D'autre part, la maladie n'est pas reproductible totalement en conditions contrôlées. Les études épidémiologiques nécessitent des données issues de nombreuses années d'observation.

Pour des raisons indépendantes de l'esca, les données des facteurs pédologiques ne sont pas toujours mesurées ni à l'échelle du cep ni de manière exhaustive car trop couteuses et/ou trop intrusives.

Dans la section suivante, nous motiverons les études de modélisation spatio-temporelle des maladies complexes des plantes et ferons un point bibliographique sur les travaux existants à l'échelle intra-parcelle pour l'esca .

1.3 Apport des analyses spatio-temporelles et de la modélisation dans le cas de pathologies complexes

1.3.1 Généralités

En épidémiologie végétale, l'analyse spatiale d'une maladie a pour objectif de décrire la distribution dans l'espace des individus malades ou des agents pathogènes au sein d'un peuplement homogène ou à plus large échelle (paysage). La répartition des individus aléatoire, régulière, ou agrégative nous apporte des informations sur le fonctionnement de la maladie (Gelfand et al., 2010). Plus précisément, dans la cas d'un profil agrégé, la taille des agrégats et leur dynamique de formation donnent des hypothèses sur les processus sous-jacents qui ont produit cette structure. Dans un agro écosystème, la structure spatiale ou spatio-temporelle est le résultat d'interactions directes ou indirectes entre agents pathogènes, l'hôte, l'environnement (biotique & abiotique), le(les) vecteur(s) et l'homme. L'analyse des profils spatiaux permet d'élaborer des hypothèses sur les processus démographiques et de dispersion du ou des agents pathogènes, sur le rôle de certains vecteurs tels par exemple la pluie ou le vent ou encore les processus de contagion (Purse and Golding, 2015). Les facteurs identifiés peuvent être testés par le biais de la modélisation. L'analyse spatiale peut aussi être utilisée pour des maladies dont la cause est inconnue. L'analyse spatiale et temporelle a par exemple été utilisée pour donner des hypothèses sur les causes de la maladie de mort subite des citrus au Brésil (Bassanezi et al., 2003). A partir des symptômes et des profils spatio-temporels, les auteurs ont montré la similitude de dynamique de la maladie avec celle d'une virose transmise par des insectes-vecteurs de type puceron. Ces méthodes ont été également utilisées dans le cas d'une maladie émergente de l'hévéa : la nécrose de l'hévéa (Peyrard et al.,

2006). Les résultats des analyses de la structure spatiale d'une maladie peuvent aussi servir à proposer des mesures de contrôle de la maladie ou à améliorer celles-ci (Spolti et al., 2012). L'analyse de la structure spatiale est un préalable indispensable à la modélisation afin de définir l'échelle d'étude la plus pertinente. Cette analyse va de pair avec l'analyse de la dynamique temporelle de la maladie.

De façon générique, la modélisation vise à une représentation simplifiée de la réalité en vue de la comprendre ou de la faire comprendre. En épidémiologie, la modélisation spatio-temporelle a pour but de construire des représentations des systèmes pathologiques en prenant en compte leurs corrélations spatiale et temporelle. D'une part, elle prend en compte l'aspect structural avec la représentation des structures spatiales et des indicateurs associés ; d'autre part, elle inclut l'aspect dynamique avec la représentation des séquences temporelles. Les objectifs de la modélisation sont de décrire, d'expliquer ou/et de prédire. Gosme (2007) présente les nombreux modèles spatio-temporels qui ont été appliqués ou élaborés en épidémiologie végétale.

Parmi les différents modèles, ceux qui nous intéressent dans le cadre de ce travail de thèse, sont les modèles qui ont pour but d'expliquer le passage d'un état à un autre, en l'occurrence d'un état sain à un état malade avec la prise en compte de phénomène de propagation interne et des variables environnementales de dimension spatiale et/ou temporelle.

En effet, l'un des objectifs de la modélisation en épidémiologie végétale est d'identifier les facteurs impliqués dans le développement de la maladie, les hiérarchiser et quantifier ces liens. Du point de vue spatial se pose la question de l'échelle d'étude pour la maladie et des différents facteurs étudiés, question d'autant plus cruciale quand il s'agit des facteurs environnementaux. Le choix de l'échelle dépend de la question posée et des réponses attendues (Bierkens et al., 2000). L'échelle d'observation et d'analyse est reconnue comme influençant les résultats et les conclusions provenant des modèles. Dillon et al. (2014), montrent dans le cas du dépérissement du chêne dû à *Phytophthora ramorum*, qu'un changement d'échelle d'étude modifie les conclusions.

Les interactions hôte-pathogène sont fortement intégrées dans un écosystème spatialement hétérogène ; l'un des enjeux est de comprendre le rôle de cette hétérogénéité de l'environnement biotique ou abiotique à différentes échelles sur le développement d'une maladie. Pour répondre à cet enjeu, les modèles dits hiérarchiques permettent d'appréhender des systèmes complexes et des questions d'épidémiologie intégrant différentes échelles (Turechek and McRoberts, 2013). En particulier, les modèles hiérarchiques spatio-temporels sont intéressants car ils permettent l'explicitation des phénomènes spatiaux et temporels en intégrant des processus latents spatiaux ou spatio-temporels (Cressie and Wikle, 2011). De tels modèles sont spécifiquement adaptés à un paradigme bayésien qui intègre dans sa conception une variabilité intrinsèque et différents niveaux d'aléa permettant la prise en compte de processus latents notamment. De plus, les méthodes d'estimation de type "fréquentiste" ne sont pas faciles et toujours disponibles pour ce type de modèles (Lele and Dennis, 2009).

Il existe bien sûr d'autres méthodes pour modéliser les données spatio-temporelles pour l'épidémiologie. Gibson (1997), par exemple, propose un modèle de Markov en temps continu avec un taux de saut qui permet de distinguer l'effet de l'infection

primaire de la transmission par les voisins. Le modèle calibré permet d'estimer la vitesse de propagation et d'autres caractéristiques de la transmission. Ce modèle est plutôt adapté pour des longues séries temporelles.

1.3.2 Analyses spatiales et modélisation de l'esca

Dans les années 2000 de nombreux articles ont concerné l'étude de la distribution spatiale des ceps exprimant les symptômes d'esca à l'échelle d'une parcelle de vigne, incluant cependant rarement la dimension temporelle (Surico et al., 2000a, Pollastro et al., 2000, Reisenzein et al., 2000, Surico et al., 2000b, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Stefanini et al., 2000, Zanzotto et al., 2013). En plus des méthodes d'ajustement de la distribution spatiale, de nombreuses méthodes concernent l'analyse du type de répartition (aléatoire, régulière, agrégative) et particulièrement des mesures d'agrégation. Les méthodes statistiques utilisées varient suivant le type de variable étudiée (binaire, comptage, continue) et utilisent le découpage de l'espace en quadrats par exemple ou des notions de voisinages et de distances. Pour l'esca de la vigne, les méthodes le plus souvent utilisées sont basées sur la comparaison du profil spatial observé avec celui sous l'hypothèse H_0 d'une distribution aléatoire. Selon les auteurs, et selon les jeux de données utilisés, les résultats diffèrent mais la situation de répartition aléatoire domine, suggérant une propagation de la maladie liée à une dissémination aérienne de propagule, plutôt qu'à une propagation de proche en proche par les outils de taille (Cortesi et al., 2000, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Surico et al., 2000a). Cependant, dans certaines situations, une répartition non aléatoire des ceps exprimant l'esca est montré (Surico et al., 2000a, Edwards et al., 2001, Pollastro et al., 2009). La description de la structure spatiale, incluant la taille des agrégats est rarement abordée dans ces études, seuls Surico et al. (2000a) par le biais de l'analyse spatiale par indice de distance (méthode "2Dclass"), caractérisent la structure agrégative par la taille des "patchs" de ceps malades.

La question de la propagation de l'esca au sein du vignoble est traitée dans deux études utilisant des modèles mathématiques. Stefanini et al. (2000) proposent un modèle statistique paramétrique pour évaluer la probabilité qu'une plante exprime des symptômes d'esca en prenant en compte les ceps de vignes symptomatiques et asymptomatiques dans le voisinage. Ils montrent une probabilité plus élevée qu'un cep devienne symptomatique s'il est situé à proximité de ceps ayant déjà exprimé des symptômes. Ils n'utilisent que le voisinage de ceps situés sur le même rang. En appliquant des modèles hiérarchiques Bayésiens à un jeu de données spatio-temporelles d'une parcelle de vigne en Italie, Zanzotto et al. (2013) testent la propagation secondaire de la maladie en fonction des directions spatiales, le long du rang de vignes ou en dehors du rang. Ils montrent une plus forte probabilité de propagation de cep à cep situés sur un même rang par rapport à celle de cep à cep situés sur des rangs différents.

1.3.3 Conclusion

Le développement des outils de modélisation statistique appliquée au domaine de l'écologie et plus particulièrement à celui de l'épidémiologie végétale, prenant en

compte la dimension spatiale et temporelle, via des processus latents par exemple, ouvre des champs d'application pour une meilleure compréhension des systèmes pathologiques complexes tel que celui de l'esca de la vigne. Le bilan bibliographique sur les études spatio-temporelles de l'esca de la vigne montre le peu d'études scientifiques sur le sujet. Les analyses spatiales réalisées à l'échelle de la parcelle montrent des résultats contradictoires : certaines montrent une répartition aléatoire, d'autre une structure agrégée. Ces analyses prennent peu en compte l'aspect temporel. Pourtant, ce dernier permet d'étudier les questions de propagation à l'échelle de la parcelle. La dynamique de la maladie dans la parcelle mérite d'être explorée et modélisée.

L'étude bibliographique nous indique également le peu voire l'absence d'étude prenant en compte les facteurs environnementaux. En plus de la description des profils spatiaux et spatio-temporels de la maladie, les analyses statistiques nous permettent d'étudier les liens entre la maladie et des facteurs environnementaux, ces facteurs spatialisés à l'échelle intra parcellaire pouvant révéler une hétérogénéité reliée à la maladie (spatiale comme l'état hydrique du sol), à l'échelle de la parcelle mais variant avec le temps (temporelle comme les variables climatiques). A notre connaissance, aucun modèle spatio-temporel appliqué à l'esca de la vigne incluant des paramètres exogènes n'a été développé.

1.4 Objectifs scientifiques du travail de thèse

Au regard de cet état de l'art, l'objectif général de cette thèse est de compléter les connaissances sur l'épidémiologie de l'esca de la vigne en décrivant le développement de l'esca au vignoble par des méthodes de statistique spatiale et spatio-temporelle. Nous visons à développer et appliquer ces méthodes pour étudier la dynamique spatio-temporelle de la maladie à l'échelle d'une parcelle de vigne et pour identifier des facteurs environnementaux pédo-climatiques liés à la maladie. L'intérêt de l'étude à l'échelle d'une parcelle est d'éliminer une des sources de variabilité de la dynamique liée aux pratiques agronomiques. En effet, celles-ci sont homogènes à l'échelle de la parcelle : peuplement génétiquement homogène, géométrie identique, type de taille...

Cet objectif général est décliné ci-dessous en spécifiant les questions d'épidémiologie liée à l'esca de la vigne et les questions méthodologiques.

Le premier sous objectif de la thèse est de mieux comprendre comment la maladie se propage dans la parcelle dans le contexte agronomique du vignoble bordelais. Nous proposons d'analyser la structure spatiale et temporelle des ceps symptomatiques dans 15 parcelles de vigne de la région de Bordeaux, caractérisés par un même cépage et de même âge, pour définir, tout d'abord, l'évolution de la prévalence annuelle et sa variation entre les vignobles. Le deuxième volet est d'analyser l'évolution temporelle de la structure spatiale de l'esca dans chaque parcelle. A l'échelle locale, nous étudierons la relation spatiale entre des ceps symptomatiques à des temps différents afin d'explorer la capacité de propagation secondaire de la maladie.

Le champ des questions à investiguer par ces méthodes est vaste et peut permettre de répondre à des questions pratiques telles que : la propagation de proche en proche le long du rang est-elle importante ? est-elle une règle générale ? Dans ce cas,

les mesures de gestion de la maladie pourraient-elles être adaptées ? Les résultats de ces analyses exploratoires sont à faire en amont de la modélisation. Les hypothèses élaborées suite à ces informations nous conduiront à faire des choix méthodologiques pour la modélisation.

Pour cela des méthodes de statistique spatiale, basées sur la recherche de dépendance spatiale, sont développées. Les informations sur la configuration spatiale de la maladie, aléatoire ou agrégative, nous permettent d'élaborer des hypothèses sur les processus de dispersion et de donner des directions pour la modélisation. Par exemple, une configuration spatiale agrégative implique souvent un processus de dépendance spatiale, et cela nous incite à intégrer l'auto-corrélation spatiale entre les individus dans la modélisation ; une agrégation sur le rang révèle une hétérogénéité spatiale et présente un challenge pour la modélisation. Grâce à l'étude des données spatiales sur une série pluriannuelle de 8 années, nous pouvons étudier la dynamique spatiale, par exemple le passage d'un profil aléatoire à agrégé, l'évolution de la taille des agrégats. Tout cela nous donnera des indications pour prendre en compte la dépendance spatio-temporelle dans les modèles. Pour définir les échelles spatiales, cette analyse, conduite sur un ensemble de parcelles de vigne, nous permettra de juger de la variabilité des résultats. Les hypothèses élaborées suite à les informations nous conduisent à faire des choix méthodologique pour la modélisation.

Le deuxième sous objectif vise à obtenir une information spatialisée à l'échelle du cep pour des facteurs d'intérêt à relier à la maladie. Les facteurs d'intérêt ont été choisis à partir d'hypothèses sur la base des résultats obtenus et de la littérature (subsection 1.2.2, ce chapitre). La première hypothèse est que l'état hydrique de la plante conditionne la sensibilité de la plante et/ou le développement de la maladie. L'hypothèse à tester est que le risque qu'un cep de vigne exprime l'esca sera d'autant plus élevé que son état hydrique est peu contraint. La deuxième hypothèse est que l'état de vigueur de la plante est un facteur de sensibilité. Nous tenterons de répondre à la question est-ce que l'hétérogénéité spatiale pour ces deux facteurs conditionne la répartition des plantes atteintes d'esca ?

Pour obtenir une information spatialisée concernant l'état hydrique, deux approches sont proposées dans la thèse : la première consiste en la spatialisation de la réserve utile du sol, qui est reliée à l'état hydrique de la plante. La deuxième est d'estimer pour chaque plante de la parcelle son état de contrainte hydrique, à partir d'un échantillonnage ponctuel régulier de mesures d'un indicateur de contrainte hydrique, le $\delta^{13}C$. Cet indicateur correspond au rapport isotopique 12C/13C des hydrates de carbone, mesuré sur les sucres du jus de raisin de vigne à maturité (Van Leeuwen et al., 2011).

Pour l'information sur la vigueur de la vigne, nous avons utilisé les données du taux d'azote assimilable du jus de raisin (Van Leeuwen et al., 2009). Ce deuxième facteur rend compte du statut azoté de la vigne, relié à un niveau de vigueur. L'hypothèse sous-jacente est que la sensibilité de la vigne à l'esca est d'autant plus importante chez des plantes plus vigoureuses.

A partir de ces données, des méthodes de géostatistiques seront développées pour obtenir une information géolocalisée au niveau de chaque cep contigu de la parcelle. Pour la spatialisation de la réserve utile du sol, elle est réalisée en utilisant des variables géophysiques mesurées entre les rangs de vigne : la résistivité électrique

apparente couplée à des valeurs de la réserve utile obtenue en des points de sondage. Dans ce cas, la procédure d'interpolation comprend une étape de caractérisation de la structure spatiale des données et une étape d'interpolation basée sur la structure spatiale estimée.

Le dernier sous objectif, qui est au coeur de la thèse, est d'expliquer la dynamique de la maladie à l'échelle de la parcelle en développant des modèles spatio-temporels. La modélisation paramétrique permet de tester et quantifier les relations entre la variable d'intérêt (apparition des symptômes ici) et donne la possibilité d'expliquer la maladie par les facteurs qui jouent sur sa dynamique spatio-temporelle.

La modélisation hiérarchique avec un processus latent nous donne la possibilité de modéliser le processus biologique complexe en tenant compte non seulement des covariables environnementales et de voisins mais aussi des dépendances spatio-temporelles entre les ceps malades.

L'ensemble des covariables constituées de différents indicateurs qui comptent les ceps malades voisins des années précédentes, que nous appelons variables "internes", permet de tester et d'identifier le rôle de la propagation secondaire dans le développement de la maladie et aussi la direction de la propagation (sur le rang notamment).

La régression sur les covariables dites "externes" qui regroupent les facteurs climatiques et les facteurs spatiaux liés à la plante ou au sol, quantifie la relation entre les facteurs pédo-climatiques et la maladie. Cette modélisation est très importante pour la connaissances de l'esca ; à notre connaissance aucune étude n'a été faite pour confirmer des liens entre la maladie et les facteurs pédo-climatiques.

De plus, la structure hiérarchique avec un processus latent spatialement ou spatio-temporellement corrélé permet de modéliser le processus complexe : il sépare les comportements spatiaux et temporels et nous permet de capturer les caractéristiques liées à la dimension spatiale et temporelle simultanément.

Enfin le paradigme bayésien facilite l'estimation des paramètres du modèle ; il permet d'intégrer les effets aléatoires expliquant l'hétérogénéité spatiale.

Nous proposons aussi une autre modélisation alternative. Au lieu de modéliser la dépendance spatiale et spatio-temporelle par des processus latents, il est possible de les modéliser par les individus eux-mêmes, c'est-à-dire en modélisant la distribution de la maladie à partir des distributions conditionnelles en un point sachant la valeur des voisins (auto-modèles). Ce type de modèles est difficile à estimer car la fonction de vraisemblance possède une constante normalisée difficilement identifiable (Guyon and Hardouin, 2002). De plus les modèles auto-logistique avec régression présentent des problèmes au niveau de l'interprétation des paramètres (Caragea and Kaiser, 2009).

Dans cette thèse, nous allons approfondir la connaissance de ce modèle et nous proposons un nouveau type de modèle plus fiable quant à l'interprétation des données. Ce type de modèle décrit une dépendance de voisinage explicite par des lois conditionnelles. Même s'il est encore en cours de développement, nous espérons qu'il soit plus adapté à décrire et expliquer une maladie complexe comme l'esca de la vigne.

1.5 Présentation du plan de la thèse

Les 5 prochains chapitres sont organisés de la façon suivante :

Le **chapitre 2** présente les jeux de données et le principe des grandes méthodes de statistique spatiale et spatio-temporelle que nous avons combinées et/ou adaptées pour répondre aux objectifs de cette thèse. Ces méthodes sont présentées en trois grandes parties : celles pour tester la distribution de la maladie (agrégée ou non notamment) au sein de la parcelle ; les méthodes pour interpoler les mesures de covariables à l'échelle du cep alors qu'elles sont échantillonnées de manière très spécifique dans la parcelle, soit à cause des rangs qui donne une structure aux données, soit parce qu'elles sont échantillonnées en très peu de points dans la parcelle. Enfin, la troisième partie comprend les modèles explicatifs de la maladie, incluant ou non des facteurs, des processus latents temporels ou spatiaux ou des dépendances explicites sur les voisins.

L'objectif du **chapitre 3** est l'étude de la distribution spatiale et temporelle de l'esca de la vigne dans le but de décrire l'histoire de la maladie, tester le caractère aléatoire ou agrégé de sa distribution afin de répondre à des questions de propagation aux voisins suivant différentes directions dont le rang qui est spécifique dans le vignoble.

Le **chapitre 4** présente les méthodes d'interpolation utilisées pour estimer certaines covariables d'intérêt en chaque point de la parcelle ou à chaque emplacement de plante, alors qu'elles n'ont été mesurées qu'en quelques points de la parcelle. Les variables étudiées se distinguent en deux groupes, les plus simples à traiter sont celles pour lesquelles on se contente d'interpoler l'échantillon mesuré pour une valeur sur toute la parcelle. Une deuxième variable, la réserve utile est traitée différemment puisqu'elle est liée à une variable (la résistivité) mesurée densément sur la parcelle. Nous développons plusieurs étapes d'interpolation, spécifiques à nos échantillonnages au sein du vignoble pour estimer la réserve utile sur toute la parcelle.

Le **chapitre 5** est au cœur des objectifs de la thèse. Il propose des modélisations de la probabilité d'occurrence (ou de première occurrence) des symptômes foliaires par des modèles logistiques pouvant intégrer des covariables, un processus latent temporellement et/ou spatialement corrélé et un effet aléatoire dû au rang. Les covariables se décomposent en des covariables environnementales et des covariables incluant des informations sur l'état sanitaire du voisinage dans les années précédant l'année étudiée. Les modèles sont appliqués sur deux parcelles de Bordeaux et les résultats sont analysés soigneusement par rapport aux hypothèses épidémiologiques et les résultats du chapitre 3.

Le **chapitre 6** plus prospectif et méthodologique présente une nouvelle extension du modèle autologistique. En effet, dans le but d'une interprétation des paramètres de la régression sur le nombre de voisins malades l'année étudiée et les années précédentes, nous proposons un modèle autologistique doublement centré. Ce modèle qui pose encore des questions de méthodologie d'inférence n'a pas encore été appliqué à nos données.

Enfin, dans une dernière partie, nous discutons les principaux résultats de cette thèse et donnons quelques perspectives épidémiologiques et statistiques de nos travaux.

Chapter 2

Méthodes statistiques et modélisation

2.1 Introduction

Dans ce chapitre, nous introduisons les méthodes statistiques utilisées pour répondre à nos objectifs scientifiques et la description des jeux de données.

Nous disposons de deux types de données : des données spatio-temporelles de la maladie et des données relatives à l'environnement de la maladie. Ces dernières comprennent des données temporelles climatiques, des données spatialisées à l'échelle de la parcelle soit issues d'échantillonnage géoréférencés (données relatives à l'état écophysiologique de la plante ou au sol), soit des données géophysiques de conductivité électrique spatialement très dense.

A partir des données de maladie que nous possédons, nous construisons les variables d'intérêt de notre étude. Ces variables ont une géolocalisation fixe (plant de vigne à emplacement pérenne dans le temps) et donc correspondent à des données sur réseau. Nous souhaitons explorer la structure spatio-temporelle de ces variables et modéliser le risque qu'un cep exprime des symptômes à un temps t.

Les données relatives à l'environnement nous permettent de générer les covariables du modèle. Ces covariables sont utilisées pour expliquer les variations spatio-temporelles de la maladie. Concernant les données spatialisées, qui ne sont pas fournies pour chaque cep de vigne du réseau, une étape d'interpolation spatiale est nécessaire faisant appel aux géostatistiques.

Nous détaillons successivement les jeux de données utilisés (section 2), les méthodes statistiques pour l'analyse spatio-temporelle des données sur réseau (section 3 et 4) et les méthodes géostatistiques nous permettant d'estimer les covariables du modèle à l'emplacement de chaque cep de la placette de la vigne d'étude (section 5). Pour les données sur réseau, les outils statistiques sont présentés pour tester le caractère aléatoire de la répartition spatiale des plantes symptomatiques contre une hypothèse alternative de type dépendance spatiale qui nous permet de quantifier la configuration spatiale (par abus de langage, nous appellerons ces tests, "tests d'auto-corrélation" dans la suite de ce mémoire). Nous nous concentrons sur les modèles pour les données de type binaire sur un réseau. A la section 5, nous présentons les principes de krigeage pour les données géostatistiques. Cette méthode est importante pour répondre aux questions d'épidémiologie en tenant compte

des contraintes des jeux de données disponibles, nous discutons aussi des problèmes de krigeage pour des structures spatiales présentant une hétérogénéité spatiale (non homoscédastique). Nous discuterons, au sein de ces sections nos choix de modélisation pour la réponse à notre problématique épidémiologique (cf chapitre 1) et l'application aux données.

2.2 Descriptions des données

2.2.1 Des données d'enregistrement de maladie aux variables d'intérêt

Les données de maladie utilisées proviennent de suivi épidémiologique réalisé entre 2004 et 2013 dans 15 parcelles de la région viticole de Bordeaux, réparties dans différentes zones d'appellations contrôlées. Pour chaque parcelle de vigne, une placette rectangulaire de 1200 à 2300 ceps contigus est géoréférencée. Les notations de la maladie sont effectuées lors de deux passages annuels. Une première notation a lieu au printemps afin de noter l'état sanitaire global (incluant les plants manquants, jeunes, morts) et les ceps exprimant les symptômes d'eutypiose. Une deuxième notation est réalisée durant la deuxième quinzaine d'août afin de noter les symptômes d'esca. Etant donné que les plantes de vigne sont conduites à deux branches issues d'un tronc commun, les symptômes sont répertoriés sur chaque branche. Les symptômes des formes chroniques et sévères ne sont pas différenciés.

Un exemple de l'évolution de l'état des ceps de la parcelle CAS1 est donné à la figure suivante Figure 2.1.

Ainsi, les localisations des ceps peuvent représenter des unités géographiques d'un réseau fini que l'on peut munir d'un graphe de voisinage pour représenter les influences possibles entre les ceps pour l'expression de la maladie par exemple. Une variable binaire est attribuée à chaque site potentiel du réseau, correspondant à la "présence-absence" du symptôme foliaire. On s'intéresse à la configuration spatiale de la maladie et on souhaite modéliser deux modalités de dimension spatio-temporelle.

La première modalité est la première expression des symptômes foliaires de l'esca : notons que les relevés n'ayant eu lieu qu'à partir de 2004, nous n'aurons pas la certitude qu'une première expression dans l'histoire de nos données est une première expression réelle du cep. Compte-tenu de ce que nous connaissons de l'histoire des symptômes (Guérin-Dubrana et al., 2013), nous ferons cependant une approximation dans la suite de ce travail : une première expression de symptôme en 2005 ou les années suivantes sans expression depuis 2004 est considérée comme une première expression.

La seconde modalité est l'expression des symptômes foliaires (éventuellement récurrente).

Ces données seront largement étudiées tout d'abord pour tester la répartition spatiale et l'évolution de cette répartition dans le temps (cf. chapitre 3) par des méthodes de test d'auto-corrélation spatiale. Une autre partie consistera à modéliser l'occurrence des symptômes foliaires sur le réseau au cours du temps en utilisant des modèles hiérarchiques spatio-temporels (chapitre 5) et un modèle auto-logistique (chapitre 6).

Figure 2.1: État des cep pour la parcelle 12 de 2006 à 2008. Gris : sains ; bleu : symptômes esca ou BDA ; noirs : morts ; verts : morts passés ; marrons : symptômes d'Eutypiose.

2.2.2 Jeux de données relatives à l'environnement

2.2.2.1 Jeux de données pour spatialiser la réserve utile du sol

Le travail de spatialisation de la réserve utile du sol a été réalisé à partir d'un jeu de données provenant d'une parcelle de vigne située en Bourgogne à St Vallerin (Coordonnées géographiques : $42.687\ N, 4.675\ E$, Département Saône et Loire, France). La parcelle de 4000 m^2 comporte 20 rangs espacés d'un mètre. La distance entre les ceps sur le rang est également d'un mètre. La parcelle est située sur un plateau calcaire avec environ 1 % de pente. Le sol de la parcelle est argileux avec des différences spatiales de texture (Ayachi, 2010). Le taux de cailloux et de graves est d'environ 1 %.

L'acquisition des données de résistivité électrique apparente du sol a été réalisée par la société GEOCARTA, fin mars 2012, à l'aide d'un dispositif roulant entre les rangs de vigne muni d'un appareil d'enregistrement (appelé ARP Automatic Resistivity Profiling). Un total de 65000 points de mesure (22 mesures au m^2) géolocalisé a été obtenu (enregistrement de la position toutes les secondes). A chaque point de mesure, trois données sont disponibles correspondant aux mesures de résistivité électrique apparente pour trois profondeurs (notées voie 1, voie 2, voie 3).

Parallèlement, 14 échantillons de sol, répartis dans la parcelle en fonction des zones, ont été prélevés pour estimer la réserve utile, selon la méthode présentée dans le chapitre 4 (section 1).

2.2.2.2 Jeux de données pour spatialiser l'état hydrique et azoté de la plante

Dans le but d'explorer les relations spatiales entre l'état hydrique et la vigueur de la plante avec la maladie, deux indicateurs mesurés à partir du jus de raisin à maturité ont été utilisés : la valeur du rapport isotopique C13/C12 (appelé $\delta^{13}C$) et la teneur en azote. Les indicateurs ont été mesurés pour 30 échantillons répartis uniformément dans chacune des deux parcelles (CAS1 et MEN2) de la région de Bordeaux. Le protocole d'échantillonnage et de mesure est détaillé dans le chapitre 4.

2.2.2.3 Jeux de données pour la sélection de variables climatiques

Les données climatiques, température et précipitation journalière (2004-2013), proviennent de la base de données SAFRAN de Météofrance. Elles correspondent à des données interpolées à partir de mesures de stations météorologiques Météofrance. L'interpolation permet d'obtenir des données pour chaque maille de 8km $\times$ 8km (appelée maille SAFRAN). Les données de maille SAFRAN ont été utilisées en fonction de l'emplacement géographique des 2 parcelles concernées. Les indicateurs pré sélectionnés pour les modèles spatio-temporels sont présentés dans le Chapitre 5.

2.3 Cadre mathématique pour les données

Soit $S \subset \mathbb{R}^d$ un ensemble spatial. Un champ Z sur S à valeur dans l'espace d'état E est la données d'une collection $Z = \{Z_s, s \in S\}$ de variables aléatoires indexées par S et à valeurs dans E.

2.3.1 Données sur réseau

On suppose dans les deux prochaines sections que Z est un champ aléatoire à valeurs dans E défini sur un réseau discret $S \subset \mathbb{R}^2$, non-nécessairement régulier, S étant muni d'un graphe d'influence $\mathcal{G}$: pour $j \neq i$, $(i,j) \in \mathcal{G}$ signifie que j a une influence sur i. On notera N_i l'ensemble des voisins de i dans le graphes. On remarquera que E est un espace d'état général, il peut être qualitatif ou quantitatif, discret ou non. On supposera qu'il est mesurable, $(E, \mathcal{E})$ étant muni d'une mesure de référence m et on notera l'espace des configurations $(\Omega = E^S, \mathcal{E}^{\otimes S})$ et $\nu := m^{\otimes S}$ la mesure de référence sur Ω.

D'autre part, on se donne une matrice de poids positifs bornés $W = \omega_{ij}, (i,j) \in \mathcal{G}$ quantifiant l'influence de j sur i, avec, pour tout i, $\omega_{ii} = 0$ et $\omega_{ij} = 0$ si $(i,j) \notin \mathcal{G}$. Le choix de W est une étape importante qui dépend du problème considéré. (Gaetan and Guyon, 2008)

2.3.2 Données géostatistiques

Pour l'approche géostatistique, S est un sous-ensemble continu de $\mathbb{R}^d$ et on s'intéressera à la fonction de covariance ou au variogramme de Z. L'objectif central de la géostatistique est de dresser des cartes de prévision de Z par krigeage sur tout S à partir d'un nombre fini d'observations.

Table 2.1: Résumé de différents jeux de données utilisés dans la thèse.

Type	Échelle spatiale	Nombre	Échelle temporelle	Nombre	Nombre de jeux de données	Localisation (Région)	Variable d'intérêt	Chapitre
spatio-temporel	plante	1000 à 2000	année	8 à 10	15	Bordeaux	Occurrence des Symptômes foliaires	3 & 5
spatial	cm entre les rangs de vigne	xx	-	-	1	Bourgogne	Réserve Utile estimée à chaque plante	4
spatial	sondage orienté entre rang de vigne	15	-	-	1	Bourgogne	Réserve utile estimée à chaque plante	4
spatial	échantillonnage régulier	30	-	-	2	Bordeaux	$\delta^{13}C$ estimé à chaque plante	5
spatial	échantillonnage régulier	30	-	-	2	Bordeaux	azote des baies estimé à l'échelle de la plante	5
temporel	-	-	année	10	2	Bordeaux	indicateurs climatiques correlés à la maladie	5
temporel	-	-	année	10	3	Bordeaux	indicateurs climatiques correlés à la maladie	5

2.4 Données sur réseau : test d'auto-corrélation

Toute modélisation de structures d'association spatiale, est confrontée à la même situation de référence : l'indépendance. Cette hypothèse nulle que nous notons H_0 dans la suite, signifie que les valeurs de Z sont réparties aléatoirement dans l'espace.

Deux situations distinctes sont en général retenues contre cette hypothèse. Une association spatiale positive correspond au fait que les valeurs proches de Z tendent à êtres regroupées dans l'espace. Il existe un effet d'agrégation qui rend le voisinage ressemblant, ce type de phénomène est particulièrement intéressant pour notre étude sur l'épidémiologie de l'esca ; une association spatiale négative, qui se présente sous la forme d'une alternance de valeurs, de points clairement isolés (répulsion) ou encore une forme de damier ou d'échiquier. Cela correspond à une association répulsive entre les voisins. Ce cas ne nous intéresse pas pour cette thèse.

Les tests statistiques de l'hypothèse d'existence d'une auto-corrélation spatiale ne sont pas nombreux (Lejeune, 2006). On présente dans les sous-sections suivantes les résumés statistiques adaptés à la mesure d'une auto-corrélation spatiale, particulièrement pour le cas d'une variable dichotomique (binaire).

2.4.1 Forme générale des indices d'auto-corrélation spatiale

Ce qui suit est largement inspiré de Lejeune (2006). Dans le cadre des statistiques spatiales, on cherche à quantifier le degré de corrélation entre les variables de positionnement dans l'espace et les variables d'intérêt. On est alors amené à lier les données relationnelles, indiquant la configuration spatiale des points de mesure, avec l'information fonctionnelle induite par la mesure de la variable Z. Pour cela, on construit une relation quantitative entre paires de lieux de l'espace. On note T la matrice (aléatoire) représentant cette relation, on peut alors choisir parmi une des trois formes classiques ci-dessous:

$$\mathbf{T} = (T_{i,j})_{i,j \in S} \quad T_{ij} = \begin{cases} Z_i \times Z_j & \text{(produit croisé)} & (i,j) \in \mathcal{G} \\ (Z_i - Z_j)^2 & \text{(écart quadratique)} & (i,j) \in \mathcal{G} \\ |Z_i - Z_j| & \text{(écart absolu)} & (i,j) \in \mathcal{G} \end{cases}$$

On peut donc étudier l'association entre l'information locative et la variable d'intérêt.

2.4.2 Le cas d'une variable dichotomique - le test du Join-count

Soit Z une variable binaire à valeurs dans $\{0, 1\}$, nous considérons Z comme l'indicatrice d'un évènement B (ex: la présence de la maladie) et notons N (ex: l'absence de la maladie) l'évènement complémentaire. Pour chaque paire (i, j), les valeurs de z_i et z_j se résument à un tableau (Table 2.2) à quatre cases, pour décrire les évènements possibles, qui se mesurent par différentes statistiques :

Les statistiques M_{11} et M_{01} constituent deux cas particuliers : l'une s'interprète comme une similarité, l'autre comme une dissimilarité. Ces statistiques sont appelées statistiques du Joint-count. On présente ensuite deux types de tests avec la statistique M_{11} qui peuvent servir à mesurer/tester le regroupement des cas de maladie de l'esca dans la parcelle.

	$N, z_i = 0$	$B, z_i = 1$
$N, z_j = 0$	$M_{00} = \frac{1}{2}\sum_{ij,i\neq j}\omega_{ij}(1-z_i)(1-z_j)$	$M_{01} = \frac{1}{2}\sum_{ij,i\neq j}\omega_{ij}(z_i-z_j)^2$
$B, z_j = 1$	$M_{10} = \frac{1}{2}\sum_{ij,i\neq j}\omega_{ij}(z_i-z_j)^2 = M_{01}$	$M_{11} = \frac{1}{2}\sum_{ij,i\neq j}\omega_{ij}z_iz_j)$

Table 2.2: Les statistiques du Join-count

2.4.2.1 Distribution d'échantillonnage d'indice spatial dans le cas binaire

Pour le cas de données binaires, on suppose que Z est une variable de Bernoulli de paramètre $p = \mathbb{P}(Z = 1)$. Sous l'hypothèse H_0 d'indépendance spatiale, la valeur de p ne dépend pas du lieu de mesure et la probabilité attachée à chaque paire de lieux se factorise. On est donc ramené à un schéma d'échantillonnage hypergéométrique ou binomial (Lejeune, 2006). Notons

$$S_0 = \sum_{i\neq j}\omega_{ij} \quad S_1 = \frac{1}{2}\sum_{i\neq j}(\omega_{ij}+\omega_{ji})^2 \quad S_2 = \sum_{i=1}^{n}(\sum_i \omega_{ij} + \sum_j \omega_{ij})$$

Pour un tirage avec remise (binomial), dans un protocole d'observation ne définissant pas a priori le nombre de points où Z prend la valeur 1, on a :

$$E(M_{11}) = \frac{1}{2}S_0p^2, \quad Var(M_{11}) = \frac{1}{4}(S_1(p^2-p^4) + (S_2-2S_1)(p^3-p^4))$$

Pour un tirage sans remise (hypergéométrique), dans un protocole d'observation fixant a priori le nombre de points n_1 où Z prend la valeur 1, on a:

$$\begin{aligned} E(M_{11}) = &\quad \tfrac{1}{2}S_0\tfrac{n_1^{(2)}}{n^{(2)}} \\ Var(M_{11}) = &\quad \tfrac{1}{4}(S_1(\tfrac{n_1^{(2)}}{n^{(2)}} - 2\tfrac{n_1^{(3)}}{n^{(3)}} + \tfrac{n_1^{(4)}}{n^{(4)}}) + S2(\tfrac{n_1^{(3)}}{n^{(3)}} - \tfrac{n_1^{(4)}}{n^{(4)}}) + S_0^2\tfrac{n_1^{(4)}}{n^{(4)}} - (S_0\tfrac{n_1^{(2)}}{n^{(2)}})^2) \end{aligned}$$

où $n^{(p)} = \frac{n!}{p!}$, n est le nombre d'observations et n_1 est le nombre d'observations qui prennent la valeur 1.

2.4.2.2 Test de permutations

D'une façon générale, sous l'hypothèse de répartition aléatoire, la loi permutationelle d'une statistique $I(Z)$ de $Z = (Z_i, i = 1, ..., n)$, conditionnellement aux n valeurs observées $(z_i, i = 1, ..., n)$, est la loi uniforme sur l'ensemble des valeurs $I_\sigma = I(Z_\sigma)$ de I pour les $n!$ permutations de σ $\{1, 2, ..., n\}$. Le niveau de significativité bilatéral (resp. unilatéral) associé à l'observation $I(Z) = a$ est:

$$p_a = \frac{1}{n!}\sum_\sigma 1\{|I_\sigma| > a\} \quad (\text{resp. } p_a^* = \frac{1}{n!}\sum_\sigma 1\{I_\sigma > a\})$$

Lorsque l'énumération de toutes les permutations n'est pas possible, on recourt à la méthode de Monte Carlo en choisissant au hasard, pour m assez grand, m permutations $\{\sigma_1, \sigma_2, ..., \sigma_m\}$ pour lesquelles on calcule les valeurs I_σ et les seuils de Monte Carlo associés p_a^{MC} et p_a^{MC*}.

Pour tester l'hypothèse d'indépendance, sans connaître la loi commune aux Z_i, on pourra utiliser la loi permutationelle de la statistique du join-count notée JC. La justification du test est que, sous (H_0), une permutation des $\{Z_i\}$ qui ne change pas la loi globale de Z, nous avons:

$$(Z_i, i = 1, ..., n) \approx (Z_{\sigma(i)}, i = 1, ..., n)$$

L'avantage d'une méthode permutationelle est de fournir un test approché non-asymptotique sans faire d'hypothèse sur la loi de Z.

2.4.3 Autres indices classiques de dépendance spatiale

On distingue classiquement deux indices de mesure de dépendance spatiale globale sur le réseau $(S, \mathcal{G})$ pour les variables continues : l'indice de Moran évalue une corrélation spatiale et l'indice de Geary un variogramme spatial.

Definition 2.4.1 (Indice de Moran)

$$I = \frac{n}{\sum_{i \neq j} \omega_{ij}} \frac{\sum_{i,j} \omega_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i (Z_i - \bar{Z})^2}$$

si $I > 0$, il y a corrélation positive (clusters); si $I < 0$, il y a corrélation négative (répulsion); si $I \approx 0$, il n'y a pas corrélation.

Definition 2.4.2 (Indice de Geary)

$$C = \frac{n-1}{2\sum_{i \neq j} \omega_{ij}} \frac{\sum_{i,j} \omega_{ij}(Z_i - Z_j)^2}{\sum_i (Z_i - \bar{Z})^2}$$

si $C \approx 0$, il y a corrélation positive (cluster), si $C >> 0$, il y a corrélation négative (répulsion).

L'indice de Moran peut également être calculé pour des données binaires ou pour des données ordinales (Cliff and Ord, 1981). Dans le cas binaire, les Join-count statistiques (M_{11}, M_{01}) sont équivalentes à l'indice de Moran.

Dans le domaine de l'épidémiologie et de l'écologie, un autre indice souvent utilisé est la fonction K de Ripley et ses variations comme par exemple, les statistiques O-rings (Xu et al., 2009). Elle est utilisée spécifiquement pour les processus ponctuels dont la densité des points n'est pas homogène sur la parcelle. Notons $(s_1, ..., s_n)$ les positions des points observés sur S, alors cette fonction est définie par :

$$\hat{K}(r) = \frac{1}{\hat{\lambda}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}_{||s_i - s_j|| \leq r}$$

où $\hat{\lambda}$ est l'estimateur de l'intensité moyen des points : $\hat{\lambda} = \frac{n}{a(S)}$ où n est le nombre de points dans la zone étudiée S et $a(S)$ est la surface de S. Dans notre cas, nous nous intéressons aux ceps malades et nous considérons que ces ceps malades sont des points geo-référencés d'un processus ponctuel. Nous pouvons alors tester le caractère aléatoire spatialement complet (CSR : complete spatial randomness) de ce processus en utilisant la fonction K de Ripley comme une mesure de statistique. La fonction K de Ripley est proportionnelle à la statistique du Join-count avec un rapport $\frac{1}{\hat{\lambda} \cdot n}$ pour $\omega_{ij} = \mathbb{1}_{||s_i - s_j|| \leq r}$.

2.4.4 Application dans la thèse

Au chapitre 3, les tests du joint count seront appliqués pour questionner l'agrégation des cas sans direction privilégiée, sur le rang, et hors rang. Des questions de propagation entre voisins de différents ordres sont aussi posées. Si l'hypothèse nulle est toujours la même (répartition aléatoire), chaque type de question correspond à une hypothèse alternative différente et un voisinage N_i différent. De plus ces tests seront appliqués à un grand nombre de parcelles (15), pour 8 années consécutives; ce qui pose le problème des tests multiples. Des tests globaux seront construits pour pallier ce problème.

2.5 Modèles de régression spatiale sur réseau pour données binaires

2.5.1 Principes

Les modèles de régression spatiale sur réseau sont tous des cas particuliers de modèles d'équations simultanées destinées à décrire la répartition d'un indicateur Z en un lieu, en fonction de covariables et ses variations sur le reste du domaine.

Dans cette section, on présentera des modèles de dépendances spatiales pour des variables binaires.

Les modèles probit et logit sont souvent utilisés pour modéliser les données binaires, d'un point de vue épidémiologique, le modèle logit est plus interprétable que le modèle probit grâce à son lien avec les odds ratio.

Pour la régression logistique, notons $\mathbf{X}_i$ le vecteur des covariables observables liées au site i et ϵ_i un terme correspondant à la dépendance spatiale, on a:

$$\begin{array}{rcl} Z_i|\mathbf{X}_i, \epsilon_i & \sim & Bernoulli(p_i) \\ logit(p_i) & = & \mathbf{X}_i^t\beta + \epsilon_i \end{array}$$

Les différents modèles correspondent à la diversité pour modéliser p_i qui, de surcroit peut dépendre de t. On peut supposer que connaissant les covariables $\mathbf{X}_i$ et ϵ_i, les Z_i sont indépendantes. Dans ce cas le lien entre les Z_i ne vient que de l'autocorrélation (spatiale et/ou temporelle) du terme $\epsilon = \{\epsilon_i, i \in S\}$, un vecteur aléatoire de dimension n, qu'il faut aussi modéliser. C'est l'objet de la section 2.5.2. On peut aussi annuler ϵ_i et mettre des variables concernant les Z_j pour $j \in N_i$, le modèle est alors défini par ces lois conditionnelles et se pose la question de l'existence d'une loi jointe. C'est l'objet des modèle auto-logistiques de la section 2.5.3.

2.5.2 Données sur réseau : modèle de régression logistique avec dépendance spatiale

Dans cette section, on présente deux types de modèles selon deux manières de modéliser le bruit ϵ: modèles logistiques avec dépendance spatiale et son extension à la dimension temporelle. Nous commençons à introduire deux type de modèles pour le bruit qui sera à chaque fois un champ Gaussien autocorrélé : le modèle auto-régressif conditionnel (CAR) Gaussien et le champ aléatoire Markov Gaussian (GMRF).

2.5.2.1 Modèles CAR et GMRF

Pour modéliser ϵ, une manière est de le considérer comme un champ aléatoire spatialement corrélé, avec $E(\epsilon) = 0$ et $Cov(\epsilon) = \Sigma$. On peut ensuite modéliser Σ à partir d'une fonction de covariance, d'un variogramme ou encore d'un modèle auto-régressif spatial. Pour les données sur un réseau, le modèle le plus souvent utilisé est le modèle auto-régressif. On donne un bref aperçu ici pour le modèle auto-régressif conditionnel Gaussien.

Supposons que pour $i = 1, ..., n$, $\epsilon_i|\epsilon_{-i}$ est normal de moyenne et de variance conditionnelles :

$$E(\epsilon_i|\epsilon_{-i}) = \mu_i + \sum_{j \neq i} \beta_{ij}(\epsilon_j - \mu_j) \qquad Var(\epsilon_i|\epsilon_{-i}) = \kappa_i^{-1} \tag{2.1}$$

où κ_i est la précision et μ_i est la moyenne. Sans perdre de généralité, on peut supposer que ϵ_i est de moyenne nulle, on impose alors $\mu_1 = \mu_2 =, ..., = \mu_n = 0$ pour la formule suivante.

$$p(\epsilon_i|\epsilon_{-i}) \sim N(\sum_{j \neq i} \beta_{ij}\epsilon_j, \kappa_i^{-1}) \tag{2.2}$$

On appelle $p(\epsilon_i|\epsilon_{-i})$ les distributions conditionnelles complètes, ici elles sont compatibles : par le lemme de Brook (Brook, 1964), il existe une distribution unique déterminée par ces conditionnelles si ces lois conditionnelles satisfont une condition de factorisation. Sous l'hypothèse supplémentaire que

$$\kappa_i\beta_{ij} = \kappa_j\beta_{ji} \quad \text{for all} i \neq j,$$

ces distributions conditionnelles correspondent à un distribution jointe Gaussienne multivariée de moyenne 0 et de matrice de précision Q avec les éléments, $Q_{ii} = \kappa_i$ et $Q_{ij} = -\kappa_i\beta_{ij}, i \neq j$, ainsi Q est symétrique et définie positive.

Un tel système de distributions conditionnelles est connu comme un système auto-normal (Besag, 1974). Habituellement, on suppose que la matrice de précision Q est régulière ; cependant, les auto-régressions Gaussiennes conditionnelles avec Q singulière sont également d'intérêt et connues comme Auto-Régressions Conditionnelle Intrinsèque (ICAR), souvent utilisées comme loi a priori pour les modèles hiérarchiques.

Les auto-régressions conditionnelles gaussiennes avec une propriété de Markov sont également connues sous le nom de champs gaussiens de Markov (GMRF: Gaussian Markov Random Field).

Un GMRF est tout simplement un vecteur aléatoire $\boldsymbol{\epsilon}$ avec une distribution gaussienne, qui obéit à une certaine propriété d'indépendance conditionnelle (Propriété Markov). Notons $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ un graphe non-orienté avec un ensemble de nœuds $\mathcal{V} = \{1, 2, ..., n\}$ et un ensemble d'arêtes $\mathcal{E}$ qui définit la relation de voisinage entre les nœuds.

Pour tout $(i, j) \in , i \neq j, (i, j) \notin \mathcal{E}$, nous avons :

$$\epsilon_i \perp \epsilon_j | \boldsymbol{\epsilon}_{-\{i,j\}} \tag{2.3}$$

Ce qui signifie que conditionnés sur $\boldsymbol{\epsilon}_{-\{i,j\}}$, ϵ_i et ϵ_j sont indépendants si i et j ne sont pas voisins. Et pour $(i, j) \in \mathcal{E}, i \neq j$, ϵ_i et ϵ_j ne sont pas indépendants sachant $\boldsymbol{\epsilon}_{-\{i,j\}}$.

On donne une définition formelle ici:

Definition 2.5.1 (GMRF) *Un vecteur aléatoire* $\epsilon = (\epsilon_1, ..., \epsilon_n)^t \subset \mathbb{R}^n$ *est appelé un GMRF rapport avec le graphe marqué* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *avec une moyenne* μ *et une matrice de précision* Q *symétrique et définie positive, si et seulement si sa densité est de la forme*

$$\pi(\boldsymbol{\epsilon}) = (2\pi)^{-n/2}|\mathbf{Q}^{1/2}|exp(-\frac{1}{2}(\boldsymbol{\epsilon} - \mu)^t\mathbf{Q}(\boldsymbol{\epsilon} - \mu)) \tag{2.4}$$

D'après cette définition, un GMRF rapport avec le graphe marqué $\mathcal{G}$ peut être caractérisé par sa matrice de précision Q.

Theorem 2.5.1 *Soit* ϵ *de distribution gaussienne avec une matrice de précision* Q *symétrique et définie positive, alors pour* $i \neq j$,

$$\epsilon_i \perp \epsilon_j | \boldsymbol{\epsilon} \Longleftrightarrow Q_{ij} = 0$$

Cet théorème signifie que si le graphe d'association des données n'est pas dense (ce qui est plausible), la matrice de précision d'un GMRF est creuse.

En ce qui concerne les champs de Markov, le théorème d'Hammesley-Clifford et le lemme de Brook donnent l'existence d'une distribution jointe pour un GMRF et l'unicité du Champ de Markov résultant des conditionnelles. L'unicité est essentielle car, on voudrait s'assurer que les échantillons générés par ces lois conditionnelles approchent une seule distribution jointe.

2.5.2.2 Extension à la dimension temporelle

Pour prolonger les modèles GMRF et GF avec une dimension temporelle, les modèles hiérarchiques spatio-temporels sont les modèles les plus classiques.

$$\begin{array}{ll} \epsilon_{it}|\mathbf{X}_{it}, \epsilon_{it-1} \sim & Bernoulli(p_{it}) \\ logit(p_{it}) = & \mathbf{X}_{it}^t\beta + \epsilon_{it} \\ \epsilon_{it} = & \epsilon_{it-1} + \omega_{it} \end{array}$$

avec ω_{it} un GF ou GMRF spatialement corrélé.
Ce type de modèles est souvent utilisé pour les pathogènes avec un patho-système clair. Par exemple, Jousimo et al. (2014) l'ont utilisé pour modéliser la dynamique de la présence-absence, colonisation, et extinction de pathogène (mildew) pour chaque population échantillonnée en Finlande.

Cependant dans le cas de l'esca, on n'a pas de patho-système clair et la caractérisation des symptômes n'est pas stable : un cep malade n'exprime pas forcément le symptôme foliaire. Pour étudier l'état d'infection des plantes, on modélise la propagation de la maladie pour les données spatio-temporelles cumulatives (l'infection par l'esca, observée par l'apparition de symptôme foliaire, devient état absorbant). Néanmoins, le modèle hiérarchique auto-régressif d'ordre 1 peut quand même être utilisé pour étudier le processus d'expression des symptômes foliaires. L'expression et la re-expression des symptômes foliaires sont considérés comme un résultat d'interactions entre l'environnement, l'hôte et le pathogène, il est intéressant d'identifier les facteurs qui entrainent ce processus.

Pour modéliser la propagation de la maladie à partir des données cumulatives, Kaiser et al. (2014) proposent d'utiliser une séquence de champs aléatoires binaires. C'est-à-dire que, dans le modèle, au lieu d'intégrer un processus auto-régressif temporel stochastique, on intègre une covariable du passé qui est calculée par les données d'observations des années précédentes.

2.5.2.3 Bénéfices de la structure hiérarchique

Les modèles présentés dans cette section sont composés d'une régression logistique avec une composante spatiale ou spatio-temporelle aléatoire ϵ. Les modèles sont structurés hiérarchiquement grâce à ce processus latent non-observé. Ici nous donnons d'abord une présentation générale du modèle hiérarchique (Gaetan and Guyon, 2008), puis nous décrivons les structures hiérarchiques spatiales et spatio-temporelles.

La loi jointe des trois variables aléatoires peut toujours être décomposée par les lois conditionnelles successives:

$$[U, V, W] = [W|U, V][V|U][U], \tag{2.5}$$

Supposons que le processus d'intérêt ϵ est non-observé (latent) et que les données Z sont modélisées conditionnellement à ϵ, nous avons une structure hiérarchique de trois niveaux d'après Equation 2.5:

$$[Z, \epsilon, \theta_Z, \theta_\epsilon] = [Z|\epsilon, \theta_Z, \theta_\epsilon][\epsilon|\theta_\epsilon][\theta_Z, \theta_\epsilon] \tag{2.6}$$

Au niveau le plus bas pour le processus de données, la loi de Z est conditionnelle à ϵ et aux paramètres du modèle $(\theta_Z, \theta_\epsilon)$; au niveau intermédiaire pour le processus latent, le processus ϵ est défini conditionnellement à ses paramètres θ_ϵ; le troisième niveau du processus spécifie les incertitudes sur les paramètres du modèle.

Dans cette section le processus latent ϵ est soit un processus spatial (subsection 2.5.1) soit spatio-temporel (subsubsection 2.5.2.2). Il décrit un phénomène spatial ou spatio-temporel auto-corrélé (subsection 2.5.1) en spécifiant que ϵ est un champ Gaussien décrivant un phénomène spatial dans lequel deux points plus proches sont plus corrélés que deux points lointains. Dans la subsubsection 2.5.2.2 nous décrivons un phénomène spatio-temporel en explicitant le comportement spatial et temporel Cressie and Wikle (2011), la structure spatiale est caractérisée par un champ Gaussien et la dépendance temporelle est spécifiée par un processus auto-régressif d'ordre 1, c'est-à-dire un processus à un pas de mémoire.

L'inférence Bayésienne nous permet d'obtenir des distributions a posteriori et s'effectue directement à partir de ces spécifications conditionnelles de la structure hiérarchique.

2.5.2.4 Inférence et discussion

En comparant avec les modèles géostatistiques, les modèles auto-régressifs conditionnels (CAR) dans un cadre Bayésien permettent des calculs très pratiques. La spécification conditionnelle est bien adaptée aux algorithmes MCMC qui cherchent la loi posteriori par les lois conditionnelles (Banerjee et al., 2014).

Bien que les modèles soient faciles à implémenter, il subsiste de nombreuses difficultés théoriques et numériques : d'une part, la matrice de précision a besoin de conditions supplémentaires pour garantir son caractère défini positif. D'autre part, pour les modèles auto-régressifs sans propriété de Markov, la matrice de covariance ainsi que la matrice de précision sont de rang plein, l'inférence peut être très lourde à calculer, notamment pour le paradigme bayésien. De plus on doit générer la simulation de ces matrices pour chaque itération.

Le fait que la matrice de précision soit creuse dans de nombreux cas, donne un avantage à utiliser ce modèle pour l'inférence car cela réduit le coût de calcul. En fait, des opérations d'algèbre linéaire peuvent être effectuées en utilisant des méthodes numériques pour matrices creuses, ce qui entraîne un gain de calcul considérable. Par exemple, la factorisation de la matrice, qui exige habituellement $O(n^3)$ pour une matrice dense, se réduit à $O(n)$, $O(n^{3/2})$ et $O(n^2)$ pour les matrice creuses de GMRFs temporelle, spatiale et spatio-temporelle respectivement (Cameletti et al., 2013).

De plus, les propriétés de calcul de GMRFs sont améliorées en utilisant Integrated Nested Laplace Approximations (Rue et al., 2009) pour l'inférence bayésienne. Cette méthode donne une approximation de la distribution a posteriori rapide et assez précise.

Cependant, ce type de modélisation via les lois conditionnelles présente également des inconvénients importants. Au contraire des modèles géo-statistiques, la spécification via le graphe de voisinage ne garantit pas l'égalité des variances marginales, ou l'égalité des covariances pour chaque paire de voisins, sauf pour certaines structures de graphe très spéciales (Besag and Kooperberg, 1995, Lavigne, 2013).

Récemment, Lindgren et al. (2011) ont proposé d'adapter les GMRF aux données observées sur un champ continu. Ils utilisent une représentation d'éléments finis pour définir un champ de Matérn comme la combinaison linaire des fonctions de la base définies sur une triangulation du domaine. Cette présentation combine GMRF et Champ Gaussien en utilisant les équations aux dérivées partielles stochastiques.

Pour le cas de l'esca, on dispose d'un jeu de données spatio-temporelles très large. Donc le temps de calcul est un des facteur les plus importants pour l'application. La méthode INLA (Integrated Nested Laplace Approximations), premièrement introduit par Rue et al. (2009) et utilisant des SPDE, donne une inférence très pratique et rapide pour les modèles GMRFs. Pour notre travail de thèse, nous allons l'utiliser pour modéliser la maladie. Ces études sont présentées dans le chapitre 5.

2.5.3 Modèles auto-logistique markoviens

L'autre manière de modéliser la dépendance spatiale pour une régression logistique est d'appliquer directement la spécification auto-régressive conditionnelle markovienne à la forme du modèle logistique. Par conséquent, le processus binaire Z n'est plus modélisé implicitement, comme dans le cas de modèles logistiques avec un processus GMRF ou GF latent continu, mais plutôt modélisé explicitement par un modèle auto-logistique. L'avantage du modèle auto-logistique est que la spécification de la dépendance ne fait pas intervenir la fonction lien et peut donc fournir un modèle plus direct ou intuitif (Haran, 2011).

2.5.3.1 Contexte : Champ de Markov

Nous supposons spécifié un ensemble de lois conditionnelles complètes pour Z_i tel que:

$$p(z_i|z_j, i \neq j) = p(z_i|z_j, j \in N_i) \tag{2.7}$$

où N_i est le voisinage spatial de l'unité i. L'idée est d'utiliser la spécification locale pour recoller ces lois conditionnelles données à l'équation (2.7) et déterminer la distribution jointe que l'on appelle un champ de Markov.

La spécification de Gibbs est une famille générale de lois conditionnelles, qui se recollent sans condition. Elles sont caractérisées par les potentiels. Ici on donne une présentation rapide avec quelques notions introductives extraites de Gaetan and Guyon (2008), Hardouin (2011).

Definition 2.5.2 (Clique) *Une clique est un ensemble d'indices de S tel que les éléments de la clique sont voisins des autres.*

Definition 2.5.3 (Potentiels, énergie et spécification de Gibbs) *Soit Φ le potentiel défini sur un ensemble $\mathcal{A}$ de parties de S tel que :*

- *Un potentiel d'interaction est une famille $\Phi = \{\Phi_A, A \in S\}$ d'applications mesurables $\Phi_A : \Omega_A \longmapsto \mathbb{R}$ telle que, pour toute partie $\Lambda \in S$, la somme suivante existe*
$$U_\Lambda^\Phi(z) = \sum_{A \in S : A \cap \Lambda \neq \emptyset} \Phi_A(z)$$
- *$U_\Lambda^\Phi(z)$ est la fonction d'énergie, elle est finie sur $\Lambda \subset S$.*
- *$\int_{E^\Lambda} exp_\Lambda^\Phi(z_\Lambda, z^\Lambda) dz_\Lambda < +\infty$*

Alors les lois conditionnelles $\{\pi_\Lambda^\Phi(\cdots|x^\Lambda)\} = (\int_{E^\Lambda} exp_\Lambda^\Phi(z_\Lambda, z^\Lambda) dz_\Lambda)^{-1} exp(U_\Lambda^\Phi(z))$ sont cohérentes.

D'après ces définitions, si Z est un champ de Gibbs sur un ensemble de sites S fini, sa distribution jointe s'écrit de la façon suivante :

$$\pi(\mathbf{z}) = C^{-1} exp\{\sum_{i \in S} \phi_i(z_i)\} = C^{-1} expU(z)$$

où $C = \sum_{z \in \Omega} expU(z) < \infty$ et C est la constante de normalisation.

Le théorème d'Hammersley-Clifford (cf (McGrory et al., 2009) pour une démonstration) montre que si on a un MRF, alors l'équation (2.7) détermine une distribution jointe unique, et il existe une famille Φ définie sur les cliques de $\mathcal{G}$ telle que cette distribution jointe est la loi d'un champ de Gibbs de potentiels Φ (la réciproque est vraie), sous la condition de positivité qui stipule que $\pi_A(x_A|x^A) > 0$ pour toute partie A de S et configuration z de Ω.

2.5.3.2 Champ de Markov binaire : modèle auto-logistique markovien

Les auto-modèles de Besag particuliers dérivent de deux hypothèses : la première porte sur les cliques qui sont au plus d'ordre 2. En conséquence, l'énergie s'écrit comme une somme de potentiels portant sur les singletons et de potentiels portant sur les paires. La seconde hypothèse caractérise les lois conditionnelles qui doivent appartenir à une famille exponentielle :

Theorem 2.5.2 *Soit une famille de lois conditionnelles $\pi(\cdots|Z_{-i})$ d'un champ Markov de loi π, appartenant à la famille exponentielle*

$$\pi(z_i|z_j, j \in N_i) = exp[A_i(z_j)B_i(z_i) + C_i(z_i) + D_i(z_j)], \quad i = 1, ..., n \tag{2.8}$$

Alors il existe α_i et $\beta_{ij} = \beta_{ji}$ tels que

$$A_i(z_j, j \neq i) = \alpha_i + \sum_{j \neq i} \beta_{ij} B_j(z_j) \tag{2.9}$$

et $\phi^{(1)}(z_i) = \alpha_i B_i(z_i) + C_i(z_i)$, $\phi^{(2)}(z_i, z_j) = \beta_{ij} B_i(x_i) B_j(x_j)$ Inversement, les lois conditionnelles exponentielles vérifiant les équations (2.8) et (2.9) se recollent en une loi qui est un champ de Markov de potentiel la famille $\Phi = \phi_i, \phi_{ij}$.

Si $\phi_{ij} = z_i z_j$, et $\beta_{ij} = \beta_{ji}$, alors Z de loi π est un auto-modèle markovien.

$$\pi(z) = C^{-1} exp\{\sum_{i \in S} \phi_i z_i + \sum_{i \sim j} \beta_{ij} z_i z_j\}$$

où $i \sim j$ signifie que i et j sont voisins et C est une constante normalisée.

Maintenant on considère un champ de Markov binaire, pour chaque i, la loi conditionnelle $\pi_i(\cdot|z_{-i})$ est un modèle de logit de paramètres $\theta_i(x_i) = \{\alpha_i + \sum_{j \in N_i} \beta_{ij} z_j\}$,

$$\begin{aligned} \theta_i(z_i) &= \{\alpha_i + \textstyle\sum_{j \in N_i} \beta_{ij} z_j\} \\ \pi_i(\cdot|z_{-i}) &= \frac{exp z_i \{\alpha_i + \sum_{j \in N_i} \beta_{ij} z_j\}}{1 + exp\{\alpha_i + \sum_{j \in N_i} \beta_{ij} z_j\}} \end{aligned}$$

Si pour tout $i \neq j$, $\beta_{ij} = \beta_{ji}$, ces lois conditionnelles se recollent en loi jointe d'énergie U:

$$U(z) = \sum_i \alpha_i z_i + \sum_{j \in N_i} \beta_{ij} z_i z_j$$

Un tel modèle s'appelle modèle auto-logistique Markovien.

2.5.4 Extension à la dimension temporelle

Plusieurs études ont été faites pour prolonger le modèle auto-logistique de Besag à la dimension temporelle. Guyon and Hardouin (2002) ont étudié un modèle paramétrique semi-causal appelé CMCM : Chaîne de Markov de Champ de Markov, défini sur un espace d'états général E et un ensemble fini de sites $S = \{1, 2, ..., n\}$. Le modèle est défini comme suit : $\mathbf{Z} = (Z_{it}, t \in N)$ est une chaîne de Markov sur $\Omega = E^S$ et $Z_t = (Z_{it}, i \in S)$ est, conditionnellement au passé Z_{t-1}, un champ de Markov sur E^S. Pour simplifier, nous considérons des dynamiques basées sur des chaînes de Markov d'ordre 1 homogènes dans le temps, mais aucune stationnarité

dans l'espace n'est supposée. Nous notons aussi $y = Z_{t-1}$ et $x = Z_t$ deux motifs successifs dans la chaîne,
On écrit la probabilité de transition (de passer d'un motif y à un motif x en un pas de temps) via une énergie conditionnelle c'est-à-dire :

$$P(y,x) = C^{-1}(y) exp U(x|y) \tag{2.10}$$

$$U(x|y) = \sum_{A \in \mathcal{C}} \Phi_A(x) + \sum_{B \in C^{-1}, A \in \mathcal{C}} \Phi_{B,A}(y,x) \tag{2.11}$$

- Les potentiels d'interaction instantanée $\{\Phi_A, A \in \mathcal{C}\}$: $\mathcal{C}$ définit le graphe non-orienté des voisins instantanés $\mathcal{G}(\mathcal{C})$.
- Les potentiels d'interaction temporelle $\{\Phi_{B,A}(z_{t-1}, z_t), B \in \mathcal{C}^{-1}, A \in \mathcal{C}\}$. $\mathcal{C}^-$ définit un graphe orienté $\mathcal{G}^-$: $< j, i >^-$ pour $j \in B$ et $i \in A$ traduit que le site j au temps $(t-1)$ a une influence sur le site i au temps t. En général, $< j, i >^-$ n'implique pas $< i, j >^-$.

2.5.4.0.1 Dynamique auto-logistique :

$$P(y,x) = C^{-1}(y) exp\{\sum_{i \in S}(\sigma + \alpha \sum_{j \in S: <j,i>^-} y_j) x_i + \beta \sum_{i,j \in S: <i,j>} x_i x_j)\}$$

Zhu et al. (2005) généralisent aussi une forme de distribution jointe pour un modèle auto-logistique spatio-temporel qui spécifie la propriété Markov sur un voisinage général spatio-temporel. Plusieurs de ces travaux ont été basés sur cette généralisation. On parlera de ces travaux dans le chapitre 6 pour proposer une généralisation nous aussi.

2.5.4.1 Quelques méthodes d'échantillonnage pour les modèles auto-logistiques

Un algorithme de chaîne de Markov est dit exact si il renvoie un tirage qui suit exactement la distribution cible donnée lorsque l'algorithme est terminé. Généralement l'algorithme a un temps d'exécution qui est aléatoire, mais fini p.s. Moller and Waagepetersen (2003) ont explicité l'avantage du perfect sampling par rapport aux algorithmes MCMC : tout d'abord, on n'a plus besoin d'évaluer un burn-in approprié. Cette période de burn-in consiste à ignorer un nombre d'échantillons au début de l'échantillonnage, car la distribution stationnaire de la chaîne de Markov est la distribution souhaitée sur les variables, mais cela peut prendre un certain temps pour atteindre la distribution stationnaire, ce qui peut parfois être une tâche difficile. Ensuite, un échantillonnage i.i.d. est disponible par simulation parfaite, de sorte que par exemple, les variances asymptotiques des estimations de Monte Carlo peuvent être calculées très facilement.

Des algorithmes de "perfect sampling" (Moller, 1999, Propp and Wilson, 1996) donnant des échantillons de la distribution stationnaire des chaînes de Markov existent pour certains modèles tels que le modèle auto-logistiques. Les algorithmes de "perfect sampling" sont d'attrayantes alternatives aux algorithmes MCMC réguliers

mais sont généralement très couteux en calcul par rapport aux algorithmes MCMC. Ils ont été utilisés pour les modèles auto-logistiques spatiaux ou spatio-temporels dans plusieurs articles (Hughes et al., 2011, Zhu et al., 2008, Zheng and Zhu, 2008, Wang and Zheng, 2013).
Le "perfect sampling" est basé sur le couplage du passé (CFTP) Propp and Wilson (1996). Un échantillonneur CFTP pour un modèle auto-logistique peut être construit de la façon suivante. Nous voulons simuler exactement le modèle auto-logistique:

$$\mathbb{P}(Z_{i,t} = 1|\mathbf{Z}_{-i,t}, \mathbf{Z}_{-i,t-1}, \theta) = p_{it} = \frac{exp(\beta + \rho_1 \sum_{j\in N_i} \omega_{ij} Z_{jt} + \rho_2 \sum_{j\in N_i} \omega_{ij} Z_{jt-1})}{1 + exp(\beta + \rho_1 \sum_{j\in N_i} \omega_{ij} Z_{jt} + \rho_2 \sum_{j\in N_i} \omega_{ij} Z_{jt-1})}$$

où $\theta = \{\beta, \rho_1, \rho_2\}$. Le modèle est interprétable si $\rho_1 > 0$, $\rho_2 > 0$, c'est-à-dire que la fonction de répartition, F_{it} de Z_{it} diminue en $\sum_{j\in N_i} \omega_{ij} Z_{jt}$ et $\sum_{j\in N_i} \omega_{ij} Z_{jt-1}$. Notons que dans ce cas la fonction F_{it} est très simple :

$$F_{it}(z) = \begin{cases} 0 & \text{si} \quad z < 0 \\ 1 - p_{it} & \text{si} \quad 0 < z < 1 \\ 1 & \text{si} \quad z \geq 1. \end{cases} \tag{2.12}$$

Comme nous avons besoin de ρ_1, ρ_2 non-négatifs, p_{it} augmente avec $\sum_{j\in N_i} \omega_{ij} Z_{jt}$ et $\sum_{j\in N_i} \omega_{ij} Z_{jt-1}$, et alors F_{it} diminue en $\sum_{j\in N_i} \omega_{ij} Z_{jt}$ et $\sum_{j\in N_i} \omega_{ij} Z_{jt-1}$. Pour un tel modèle, le CFTP se déroule comme ci-dessous (Moller (1999)).

Notons $\mathbf{L}_M(m, k)$ et $\mathbf{U}_M(m, k)$ la $k-$ième observation à l'instant m des chaines supérieure et inférieure, respectivement, où ces chaines sont commencées à l'instant M du passé. Et $k = n(t-1) + i, n =$ la taille du réseau correspondant au i-ième site dans un lattice à l'année t. Nous fixons $M < 0$ et mettons $\mathbf{L}_M(M, \cdot) = 0$ et $\mathbf{U}_M(M, \cdot) = 1$. Alors, non avons:

$$\begin{aligned} \mathbf{L}_M(m, k) &= F_{it}^{-1}(\mathbf{R}(t, i))|\mathbf{L}_M(c, n(t-2) : n(t-1) + i - 1), \mathbf{L}_M(c, n(t-1) + i + 1 : nt) \\ \mathbf{U}_M(m, k) &= F_{it}^{-1}(\mathbf{R}(t, i))|\mathbf{U}_M(c, n(t-2) : n(t-1) + i - 1), \mathbf{U}_M(c, n(t-1) + i + 1 : nt) \end{aligned} \tag{2.13}$$

où les $\mathbf{R}(t, i)$ sont des variables aléatoires uniformes standard indépendantes et

$$F_{it}^{-1}(p) = \begin{cases} 0 & \text{if} \quad p \leq 1 - p_{it} \\ 1 & \text{if} \quad p > 1 - p_{it}. \end{cases} \tag{2.14}$$

Si $\mathbf{L}_M$ et $\mathbf{U}_M$ fusionnent à l'instant $m_0 \leq 0$, nous retournons $\mathbf{L}_M(0, \cdot)$ comme un échantillon de $\pi(\mathbf{Z}_t|\mathbf{Z}_{t-1}, \theta)$, la loi jointe des Z_i conditionnelles à l'année précédente. Sinon, nous doublons M et recommençons, utilisons de nouvelles variables aléatoires uniformes pour $M, M+1, ..., \frac{M}{2} - 1$, mais nous réutilisons les variables aléatoires générées précédemment pour des temps $\frac{M}{2}, \frac{M}{2} + 1, ..., -1$.

2.5.4.2 Un modèle auto-logistique de régression spatio-temporelle

L'objectif est de construire un modèle auto-logistique de régression spatio-temporelle pour expliquer les données de la maladie. Par rapport aux GMRF, il y a moins d'études sur les modèles auto-logistiques et, de plus, les modèles auto-logistiques

avec covariables présentent des difficultés d'interprétation. Dans le chapitre 6 de la thèse, nous avons développé un nouveau modèle auto-logistique spatio-temporel, centré en deux étapes, avec régression et montré son intérêt par rapport aux modèles précédents.

2.5.4.3 Application aux données

Les premiers modèles présentés dans cette section sont bien sûr appliqués aux données de maladie de l'esca. La variable d'intérêt Z est la variable "expression des symptômes" ou "première expression des symptômes". L'objectif est de réunir l'ensemble des éléments de compréhension de la maladie dans un même modèle (corrélation spatiale, au rang ; autocorrélation temporelle ; ...) et de tester l'effet de cofacteurs environnementaux. Plusieurs de ces facteurs ont bien souvent été "préparés" en amont de la modélisation car il n'étaient pas connus à l'échelle du cep. Les résultats escomptés de ces modèles sont évidemment au coeur des objectifs de la thèse. Le modèle autologistique est en cours de développement pour être exploitable.

2.6 Géostatistique, krigeage

Contrairement aux modèles sur réseau qui traitent des données spatiales discrètes, la théorie géostatistique suppose que les données observées sont un échantillon de n réalisations d'un processus stochastique spatial continu indexé: $Z(s), s \in D$. La géostatistique permet de réaliser l'interpolation spatiale et comparer différentes techniques d'interpolation. Après une étape de modélisation de la variabilité spatiale à travers le variogramme, elle permet de fournir une mesure de la précision de cette interpolation en tout point de l'espace (Lejeune, 2006). L'objectif commun de la géostatistique est de prédire la valeur de $Z(\cdot)$ à des endroits non-échantillonnés en utilisant les observations disponibles. La prédiction repose sur l'estimation de la structure spatiale, particulièrement la structure du seconde ordre.

Nous présentons brièvement les principes des modèles géostatistiques classiques dans la suite. Un modèle géostatistique classique se décompose en un processus spatial par une fonction moyenne et un processus d'erreur.

$$Z(s) = \mu(s) + e(s)$$

Le processus aléatoire $e(s)$ est intrinsèquement stationnaire si ses incréments sont stationnaires:

$$\begin{aligned} E(e(s+h) - e(s)) &= 0 \\ \gamma_h &= \tfrac{1}{2} Var[e(s+h) - e(s)] \end{aligned}$$

où γ est la fonction de variogramme qui ne dépend que du décalage h ($||h||$ si e est isotropique). Un processus aléatoire stationnaire de second ordre de fonction de covariance $C(\cdot)$ est intrinsèquement stationnaire (la réciproque est fausse), de variogramme donné par:

$$\gamma(h) = C(0) - C(h)$$

Nous répétons les objectifs de l'analyse géostatistique explicitement ici : caractériser la structure spatiale par l'estimation de $\mu(s)$ et $C(\cdot)$ ou $\gamma(\cdot)$ et effectuer la prédiction par combinaison linaire des observations.

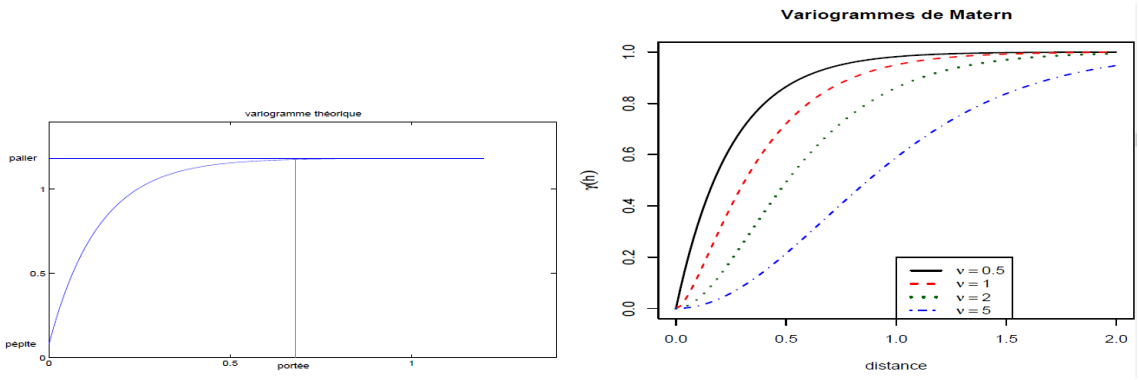


Figure 2.2

Ici on donne plusieurs modèles de variogrammes isotropiques classiquement utilisés en géostatistique :

- *Sphérique*: $\gamma(h) = \begin{cases} C(\frac{3}{2}\frac{|h|}{a} - \frac{1}{2}\frac{|h|^3}{a^3}) & \text{si} \quad 0 \leq |h| \leq a \\ C & \text{si} \quad |h| \geq a \end{cases}$
- *Puissance*: $\gamma(h) = C|h|^\alpha, \quad \alpha < 2$
- Class de Matérn

$$\gamma(h) = C[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}(\frac{2\nu^{1/2}h}{\rho})\mathcal{K}_\nu(\frac{2\nu^{1/2}h}{\rho})] \tag{2.15}$$

Où $\mathcal{K}_\nu$ est la fonction de Bessel modifiée de 3ème espèce, d'ordre ν et ν est un paramètre qui règle la régularité en 0, si $\nu = 1/2$, c'est un modèle exponentiel; si $\nu \to +\infty$, c'est un modèle Gaussien.

Quel que soit le modèle, il est intéressant de définir les paramètres du variogramme:

- Palier : $C = lim_{h\to\infty}\gamma(h)$
- Portée : a t.q. $||h|| \geq a \to C - \gamma(h) < \epsilon$
- Pépite : $\tau = lim_{h\to 0}\gamma(h)$

Si le variogramme d'un processus $\gamma(h)$ ne dépend que de la magnitude de h, le variogramme est isotropique. Avec un modèle isotrope estimé par le variogramme empirique, on peut prédire la valeur de Z à un site s_0 non échantillonné comme moyenne pondérée des valeurs observées des autres sites voisins. Et le pondération est calculée en fonction de $\gamma(h)$, on appelle cette procédure le krigeage.

Cependant, dans la pratique, les variogrammes empiriques sont souvent de forme irrégulière sans modèle de variogramme adapté. On distingue ces variogrammes anisotropes généralement en deux types :

2.6.0.3.1 Anisotropie

- Anisotropie géométrique : la portée varie selon une ellipse.
- Anisotropie zonale : le palier dépend de la direction.

Pour l'anisotropie géométrique, une solution pratique est de décomposer le variogramme selon les directions et identifier les portée associées avec ces directions. Puis nous transformons le vecteur de décalage $\vec{h}$ en un vecteur isotropique équivalent.

Dans ce travail de thèse, on utilise des approches géostatistiques (krigeage) pour affiner l'estimation des covariables à l'échelle du cep pour différentes variables. Le modèle utilise une variable auxiliaire (pour la réserve utile) ou non (l'azote, DC13).

Comme variable auxiliaire, les données de résistivité sont utilisées pour estimer les valeurs de la réserve utile du sol. Ces données présentent l'anisotropie zonale et comme il n'y a pas une solution générale pour l'anisotropie zonale, on a développé une solution adaptée à la spécificité de données, cette méthode est présentée au chapitre 4.

2.7 Discussion

Les méthodes présentées ici sont choisies en fonction des questions épidémiologiques qui se posent au sujet de l'esca et de la spécificité des données disponibles pour répondre à ces questions. Elles sont clairement reliées par une problématique de statistique spatiale mais aussi spatio-temporelle. Cependant, peu de recul dans le temps sur les données (entre 8 et 12 ans) nous oblige à être prudent avec l'aspect temporel. L'originalité de toutes ces méthodes provient de leur adaptation et de leur combinaison pour traiter les questions posées et tenir compte du format des données : adaptation des voisinages pour répondre aux questions traitées dans la méthode de Join Count, combinaison de tests multiples dans le cas de l'étude de l'indépendance spatiale (chapitre 3), la gestion de l'anisotropie et de la sparsité de la variable d'intérêt dans le chapitre 4, la souplesse des modèles INLA pour tester l'effet de différentes variables (chapitre 5) et enfin le développement d'un nouveau modèle auto-logistique au chapitre 6. Cependant, il faut encore étudier les aspects théoriques de ce modèle auto-logistique. Et comme la structure de ce modèle auto-logistique est complexe, la simulation et l'estimation pour une base de données de grande taille sont très longues. Nous testons ce modèle sur les simulations des données de lattice 20×20 sur environ 10 ans. Nous souhaitons appliquer ce modèle à des données de maladie de taille plus grande.

Chapter 3

Analyse de la distribution spatiale de l'esca au cours du temps

Introduction

L'émergence et le développement de l'esca provient d'interactions complexes entre différents types et quantités d'inocula, ainsi que d'autres facteurs, tels que la physiologie de la plante, un environnement abiotique (climat et sol) et les pratiques culturales (Bertsch et al., 2013, Lecomte et al., 2011, Murolo and Romanazzi, 2014). L'identification et le classement des facteurs qui dirigent la propagation de l'esca restent un sujet d'étude majeur dont l'objectif est de comprendre les étapes fondamentales de l'épidémie et de trouver des pratiques tactiques ou stratégiques de contrôle de la maladie.

Des méthodes de statistique spatiale et de modélisation mathématique spatiale ont été utilisées pour décrire la distribution et la propagation de l'esca dans les vignes. Plusieurs analyses de statistique spatiale ont révélé que ce qui prévaut dans la plupart des parcelles est une répartition aléatoire (Cortesi et al., 2000, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Surico et al., 2000b), ce qui suggère que la propagation de l'esca dans une parcelle est essentiellement aérienne, via des sources externes et/ou internes d'inoculum, plutôt que la conséquence d'une contamination par les outils d'élagage le long des rangs (Surico et al., 2000b). Cependant, dans certaines parcelles, des motifs agrégés de vignes exprimant l'esca ont été observés (Edwards et al., 2001, Pollastro et al., 2009, Surico et al., 2000b), ce qui amène à la question d'une éventuelle seconde propagation de l'esca au cours du temps. Seules deux études, combinant la distribution spatiale et la dynamique temporelle de l'esca Stefanini et al. (2000), ont observé à partir de leurs données une légère augmentation de la probabilité de l'expression d'un symptôme lorsque des vignes infectées se trouvent dans un voisinage (le long du même rang dans la culture), mais ces études n'ont pas traité de longue période temporelle. Plus récemment, Zanzotto et al. (2013) ont étudié la seconde propagation de l'infection de l'esca en se basant sur plusieurs directions de propagation (le long des rangs ou perpendiculairement aux rangs). Leurs résultats ont montré une probabilité plus importante d'expression de l'esca au cours du temps et une plus grande propagation selon les rangs plutôt que selon les rangs adjacents.

L'analyse spatiale utilisant des données enregistrées sur plusieurs années devrait

nous aider à distinguer les différentes composantes en caractérisant l'augmentation de la taille des clusters de vignes symptomatiques et en testant la relation spatiale entre les vignes déjà symptomatiques et celles symptomatiques pour la première fois. Les résultats peuvent mener à une compréhension de l'implication des vignes anciennement symptomatiques localisées le long des rangs, renforçant le besoin de mesures sanitaires. Plus généralement, une telle étude nous permet de décrire l'évolution de la structure spatiale de la maladie au cours du temps et la dépendance statistique entre les vignes symptomatiques pour guider les futures recherches de facteurs de risque. De plus, les résultats devraient aider à définir la meilleure échelle spatiale pour modéliser la propagation de l'esca dans une parcelle. Nous présentons dans ce chapitre une méthode statistique permettant d'analyser la configuration spatio-temporelle de la maladie de l'esca pour les vignobles en France d'une manière exploratoire.

Résumé

Pour évaluer la capacité de l'esca à se propager dans les vignes de la région de Bordeaux, des enregistrements sur 8 années de 15 parcelles de la région et contenant chacune entre 1 200 et 2 300 vignes voisines de Cabernet Sauvignon, ont été utilisés pour réaliser des analyses statistiques spatiales. Plusieurs tests non-paramétriques, basés sur la statistique "join count" et sur des méthodes de permutation ont été développés pour caractériser la structure spatiale de vignes symptomatiques en étudiant plusieurs hypothèses, une propagation sans direction privilégiée ou une propagation le long des rangs. Parmi les parcelles, un grand nombre de motifs spatiaux, d'aléatoires à fortement structurés, avec de nombreux taux de prévalence augmentant sur la durée, ont été observés. Dans quatre parcelles, la complexité du motif de distribution de l'esca indiquait différents niveaux d'agrégation. Au contraire, dans les autres parcelles, seuls de petits agrégats composés de deux ceps adjacents ont été observés. De plus ces agrégats étaient localisés le long des rangs et sans augmentation de taille dans le temps, excepté sur une parcelle. Une analyse de la dépendance spatiale entre les ceps précédemment et nouvellement infectés dans un voisinage à l'ordre k (k=1 à 5) a montré que, pour 6 des 15 parcelles, les ceps nouvellement symptomatiques étaient situés à proximité des ceps précédemment symptomatiques, mais sans orientation ni ordre de voisinage privilégiés. Ces résultats indiquent une légère, sinon aucune, propagation secondaire à partir des ceps symptomatiques.

Le travail présenté dans la suite correspond à un travail en collaboration avec Florent Bonneu[1], Joël Chadoeuf[2], Delphine Picart[3] et mes co-encadrantes. Il a été soumis en juin dernier à Phytopathology.

[1]Université d'Avignon
[2]INRA Avignon
[3]INRA Bordeaux

Spatial and Temporal Pattern Analyses of Esca Grapevine Disease in Vineyards in France

Shuxian Li[a,b] , **Florent Bonneu**[c], **Joël Chadeuf**[d], **Dominique Picart**[e] **Anne Gégout-Petit**[f] and **Lucia Guerin-Dubrana**[a,b]

[a]Université de Bordeaux, ISVV, UMR-1065 INRA
[b]Bordeaux Sciences Agro, Gradignan, France
[c] Université d'Avignon (Laboratoire de Mathématiques-EA2151), F-84914 Avignon, France
[d] INRA –Statistics, UR1052, F-84914 Avignon, France
[e] INRA UMR ISPA, F-33140 Villenave d'Ornon, France
[f]Institut Elie Cartan, Université de Lorraine, INRIA BIGS Nancy Grand-Est, Nancy, France

Abstract

To assess the capacity of esca to spread within vineyards of the Bordeaux region, over 8 years of annual records, containing between 1,200 and 2,300 contiguous Cabernet Sauvignon vines from 15 mature vineyards, were used for spatial statistical analyses. Several non-parametric tests, based on join count statistics and on permutation methods, were developed to characterize the spatial structure of esca-symptomatic vines based on non-directional or along-row directional assumptions. Among vineyards, a large range of spatial patterns, from random to strongly structured, associated with various prevalence rates that increased over time, were observed. In four vineyards, the complex esca distribution pattern indicated different levels of clustering. By contrast, in other vineyards, only small clusters of two adjacent symptomatic vines were observed, and they were localized along rows, without enlargement over time, except in one vineyard. An analysis of spatial dependence between previously- and newly-symptomatic vines within k-order neighborhoods ($k = 1$ to 5), showed, for 6 of the 15 vineyards, that the newly-symptomatic vines were located close to previously-infected vines, without a favored orientation or neighbor order. These results indicated only a slight, or no, secondary spread from symptomatic vines.

Additional keywords: fungal disease; join count statistics

3.1 Introduction

Esca, one of the grapevine trunk diseases (GTDs), causes extensive damage in vineyards worldwide, resulting in major economic losses (Bertsch et al., 2013, Mugnai et al., 1999). In France, it is considered by winegrowers as a great threat. The National Grapevine Trunk Disease Survey, conducted from 2003 to 2008, showed that the mean incidence of esca in vineyards by region varied from 2.3 % to 8.2 % (Bruez et al., 2013) and led to vine decline and death.

Grapevine esca, also called esca proper, is defined as a complex dieback disease associated with pathogenic fungi that degrade the woody part of the vine that also exhibits discolored foliar symptoms (Mugnai et al., 1999, Surico et al., 2008). Two forms of the disease are described as follows: a chronic form, characterized

by the "tiger stripe" appearance of the affected leaves, and an acute form (also called apoplexy), presenting the dieback of one or more shoots, combined with leaf drying (Mugnai et al., 1999, Surico et al., 2008) in midsummer. Esca foliar symptoms, observed more frequently in 12–18-year-old adult vines (Fussler et al., 2008), are commonly associated with two vascular pathogenic *ascomycetes*, *Phaeomoniella chlamydospora* and *Phaeoacremonium spp.*, growing within the wood and producing specific necroses, central necrosis and black punctuations. Additionally, in European regions, it is also associated with *Fomitiporia mediterranea*, which causes white decay (Fischer, 2002, Mugnai et al., 1999). Other Ascomycete pathogenic fungi, *Eutypa lata*, causing the well-known trunk disease Eutypa dieback, and Botryosphaeriacea fungi may be involved in esca, as they may be isolated from sectorial necroses in vines showing external foliar esca symptoms (Larignon and Dubos, 1997, White et al., 2011). Lecomte et al. (2012) showed that some foliar symptoms of vines showing esca symptoms, overlapped with some those of black dead arm (BDA), a GTD associated with several species of Botryosphaeriaceae (Larignon and Dubos, 2000). All of these fungi are disseminated, over long or short distances, and cause infection using different methods, making the study of disease spread difficult. For example, the *Ascomycetes*, *P. chlamydospora* and *Phaeoacremonium spp.*, which are endophytic fungi, can be transmitted via propagation materials. Consequently, the infected young plants may represent a primary source of inoculum introduced into the vineyards (Gramaje and Armengol, 2011). Over the vine's lifespan, endo-inocula, like pycnidia and conidiphores of *P. chlamydospora*, produced at the vines surface (bark and pruning wounds), are airborne dispersed (Larignon and Dubos, 2000) or may be transmitted via cutting tools (Agustí-Brisach et al., 2014) and/or insects (Moyo et al., 2014) to the other vines. However, the aerial inoculum of *F. mediterranea* is assumed to come from external sources because this fungus is never present in young nursery plants (Gramaje and Armengol, 2011).

The emergence and development of esca results from complex interactions between types and amounts of inocula, and also other factors, such as plant physiology, abiotic environment (climate and soil) and cultural practices (Bertsch et al., 2013, Lecomte et al., 2011, Murolo and Romanazzi, 2014). The identification and ranking of factors that drive esca spread remain a major subject of study in an attempt to understand the stages underlying the epidemic and to deploy tactical or strategic control management practices for this disease.

Spatial statistics and spatial mathematical modelling have been used to describe esca vine distribution and spread. Several spatial statistical analyses, based on annual data from different vineyards in European countries, revealed that the random distribution of esca prevails in most vineyard situations (Cortesi et al., 2000, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Surico et al., 2000b), suggesting that the spread of esca within the vineyard is mainly airborne, via external and/or internal sources of inoculum, rather than due to contaminated pruning tools along the vine agronomic row (Surico et al., 2000b). However, aggregated patterns of esca expressing vines have been observed in certain vineyards situation (Edwards et al., 2001, Pollastro et al., 2009, Surico et al., 2000b), leading to the question of whether there is a secondary spread of esca over time. Only two studies, using spatial modelling and several years of recorded data, have investigated this question. Stefanini et al. (2000) used a parametric statistical model to determine the

probability of vines expressing esca foliar symptoms, taking into account the symptomatic and asymptomatic vines in the neighborhood. The authors observed from their data, for one vineyard in Italy, a slight increase in the probability of symptom outbreak when infected vines were present in close vicinity (along the same agronomic row) but they did not explore a long time period. More recently, Zanzotto et al. (2013) analyzed 17 years of data from one single vineyard surveyed from planting. Using Bayesian spatio-temporal methods, they investigated the secondary spread of esca infection based on different spatial directions (along- or across-rows). Their results showed a higher probability of esca expression over time and a greater spread along rows, rather than among adjacent rows.

To our knowledge, apart from the studies of Stefanini et al. (2000) and Zanzotto et al. (2013), which both used mathematical models applied to spatio-temporal data from only one vineyard, there are no other studies combining the spatial distribution and temporal dynamics of esca. Spatial analyses using data recorded over several years should help us to differentiate the components by characterizing symptomatic vine cluster enlargement over time and by testing the spatial relationship between the previously- and newly-symptomatic vines. The results may lead to an understanding of the involvement of previously symptomatic vines localized on rows, reinforcing the need for specific sanitary measures. More globally, such a study allows us to describe the scale of the disease's spatial structure over time and the statistical dependence among diseased vines to guide future risk factor research. Also, the results should help define the most relevant distance scale for modeling esca spread within a vineyard.

To address these issues, we analyzed the distribution patterns of vines expressing esca and esca spread over 8 years (6 years for one vineyard), using recorded data of esca symptomatic and non-symptomatic vines of the same susceptible cultivar, Cabernet Sauvignon, planted during the same period from 15 mature vineyards. Several types of statistical methods are available to measure spatial structure and intensity using binary data in epidemiological domains (Madden et al., 2007). During the last two decades, a range of different spatial statistical methods were applied to esca based on the studies objective. For instance, to test the overall aggregation of esca in vineyards, Reisenzein et al. (2000) compared the expected random and observed distributions directly using the Chi-square test, while Cortesi et al. (2000) used nearest neighbor methods, which compared the average distances between observed symptomatic vines with random distributions. To quantify the spatial pattern, Surico et al. (2000b) combined several methods, dispersion index, quadrat method and ordinary runs, to detect the aggregation along the row. The two-dimensional distance class method (Surico et al., 2000b) proposed by Gray et al. (1986) was applied to quantitatively describe the disease at each distance class. In our study, another non-parametric method based on join counts (JCs) and permutation tests (Monte Carlo) was used. This general method is well adapted for data exploration because no assumptions and no prior knowledge of the processes of disease spread are needed. Moreover, the method allows for the testing of different hypotheses using a large dataset comprised of more than a thousand mapped data points, from an irregular lattice. In our cases, by developing tests based on JCs to analyze the spatial dependence over time between esca vines separated by a specific distance, we attempted to address different epidemiological questions on esca

disease.

The objective of our study was to describe how the disease spreads in the Bordeaux regions commercial vineyards by answering the following questions: What is the rate of the disease's temporal progress and how does it vary among vineyards? Is the esca disease always randomized or not? How does the structure of its spatial distribution vary over time? The capacity of local spread along the row or in all direction over time is investigated, as well as the relative importance of local disease spread from vines that have previously expressed foliar symptoms.

3.2 Materials and methods

3.2.1 Monitored vineyard

Fifteen commercial vineyards in the Bordeaux region, belonging to several owners, were used to monitor esca disease for 8 consecutive years, from 2004 to 2011 (except for one vineyard that was monitored for 6 years, from 2006 to 2011) (Table 3.1). The 15 vineyards were ordered from 1 to 15, depending on the disease prevalence in 2011 (except for vineyard 13). All of the vineyards were planted between 1985 and 1990 with the cultivar, Cabernet Sauvignon (Vitis vinifera 130 L.), and were trained in accordance with the Guyot method. The vines were grafted onto a variety of rootstocks, depending on the particular vineyard. Within plots, the distance between rows varied from 1–3.5 m, and the distances between vines within each row from 0.8–1.2 m. The number of living vines per vineyard, monitored at the beginning of the observation period, varied from 1,289 to 2,281. As the vineyard plots did not form a regular lattice, with rows perpendicular to columns, the alignments of the first vine of each row, the position (x, y coordinates in meters) of each surveyed vine within each plot, was estimated on an orthogonal basis, using both inter-row and inter-vine distances.

3.2.2 Data collection and temporal disease progress

In 2004 (in 2006 for one vineyard), at the end of August, all of the contiguous vines from each of the 15 vineyards were individually surveyed for foliar esca expression. Symptoms of esca included "tiger-stripe" patterns (chronic form), and the wilting of some or all the vine branches (acute form). Newly-planted, and dead or missing plants, were also recorded. From 2005 (or 2007) to 2011, foliar esca symptoms were recorded by individual vine surveys.

The prevalence of esca, defined as the ratio between the total number of cumulated vines exhibiting esca symptoms since 2004 (or 2006) and the number of living vines counted in 2004 (or 2006), was then calculated.

3.2.3 Spatial point pattern analysis based on JC statistics

Spatial point pattern analysis based on JC statistics. The spatial patterns of symptomatic vines in each year were analyzed using statistical tests based on JC statistics adapted for binary data (Moran, 1948).

3.2.3.1 Distance tests.

A set of statistical tests was designed for a lattice to detect dependence between symptomatic vines separated by a certain distance, regardless of the orientation or location on the same row, respectively, called the Omni-directional Distance test and Row Distance test. For a lattice of grapevines l, the spatial locations of vines are $s_{il} = (x_{il}, y_{il})$, the coordinates of vine i. We denote the status of vine i associated with the spatial unit s_{il} at time t by c_{ilt}. c_{ilt} is defined by:

$$c_{ilt} = \begin{cases} 1 & \text{if vine i of lattice l expresses esca symptoms at or before year t.} \\ 0 & \text{if vine i of lattice l didn't express esca symptoms at year t or before} \end{cases}$$

To determine whether esca diseased vines in a pair of sites separated by a certain distance are dependent, the number of symptomatic vine location pairs separated by this distance are counted and compared with the expected number of pairs under the null hypothesis of random distribution. For each vineyard, the JC statistic is determined, regardless of the orientation or the along-row direction, for each distance class $(r-1, r]$ (bigger than $r-1$ and smaller than or equal to r), with r varying from 1 or 2 to 15 m, depending on the smallest distance between two vines in the vineyard. These consecutive distances cover the range of distances between adjacent vines to the vineyard radius.
For the Omnidirectional Distance test, the JC statistic is defined by:

$$JC^{omni}_{distance}(l, t, r) = \frac{1}{2} \sum_{i \neq j} c_{ilt} c_{jlt} \omega^{omni}_{lij}(r)$$

$$\omega^{omni}_{lij}(r) = \mathbf{1}_{\{r-1 < |s_{il} - s_{jl}| \leq r\}}$$

with $\omega^{omni}_{lij}(r)$, a weight with value 1 if the disease between the two plants at sites s_{il} and s_{jl} belongs to the distance class $(r-1, r]$ and 0 for all other cases. The null hypothesis, H_0, is that the cases of esca are distributed at random in the lattice, and the alternative hypothesis, H_1, is that pairs of symptomatic vines belonging to the distance class are significantly more frequent than expected from a random distribution.
A similar JC statistic is computed for the row distance test, $JC^{row}_{distance}$, using a new weight defined by $\omega^{row}_{lij}(r)$,

$$\omega^{row}_{lij}(r) = \mathbf{1}_{\{r-1 < |s_{il} - s_{jl}| \leq r \text{ and } x_{il} = x_{jl}\}}$$

To provide a direct evaluation of the unilateral right-sided departure of the observed pattern from H_0, we computed the same JC statistic on the observed data and on $1,000$ simulations by fixing the missing, dead and one-year-old vines, and by reallocating the remaining locations randomly. An approximate P value was computed in corrected form according to Phipson and Smyth (2010).

3.2.3.2 Neighbor tests.

To study the local dependence between newly-diseased plants and previously-diseased plants, based on the order of neighbors located around or on the same row, two complementary tests called the Omni-directional Neighbor test and Row Neighbor test

respectively, were developed.

At each year t, the spatial dependence between previously- and newly-symptomatic vines situated in a close neighborhood was analyzed by considering the k-th neighborhood order, with k varying from 1 to 5. The term "previously" defines vines that had previously expressed esca symptoms before t, and "newly" defines vines that expressed esca symptoms at t for the first time. To determine k-th neighborhood order, regardless of the direction or the row, the non-equal distance between two adjacent rows and between two adjacent vines along the row were taken into account, and an elliptical band, shaped by the distance between vines located along two adjacent rows Δx_l and the distance between two consecutive vines along the same row Δy_l, is considered Figure 3.1. Two vines at sites s_{il} and s_{jl} of the same vineyard l are neighbors of order k if their coordinates satisfy $(k-1)^2 < \frac{(x_{il}-x_{jl})^2}{(\Delta x_l)^2} + \frac{(y_{il}-y_{jl})^2}{(\Delta y_l)^2} \leq k^2$. The JC statistics count the vine pairs, including a newly-and a previously-diseased vine, according to the given definitions, if $c_{ilt} - c_{il(t-1)} = 1$, $c_{ilt} = 1$ and $c_{il(t-1)} = 0$.

$$JC^{omni}_{neighbour}(l,t,k) = \frac{1}{2}\sum_{i \leq j}(c_{ilt} - c_{il(t-1)})c_{jl(t-1)}\omega^{omni}_{lij}(k)$$

$$\omega^{omni}_{lij}(k) = \mathbf{1}_{\{(k-1)^2 < \frac{(x_{il}-x_{jl})^2}{(\Delta x_l)^2} + \frac{(y_{il}-y_{jl})^2}{(\Delta y_l)^2} \leq k^2\}}$$

The JC statistics for the row neighbor test is defined by changing $\omega_{lij}(k)$. The number of observed pairs was compared with the expected number of pairs of the same category under the null hypothesis : the newly-symptomatic vines were randomly distributed among the asymptomatic vines at the previous date. The significance test described below was also performed using permutation tests. In the latter case, the field simulations were generated by fixing the previously-symptomatic vines in year t (those of $c_{jl(t-1)=1}$), as well as the dead, missing and one-year-old plants, before each permutation.
Thus, to detect a general trend in the spatio-temporal data sets, we built several groups of global tests as suggested in Thébaud et al. (2005). These global tests synthesized the deviations of observed JC statistics from the means of JC statistics computed under H_0. For the groups of global tests, after testing the globally given hypothesis, each was performed with Bonferroni corrections (Bland and Altman, 1995).
For each vineyard, the aggregation patterns at different scales and different years could be analyzed using these individual tests in an exploratory fashion and several significance levels (0.05, 0.01 and 0.005) have been used to present more detailed results.

3.2.3.3 Global tests.

We performed groups of global tests using specifically defined statistics, which summed up the deviations of the observed JC statistics from means of permuted deviations computed on JC statistics under the null hypothesis.

Global Distance test per vineyard. To detect the spatial non-randomness in different vineyards, a global distance (GD) test was performed for each vineyard using

the : in the vineyard l, the esca cases are distributed randomly and globally in all of the years included. The definition of each test is as follows:

$$JC^{global}_{distance}(l) = \sum_{t=2004}^{2011} \sum_{r=1}^{15} |JC^{omni}_{distance}(l,t,r) - \overline{JC^{omni}_{distance\_sim}(l,t,r)}|W(l,t,r)$$

A weighted sum of differences between the observed JC statistics and the means of the simulated JC statistics for all of the distance classes, from 1 m to 15 m, and in each year from 2004 (or 2006) to 2011. We summed positive values to avoid the negative compensations that may influence the sensibility of the global test.

The weight is defined as the inverse of the averaged simulated JC statistics, $W(l,t,r) = \frac{1}{\overline{JC^{omni}_{distance\_sim}(l,t,r)}}$, to balance the JC statistics from different scales. The JC statistic calculated for larger distance classes or later years will count more in a non-weighted mean. Moreover, because the statistical test is meaningless for the distance class/year that has few pairs, we used 0 for the weights of the distance classyear in which the number of diseased pairs was less than 10.
With this construction of the global statistic, the global test was bilateral for H_1: the esca cases are non randomly distributed for all of the years and for all of the distances in the vineyard l. (The observed global statistic is significantly different from the statistic under the random distribution.)
To determine whether the non-randomness was at large scales or at small scales, and or along-row or not, for the vineyards showing significant non-randomness, we performed a group of tests containing four global tests, two at small scales, Global Distance Small (GDS) tests and two at large scales, Global Distance Large (GDL) tests by summing the differences computed using Omni-directional and Row Distance JC statistics, respectively, from 1m to 5m and from 5m to 15m respectively, in the same manner as $JC^{global}_{distance}(l)$.

Global Distance test per vineyard per year. To further explore the temporal evolution of the esca vines spatial distribution, for each vineyard showing significant non-random patterns, a global test was performed for each year with H_0: For vineyard l and year t, the esca vines are randomly distributed. The definition is as follows:

$$JC^{global}_{distance}(l,t) = \sum_{r=1}^{15} |JC^{omni}_{distance}(l,t,r) - \overline{JC^{omni}_{distance\_sim}(l,t,r)}|W(l,t,r)$$

Global Neighbor test per vineyard per order. To determine the neighbor order in which the newly diseased vines are distributed conditionally on the previously-diseased vines, groups of global tests containing 10 tests (omni-directional neighbor orders) with H_0: For vineyard l and neighbor order k, the esca vines are randomly distributed, were performed for each vineyard. The definition is as follows:

$$JC^{global}_{distance}(l,k) = \sum_{t=2004}^{2011} |JC^{omni}_{distance}(l,t,k) - \overline{JC^{omni}_{distance\_sim}(l,t,k)}|W(l,t,k)$$

The codes to perform these tests were written in R, and the package "Spatstat" was mainly used to generate spatial functions (R Core Team, 2013).

3.3 Results

Temporal progress of the disease. In 2004, the percentage of esca vine varied between 0.1 and 11.7 %, and the percentages of dead, missing or one-year-old plants was between 0.23 % and 13.85 % (Table 3.1). The temporal esca progression, shown in Figure 3.2 for the 15 vineyards, greatly differed depending on the vineyard, with disease prevalence varying between 1.34 % and 45.3 % at the end of the survey. For each vineyard, the disease's progress was mostly linear with slight variations depending on the year. Some of the vineyards showed linear disease rates in the first years of the survey followed by a slow down.

Global spatial analysis within vineyards. The GD test per vineyard, using the results of the $JC^{omni}_{distance}$, showed that 9 (2, 4, 7, 8, 9, 11, 12, 13 and 14) out of the 15 vineyards had significant P values (< 0.0033), after Bonferroni corrections, indicating that the esca distribution differed from a random pattern in these vineyards Table 3.2. When looking at individual vineyards with significant global tests, we focused on the GD test per year. The results showed that the year of transition from a random to a non-random distribution or, in some cases, conversely, varied according to the vineyard. Only four vineyards (7, 8, 12 and 13) showed a non-random disease distribution from the first or second year of recordings, indicated by significant values ($P < 0.00625$), after Bonferroni correction, for each year.

Spatial pattern at small and large scales. Table 3.3 presents the results of the distance tests, to distinguish the statistical dependence between esca vines at small (GDS tests) and large scales (GDL tests), oriented along the row, or not, for each vineyard. The four vineyards 7, 8, 12, 13 in addition to the vineyard 2 showed a significant P value ($P < 0.0125$) using the Omni-directional GDS and GDL tests, indicating a complex structural pattern. Within four vineyards (4, 11, 14 and 15), esca vines were significantly aggregated at the small scale because significant P values ($P < 0.0125$) were found only for the GDS tests. By contrast, the vineyards 1, 5 and 9, revealed a disease pattern only structured at the large scale, as shown by the Omni-directional GDL tests. The Row GDS tests showed similar results as the Omni-directional tests. Only vineyard 9 had a significant value ($P < 0.0125$) for the Row GDS test but not for the Omni-directional GDS test. This indicated the occurrence of symptomatic vine clusters solely oriented along rows in this vineyard. Eight vineyards (2, 3, 4, 6, 7, 8, 11 and 13) showed significant values ($P < 0.0125$) for the Row GDL tests.

Spatial pattern over time. Figure 3.3, 3.5, 3.6 present the evolution of the esca disease pattern over time within nine vineyards using the results of statistical tests for each distance class and each year. Only the vineyards with significant values ($P < 0.0125$) for the Omni-directional or Row GDS tests, except vineyard 2, with a very low prevalence, are shown. The vineyards 7, 8, 12 and 13 showed a great number of low P values from 2006, 2007 or 2008, revealing a tendency toward spatial dependency ($P < 0.005$ or $P < 0.01$) for distances between 1 and 10 m (Figure 3.3). This can be explained by clusters of esca vines at the small scale, distant from each other up to 10 m, as illustrated by the map of the vineyard 7 (Figure 3.5). Among

the vineyards, only 7 showed a great number of extremely low P values < 0.005 for the small distance tests and for the two first recording years. Figure 4 illustrates the results of tests for each distance class, and for 2007 and 2010, the number of pairs of symptomatic vines observed was always greater than those from the simulations. This indicated that the esca vines aggregated at different scales.

From 2004 or 2006 to 2008 or 2009, for the vineyards 7, 8, 12 and 13 the number of low P values, and also the maximal distance having a low P value, increased, but without an observed continuum over time (Figure 3.3). These results indicated no clear spatial extension of the clusters. Similar results were found with the row distance test (data not shown).

By contrast, the Omni-directional tests, for the other five vineyards (4, 9, 11, 14 and 15), showed that low P values were mostly obtained for the minimum distance class between esca vines (Figure 3.6, Figure 3.4).

This corresponded to the class between two adjacent vines located on the same row, indicating small clusters oriented along rows. Over time, there was no great increase in the number of low P values, except for vineyard 14. In this vineyard, the increase was associated with an increase in the maximum distance, suggesting an increase in the size of the esca vine cluster over the years. Row test results were consistent with Omni-directional test results. The P values were mainly obtained for minimum distances between two vines, corresponding to the adjacent vines on the row. For vineyards 9 and 14, in 2011 and 2010, respectively, an interaction was observed between two vines 2 or 3 m apart in the row, demonstrating the existence of small symptomatic vine clusters located along the rows that expanded very slowly over the years.

Location of new symptomatic vines in the close neighborhood. Five (7, 8, 12, 13 and 14) of the 15 vineyards showed significant P values with Bonferroni corrections (< 0.0033) for one or both Global Neighbor tests (Omni-directional and Row; Table 3.4). This result indicated, for these vineyards, the location of new symptomatic vines in a closed environment of previously-expressed symptomatic vines. When looking at individual vineyards and considering the Omni-directional Global Neighbor test, a significant effect of that closed environment was found, independent of the neighbor order, for the vineyards 7 and 13. For the other three vineyards (8, 12 and 14), in addition to vineyard 5, the number of significant P values varied between one and three. Globally, a slight effect of the neighbor order 1 was observed. Vineyards 7 and 13 showed significant P values for the Row Neighbor test without any advantage for a specific neighbor order in comparison with the Omni-directional Neighbor test.

3.4 Discussion

In this study, by conducting spatial point analyses from data records of contiguous vines of the same cv. Cabernet Sauvignon from 15 adult commercial vineyards in the Bordeaux region, a better understanding of esca spread capacity within a vineyard was obtained. The results indicated a large range of spatial patterns, from random to strongly aggregated structures, among vineyards. Various rates of disease increase were also observed over time. However, no clear relationship between these two factors was shown.

The secondary local spread of esca from symptomatic vines to their neighbors was investigated using Neighbor tests. For one third of the vineyards, the locations of newly-symptomatic vines were significantly related to those of previously-symptomatic ones, indicating that disease propagation from a symptomatic vine to its neighbors at the local scale occurred but was not systemic. The assertion of a weak potential for secondary local spread from neighboring symptomatic vines can be supported by two arguments. First, the vineyards showing a significant global effect at small distances frequently revealed a significant spatial relation between two vines at the minimum distance. This indicated the existence of small clusters of adjacent vines located along the rows. However, the analyses of multi-annual data revealed no enlargement of the small clusters over the years, except for one vineyard (14). Secondly, within vineyards with a stronger aggregated structure, the results from the analyses for each distance class and each year revealed various patterns without demonstrating a continuum of cluster size increases over time. The existence of small clusters without enlargement, in some cases, or the significant effect of the neighbor test, in other cases, may be more due to a local environmental effect than to a secondary disease spread. Thus, we concluded there is little to no secondary spread of esca.

The results of the Omni-directional Row Distance tests allowed us to generally conclude that the spatial structure of the disease lacks directional structure along the rows. Only one vineyard (9) showed significant aggregation structure at the small scale along the rows, without cluster enlargement. This vineyard is also characterized by the greatest distances between two adjacent rows among the 15 vineyards and consequently, a greater space between rows than between vines on the same row. This singular feature may result in the directional pattern along the rows. These results could be explained by three hypotheses: (i) local transmission through the short-distance spread of pathogenic fungi from one vine to another on the same row, because the distances between vines in a row was shorter than those between vines in two adjacent rows (see Table 3.1); (ii) fungal transmission via cutting tools; and (iii) similar locally conducive environmental conditions along the rows. The lack of increase in cluster size over time suggests the last hypothesis.

The combined results of our study contradict the suggestion of Zanzotto et al. (2013), who applied Bayesian spatio-temporal models to data from only one vineyard. Their results seemed to support a higher probability of infection along rows rather than among adjacent rows, suggesting an along-row spread of the disease, and they recommended measures to avoid contamination by pruning tools. Although our results do not confirm those of Zanzotto et al. (2013), the capacity to transmit pathogenic fungi through pruning shears, as shown by Agustí-Brisach et al. (2014), leads us to the same recommendation. This is in addition to all other preventive measures to decrease or eradicate all inoculum sources within the vineyards and their surroundings.

Further studies might focus on four of the surveyed vineyards (7, 8, 12 and 13), which showed, from the first years of the survey, highly significant P values for short distance classes, thereby confirming an aggregation structure. We could interpret the significance tests for long-distance classes as showing the relationships between two spatially distant diseased vine clusters. To explain the spatial structure at the small and large scales, the patterns could be related to the spatial heterogeneity of the

environment within these vineyards or the surroundings. Our first exploration of the spatial heterogeneity of soil suggested a role for this factor in explaining the spatial distribution of esca vines. Soil composition impacts the vine's physiological state and, consequently, its response to fungal infection.Surico et al. (2000b) observed heterogeneity in the esca distribution among vines, which they linked to the vineyard topography. They observed a lower number of diseased vines when the slope was steeper because this gave rise to a low soil water content.

The temporal progress of esca prevalence from 2004 could be empirically approximated by linear temporal curves for the majority of vineyards. These temporal results agreed with those of Marchi et al. (2006), who observed a constant increase in disease prevalence over their 10-year survey of mature vines. The conclusion of little or no secondary spread of esca corroborated the results of temporal progress, which can correspond to a monomolecular model. In that case, the rate parameter of the model is constant and linked to the amount of inoculum available for the infection of disease-free plants. The rate parameter varied slightly over the years, which was possibly related to the climatic variation influencing foliar symptom expression (Marchi et al., 2006, Surico et al., 2000b). In vineyards with the highest prevalence, a decreasing rate of disease progress over the last three years was observed, probably because fewer asymptomatic vines were available for foliar symptom expression. Because data recording only began when the vines were between 14 and 19 years of age, and the initial temporal disease pattern was unknown, we need to consider the relative prevalence values. Previously symptomatic plants could have been pruned, removed or even have died, in 2004, when recording started. A better estimation might have been obtained by considering the dead, missing and newly planted vines if vine mortality had only been caused by esca, but the causes of mortality were not available.

The random or weakly aggregated pattern in most of the vineyards in our study corroborated the results of most of the previous spatial analyses, even though they were conducted in different countries under different agronomical conditions. The random pattern may be due to the contribution of airborne pathogenic fungal spores, causing internal necrosis and associated with foliar symptomatic expression. Inocula originated from internal and external sources. Among the internal sources, epiphytic fungi present on the bark and endophytic fungi, such as P. chlamydospora, Phaeoacremonium aleophilum and Botryosphaeriaceae, already present in internal non-symptomatic woody tissues of the trunk or cordon (Bruez et al., 2013) can also play roles.

Despite the vines being the same cultivar, Cabernet Sauvignon, and of similar ages among the vineyards, the contrasting esca prevalences between vineyards, could be explained by multiple variables, such as micro-climatic factors, the level of primary infection, the external sources of inocula, and the clonal and rootstock genetics associated with viticultural practices. The choice of grapevine material, even for the same cultivar, such as the type of clone or the variety of rootstock, might influence vine vulnerability to GTDs. For instance, Murolo and Romanazzi (2014) found higher esca incidence on Fiano and Sauvignon when these were grafted onto rootstock SO4 rather than onto 1103P. They also found significant differences in esca foliar expression among Sauvignon clones. Additionally, both the training system and pruning practice have an influence on esca incidence (Lecomte et al.,

2011). For example, the positive impact of double pruning (or pre-pruning) on reducing grapevine trunk disease incidence in spur, pruned vineyards has been shown (Weber et al., 2007). In our survey, despite the same training system (Guyot) being used, variable winegrowing pruning practices, which were more or les conducive to esca foliar expression, might explain the differences in temporal esca dynamics in the vineyards. Additional research is needed to better understand the variability of disease spread among vineyards and vineyards showing a weak rate of esca increase should be analyzed. Studies on these vineyards should allow us to identify the environmental and agronomic practices that reduce the disease risk.

In conclusion, the hypothesis of spatial and temporal variations of esca distribution being attributable to different environments, as presented above, needs to be explored. Further studies, focused on identifying and evaluating the key environmental factors for esca can be performed using spatial and temporal modeling at the vine and vineyard scale.

3.5 Acknowledgements

The authors wish to acknowledge all the vine-growers who participated in this study, and Sylvie Bastien for her excellent technical assistance. They thank Pr. Xiangming Xu (East Malling Research, UK) for a critical review of a preliminary version of the manuscript. They are also grateful to Ray Godfrey for his kind assistance in improving the English of the manuscript. This study was funded by Bordeaux Sciences Agro, the Regional Council of Aquitaine, the JEAN POUPELAIN Foundation, the French Ministry of Agriculture and the Food-processing industry and Forest (CASDAR V907).

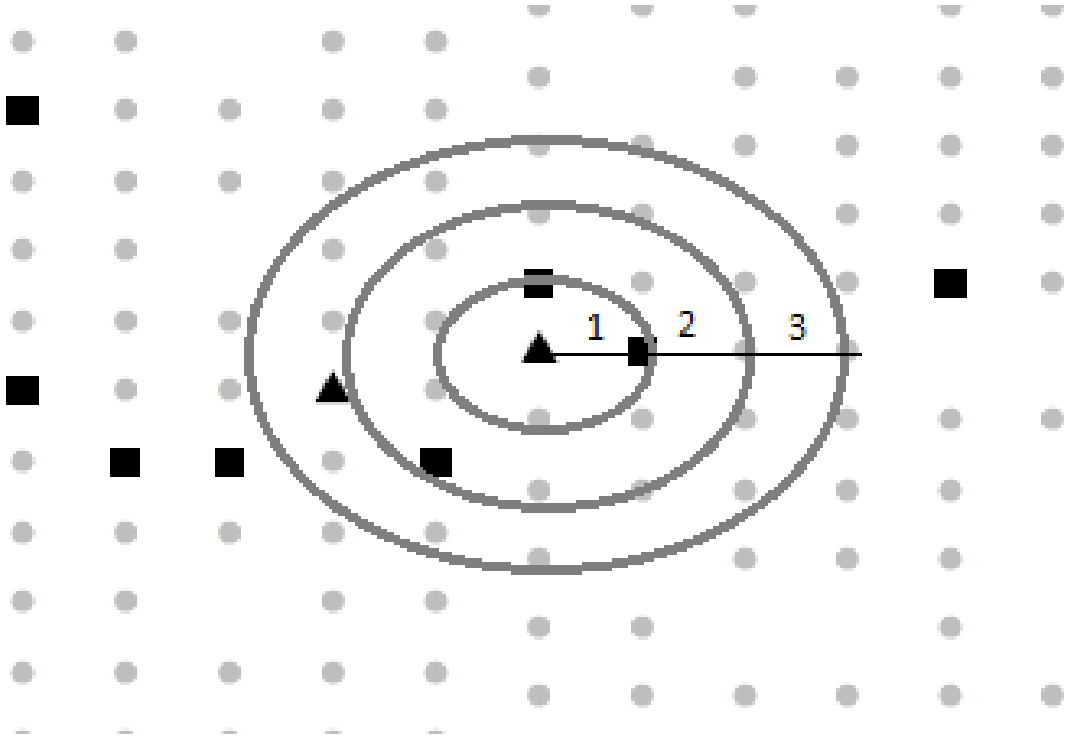


Figure 3.1: Spatial distribution of esca diseased vines (black) and healthy vines (grey), and the locations of the 1st-3rd neighbors, represented by the three circles, of one newly-diseased vine (black triangle).

Table 3.1: Location and characteristics of the 15 surveyed vineyards and their sanitary state in 2004

Vineyard number	Commune	Planting Year	Rootstock	Inter-vine	Inter-row	% esca vine in 2004	% dead, missing and young plants	Vines used for permutation	Total vine numbers per plot
1	Léognan	1988	101-14	1	1.2	0.10	2.80	1944	2000
2	Capitourlan	1989	101-14	1	1.8	0.23	0.23	1286	1289
3	St Emilion	1989	101-14	1.2	1.5	0.71	2.41	1981	2030
4	St Philippe d'Aiguille	1987	101-14	1.2	2.2	0.42	2.85	1433	1475
5	St Emilion	1989	101-14	1.2	1.5	1.00	1.48	2000	2030
6	Margaux	1987	101-14	1	1.5	2.69	12.37	1786	2038
7	Canéjan	1988	-	1	1.4	3.36	9.30	1814	2000
8	Martillac	1989	101-14	1	1.4	1.53	1.80	1964	2000
9	Beautiran	1990	-	1	1.8	2.97	5.37	1919	2028
10	St Philippe d'Aiguille	1985	101-14	1.2	2.2	4.63	2.38	1965	2013
11	Espiet	1989	SO4	1.2	3.5	4.81	13.85	1766	2050
12	Castres	1989	3309	1.2	1.4	10.88	6.93	2123	2281
13	Margaux	1987	3309	1.2	1.2	7.57*	10.10	1834	2040
14	Gradignan	1989	3309 C	0.8	1.5	9.21	2.06	1998	2040
15	Galgon	1987	3309 and Paulsen	1.2	2.9	11.70	2.90	1942	2000

* % esca vine in 2006

Table 3.2: P values of Omni-directional Global Distance tests for each vineyard and P values of Omni-directional Global Distance test per year. Significant P values are marked in bold, with the significance level adjusted by a Bonferroni correction performed within each group (P value = 0.05/15 = 0.0033 for the global tests per vineyard and P value = 0.05/8, except for vineyards 1 and 2, P value = 0.05/5, and vineyards 4 and 13, P value = 0.05/6). F: few diseased pairs (not tested), ND: No recorded data

Vineyards	Global test	2004	2005	2006	2007	2008	2009	2010	2011
1	0.004	F	F	F	**0.003**	**0.004**	0.03497	0.68931	0.7043
2	**0.001***	F	F	F	0.04096	0.01598	**0.005**	**0.002**	**0.001**
3	0.21678	-	-	-	-	-	-	-	-
4	**0.001**	F	F	0.3047	0.02697	0.01499	**0.005**	**0.001**	**0.002**
5	0.04496	-	-	-	-	-	-	-	-
6	0.04695	-	-	-	-	-	-	-	-
7	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
8	**0.001**	**0.002**	**0.004**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
9	**0.002**	0.38362	0.18781	0.3007	0.43257	0.00899	0.005	0.001	0.003 -
10	0.26873	-	-	-	-	-	-	-	-
11	**0.001**	0.74226	**0.001**	0.004	**0.001**	0.01598	0.01499	0.12787	0.1978
12	**0.001**	0.14885	0.02597	$<$ **0.001**	0.002	$<$ **0.001**	$<$ **0.001**	**0.001**	**0.001**
13	**0.001**	ND	ND	**0.002**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**
14	**0.001**	0.44456	0.10989	0.01199	0.01698	0.005	0.004	0.01099	0.005
15	0.005	-	-	-	-	-	-	-	-

* Significant P value

Table 3.3: P values of the global tests for each vineyard summing statistics for all classes of omni-directional distances from 1 to 5 m (small distance classes) or all classes of omni-directional distance from 5 m to 15 m (large distance classes). Bonferroni corrections performed on four tests for each vineyard; the significance level was reduced to $0.05/4 = 0.0125$.
F: few diseased pairs (not tested), ND: No recorded data

	$1m \leq r \leq 5m$	$r > 5m$	$1m \leq r \leq 5m$	$r > 5m$
	Omni	Omni	Row	Row
1	F	**0.004***	F	F
2	**0.00599**	**0.001**	F	**0.00699**
3	0.02198	0.71528	0.03796	**0.001**
4	**0.001**	0.42757	**0.001**	**0.001**
5	0.77822	**0.004**	0.01698	0.2987
6	0.30669	0.01898	0. 4046	**0.002**
7	**0.001**	**0.001**	**0.001**	**0.001**
8	**0.001**	**0.001**	**0.001**	**0.001**
9	0.02997	**0.00599**	**0.001**	0.08991
10	0.16983	0.48551	0.25475	0.0969
11	**0.001**	0.38462	**0.001**	**0.001**
12	**0.001**	**0.001**	**0.001**	0.13686
13	**0.001**	**0.001**	**0.001**	**0.001**
14	**0.001**	0.59341	**0.001**	0.89011
15	**0.005**	0.1958	0.18581	0.41259

* Significant P value

Table 3.4: P values for each Omni-directional or Row Neighbor test (Global Neighbor test and tests for each neighbor order from 1 to 5, combining eight years (2004–2011) of data. Bold numbers: significant P values. The significance level of P values was adjusted using a Bonferroni correction with the corresponding groups: P value = 0.05/15 = 0.0033 for Omni-directional and Row Global Neighbor tests per vineyard, and P value = 0.05/5 = 0.01 for Omni-directional and Row Global Neighbor tests per vineyard per neighbor order.

			Neighbor order									
			1		2		3		4		5	
Vineyards	Omni.	Row	Omni.	Row	Omni.	Row	Omni.	Row	Omni.	Row	Omni.	Row
1	0.37762	0.05395	-	-	-	-	-	-	-	-	-	-
2	0.20979	0.23177	-	-	-	-	-	-	-	-	-	-
3	0.39161	0.77722	-	-	-	-	-	-	-	-	-	-
4	0.14985	0.08092	-	-	-	-	-	-	-	-	-	-
5	0.04196	0.02198	0.04795	0.09091	0.62937	0.61938	0.55145	0.52547	0.93007	0.005	0.002	0.14286
6	0.46553	0.62737										
7	0.001*	0.001	0.001	0.004	0.001	0.1029	0.001	0.001	0.001	0.001	0.001	0.96503
8	0.001	0.01998	0.04196	0.22378	0.22178	0.2028	0.03197	0.10989	0.18681	0.27972	0.003	0.01998
9	0.03097	0.25475	-	-	-	-	-			-		-
10	0.96703	0.81818	-	-	-	-	-			-		-
11	0.64735	0.68332	-	-	-	-	-			-		-
12	0.001	0.002	0.001	0.16583	0.61039	0.08891	0.08891			0.17582		0.08991
13	0.001	0.003	0.004	0.002	0.00999	0.001	0.001			0.58042		0.06893
14	0.001	0.002	0.001	0.02498	0.1978	0.13886	0.13886			0.18082		0.87413
15	0.37662	0.57243	-				-		-			

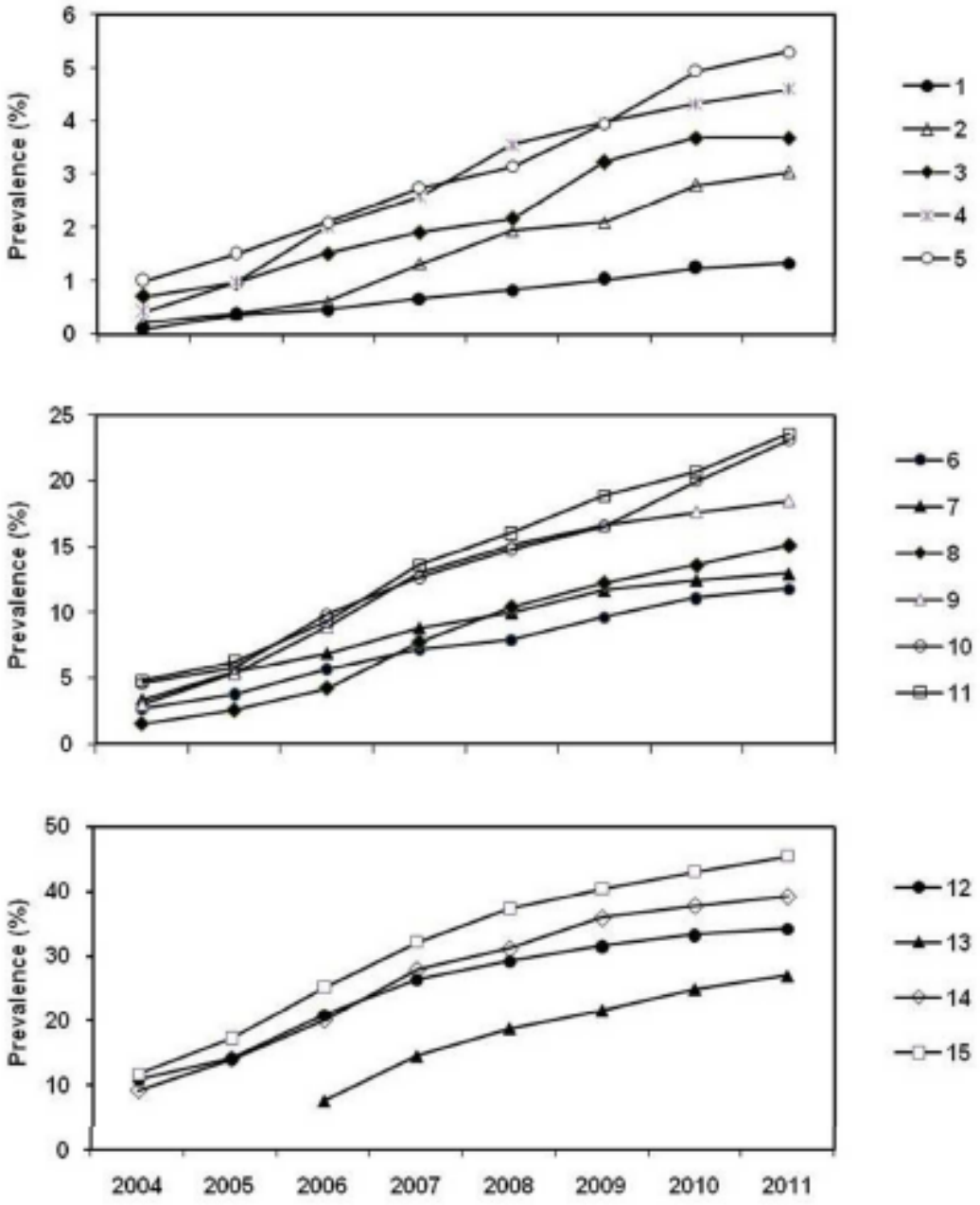


Figure 3.2: Yearly prevalence of grapevine esca between 2004 and 2011 for the vineyards (1 to 15), except for Vineyard 13 (from 2006 to 2011).

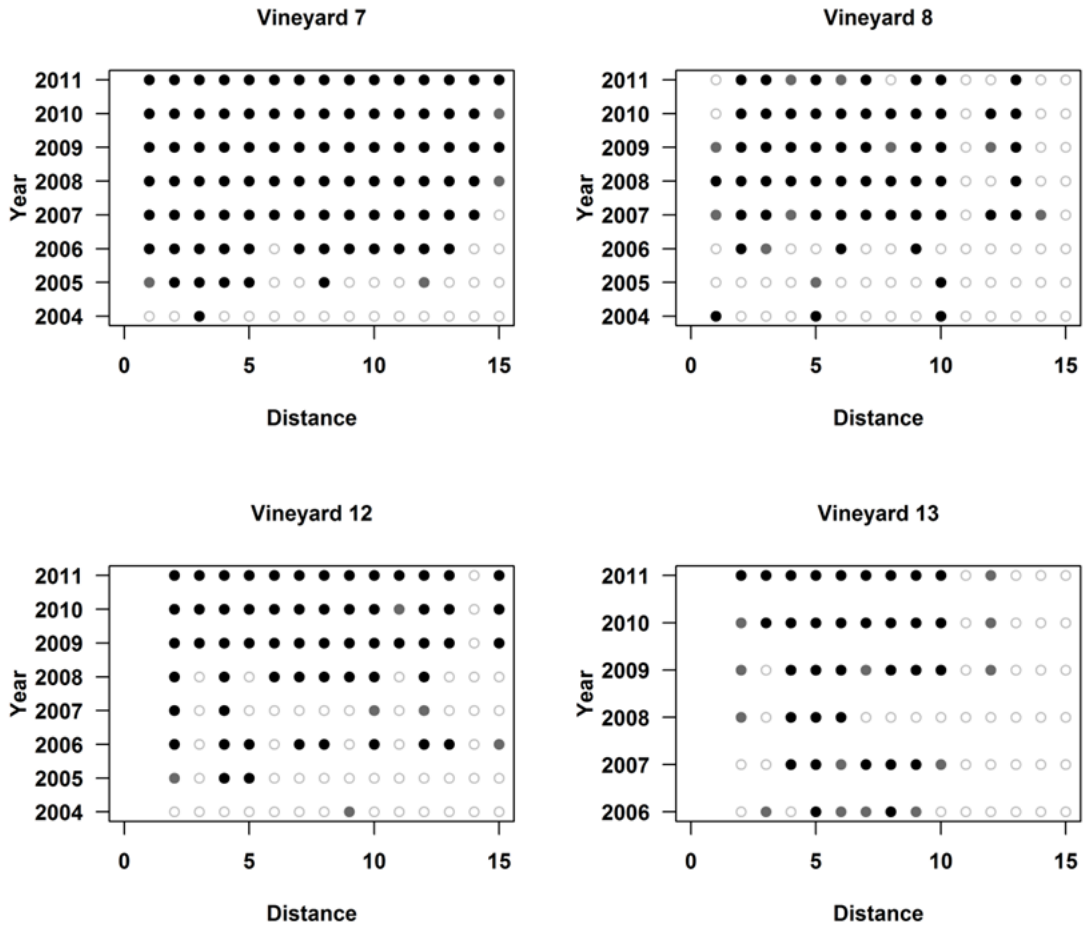


Figure 3.3: Significant P values of Omni-directional Distance tests from 1m to 15 m for the vineyards 7, 8, and 12 over eight years (2004 to 2011) and for vineyard 13 over six years (2006 to 2011). Black points indicate P values less than 0.005; grey points, less than 0.01; and white points, greater than 0.01, corresponding to strong, less strong and no rejection of their null hypotheses, respectively.

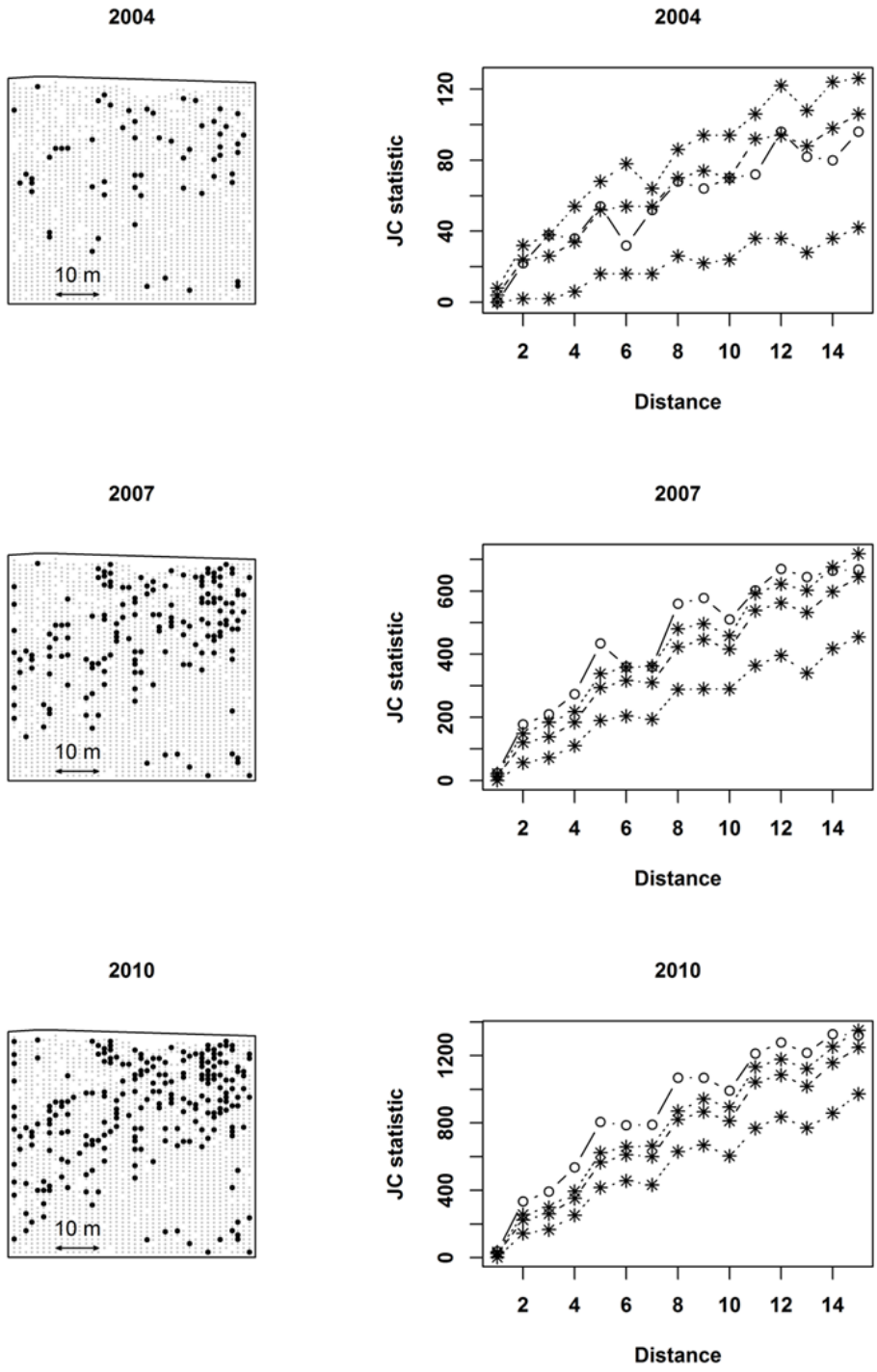


Figure 3.4: Distribution of esca-expressed vines and the JC statistics fit for to vineyard 7 at for years 2004, 2007 and, 2009. The black curve with circles corresponds to the observed JC statistics at each distance. The dashed line with stars corresponds to the maximum, 95th quantile (dotted dash line), the minimum value of JC statistics calculated by 1,000 simulations under the null hypothesis. If the circle is above the star of the 95th quantile, the test is significant at significance level 0.05 without adjustment.

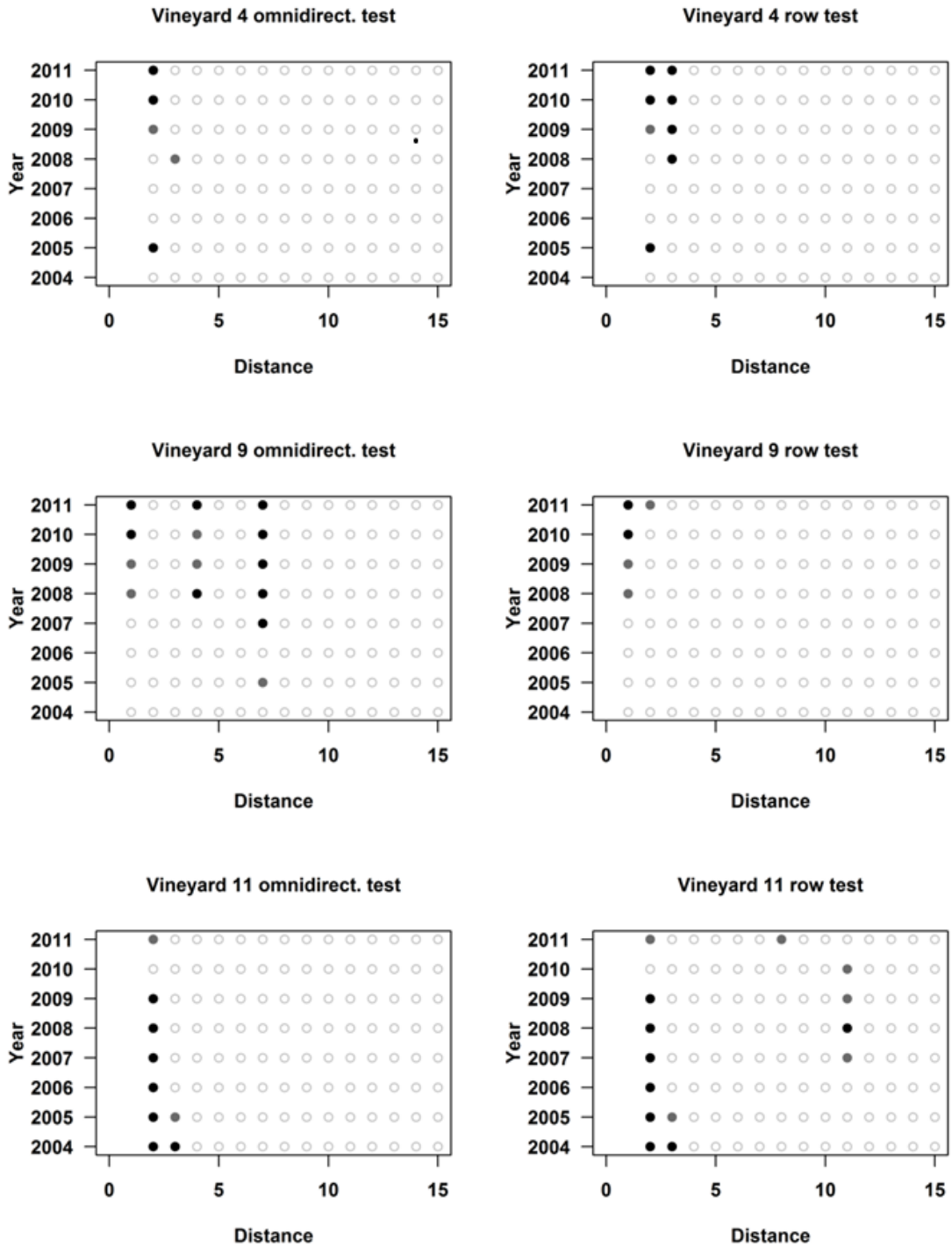


Figure 3.5: P values of Omni-directional (left side) and Row Distance tests (right side) for vineyards 4, 9 and 11over eight years (2004 to 2011). Black points indicate P values less than 0.005; grey points, less than 0.01; and white points, greater than 0.01, corresponding to strong, less strong and no rejection of their null hypotheses, respectively.

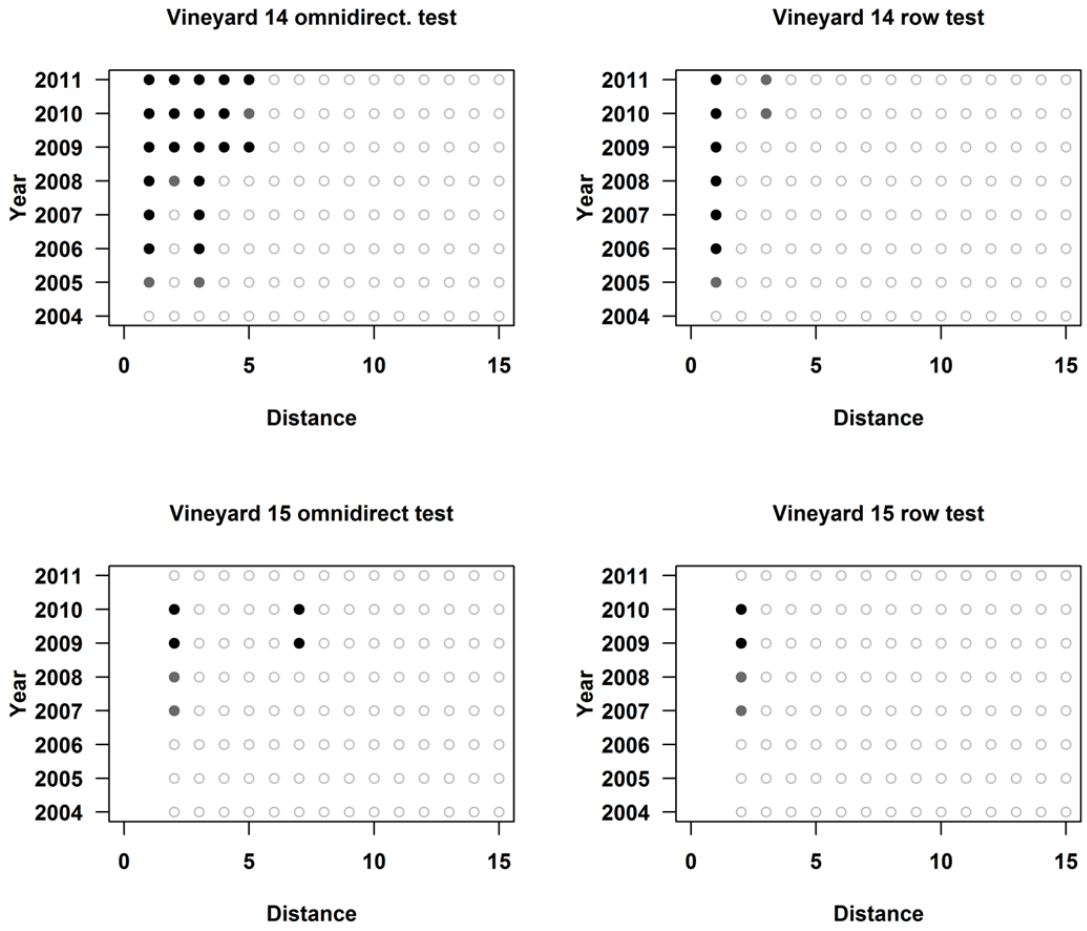


Figure 3.6: P values of Omni-directional (left side) and Row distance tests (right side) for vineyards 14 and 15 over eight years (2004 to 2011). Black points indicate P values less than 0.005; grey points, less than 0.01; and white points, greater than 0.01, corresponding to strong, less strong and no rejection of their null hypotheses, respectively.

Chapter 4

Méthodes Géostatistiques pour l'estimation des covariables à l'échelle du cep

Introduction

Dans ce chapitre, nous présentons les méthodes géostatistiques pour l'estimation des covariables à l'échelle du cep. Le chapitre comprend deux parties distinctes. Dans la première, notre objectif est de proposer des méthodes de statistique spatiale pour obtenir des estimations spatiales d'une variable de l'état hydrique du sol : la réserve utile du sol. Ceci pour chacun des ceps, et à l'échelle de la parcelle. Pour ce faire, nous utilisons les enregistrements des données de résistivité électrique comme variable auxiliaire.

Combinée avec les données de climat, la réserve utile peut être utilisée comme paramètre d'entrée pour le modèle Lebon (Lebon et al., 2003). La sortie de ce modèle est une variable spatio-temporelle qui mesure l'état de contrainte hydrique de la plante suivant sa position et les données climatiques. Cette variable peut être utilisée dans le modèle spatio-temporel pour prédire et/ou expliquer la variation spatio-temporelle de la maladie.

Pour obtenir l'information spatiale sur la réserve utile au cep, nous aurions besoin d'un très grand nombre de mesures de réserve utile, qui sont impossibles à recueillir compte-tenu du coût prohibitif et surtout du caractère invasif des mesures. Pour pallier cet inconvénient, nous utilisons les enregistrements très denses de la résistivité apparente, qui est connue pour être liée à la réserve utile et que nous utilisons comme covariable externe pour la prédiction spatiale de la variable d'intérêt, lorsque cette variable est faiblement échantillonnée (Bourennane and King, 2003, Bourennane et al., 2012, Wackernagel, 2003). Afin de prédire spatialement la réserve utile en utilisant les données de résistivité, une méthode statistique a été développée pour prendre en compte les spécificités suivantes : (1) une aire géographique structurée en rangs de ceps, (2) un échantillon de données de résistivité structuré suivant la même géométrie : des mesures en lignes espacées de plusieurs mètres, (3) un échantillon de réserve utile très épars du fait du caractère intrusif des mesures, qui ne sont pas adaptées à la plantation pérenne.

Nous avons combiné plusieurs méthodes statistiques ou géostatistiques standard

afin de traiter la forme particulière de nos données : variogramme local, krigeage median polish, krigeage avec dérive externe et modèles de régressions classiques. L'objectif principal de cet article est de trouver un bon modèle pour reproduire la corrélation entre les données de réserve utile et de résistivité électrique mesurées sur une parcelle de la région Bourgogne en France. Pour atteindre cet objectif, nous considérons deux sous-objectifs : (1) caractériser la structure spatiale locale des données de résistivité et les interpoler aux points de mesure de la réserve utile, (2) paramétrer la relation entre la réserve utile et la résistivité afin d'utiliser cette relation pour des prédictions.

Dans une seconde partie, nous décrivons la prédiction de deux indicateurs liés aux plantes : l'azote total et le $\delta^{13}C$ mesurés à partir du jus de raisin pour quelques dizaines de ceps. Le premier nous donne des informations sur l'état azoté de la plante, qui est relié à la vigueur de la plante et le deuxième nous donne une information sur l'état hydrique de la plante. Les prédictions de ces indicateurs à l'échelle du cep sont effectuées par krigeage ordinaire. Les prédictions obtenues sont incluses dans les modèles statistiques présentés dans le chapitre 5.

La première partie de ce chapitre correspond à une version préliminaire d'un article en collaboration avec Pierre Curmi[1] et mes co-encadrantes. Elle a bénéficié de la lecture bienveillante de Benjamin Bois que nous remercions. Nous voudrions la soumettre prochainement à Spatial Statistics.

Résumé

La réserve utile est un paramètre du sol dont les mesures sont invasives et couteuses. Le nombre de mesures de réserve utile est donc limité. Au contraire, la résistivité électrique peut être facilement et densément mesurée pour décrire la variabilité de la structure et des propriétés du sol. De plus la résistivité et la réserve utile sont corrélées indépendamment du temps. C'est pourquoi dans cette étude, nous étudions une méthode pour raffiner les estimations de la réserve utile à partir de peu de sondages de réserve utile, tout en utilisant les données de résistivité électrique . En revanche, le sondage des données de résistivité est très spécifique puisqu'il est effectué le long des rangs de vignes induisant une structure spatiale complexe. Pour mener à bien notre objectif, plusieurs méthodes statistiques sont utilisées: un variogramme et un "krigeage one-way median polish" ont été utilisés pour caractériser la structure spatiale localement en ligne et pour réaliser le krigeage local pour la résistivité autour des sondages de réserve utile. La corrélation entre les estimations de réserve utile et de résistivité est ajustée par un krigeage à dérive externe et des modèles de régression. Les résultats montrent d'une part que le "krigeage one way median-polish" est efficace pour traiter la structure en ligne avec un variogramme cyclique et d'autre part qu'il y a une relation linéaire entre le logarithme de la réserve utile et la résistivité. De plus le modèle GLM avec résidus de loi log-gamma possède les performances suffisantes pour la validation croisée, même avec peu de mesures de réserve utile, ce qui permet des prédictions plus précises de cette variable hydrique. Ceci confirme le fait que l'idée d'échantillonner la réserve utile à partir

[1]Agrosup,UMR 1347 Agroécologie, Dijon, France

de la distribution de la résistivité est efficace pour récupérer la corrélation entre les variables tout en éliminant la structure spatiale.

Spatial prediction of available water capacity using geostatistical Models and soil resistivity data

Shuxian Li[a,b] , **Pierre Curmi**[c] , **Lucia Guerin-Dubrana**[a,b] and **Anne Gégout-Petit**[d]

[a]Université de Bordeaux, ISVV, UMR-1065 INRA
[b]Bordeaux Sciences Agro, Gradignan, France
[c] Agrosup, UMR 1347 Agroécologie, Dijon, France
[d]Institut Elie Cartan, Université de Lorraine, INRIA BIGS Nancy Grand-Est, Nancy, France

Abstract

Available water capacity (AWC) is a perennial soil parameter of which the measurements are invasive and time consuming. Electrical resistivity (ER) data, which can be easily and densely measured to describe the variability of soil structure and properties, has a time-invariant correlation with AWC. However, ER measurements are distributed along the rows of vine plantation and the number of AWC measurements is limited. In this study, with few AWC surveys, an effective approach is investigated to refine the estimation of AWC using electrical resistivity. Several statistical methods are combined: variogram and one-way median-polish kriging have been used to characterize the local spatial line-structure and perform the local kriging for resistivity data around AWC surveys; the correlation between between AWC and ER estimations are fit by kriging with external drift and regression models. The results showed that on one hand, one-way median-polish kriging is proved effective to deal with the line-structure with a cyclic variogram, and on the other hand, there is a linear relationship between logarithm of AWC and ER and moreover, even with few AWC measurements, the GLM model with log-gamma distributed residuals present the adequate performances from cross validation, which ensures the further prediction of water route. This confirm the fact that the sampling plan of AWC according to ER distribution is effective to recuperate the correlation between variables but we lost spatial structure in this way.

Additional keywords: Geostatistical model; kriging; spatial regression

4.1 Introduction

Spatialized soil water availability for plants at field scale is a challenge issue for agriculture because it helps to reach the water regime of the plant, using modeling. The hydric state of the plant is a limiting component of crop production. A water stress may result in lower yields (Kang et al., 2009) or induce plant susceptibility to pests (Garrett et al., 2006). Knowledge of the spatial variability of the water content of the soil provides information, for example, to manage irrigation strict requirement of the plant, to delineate areas of use of variety or rootstock most suitable for the availability soil water, to adapt plants protection methods.

The electrical resistivity geophysical technology has showed its interest in the study of the physical properties of the soil , to assess the soil volume wetness and the transpirable soil water at different spatial scales Celano et al. (2011), Michot

et al. (2003), Brillante et al. (2014). This non intrusive technology is very useful in the case of soil exploration in the 2D or 3D. Using a specific equipment, it has been developed to record spatially continuous measurements for the large scale soil prospection in a short time and at a high spatial resolution (Samouëlian et al., 2005).

In the present study, our objective was to compare spatial statistical methods to spatially estimate a soil water variable: the available capacity of the soil for a vine at vineyard scale and high resolution using the recording continuous data of electrical resistivity, as an ancillary data.

The available water capacity of a soil (or Soil Water Holding Capacity : SWHC) has been defined as the difference between the maximum amount of soil water, excluding free water that a soil is able to store in the root zone and the volumetric soil water content at the permanent wilting point (the amount below which water is so strongly retained that plants are unable to absorb it). These variables may be estimated with pedotransfer function, using texture, bulk density of clods (Bruand et al., 2004). Taking into account the thickness of each soil profile horizon and plant characteristics, the available soil water content may be estimated, in field condition at each sampling point. To perform spatial information about this variable, we need a great number of sampling leading to prohibitive costs as well as invasive and time consuming measurements. The spatial continuous records of the apparent resistivity, used as an external drift can improve the spatial prediction of the variable of interest when this variable is sparsely sampled (Bourennane and King, 2003, Bourennane et al., 2012, Wackernagel, 2003). Moreover, electrical resistivity could be a useful tool to spatially estimate the time-invariant variable, such as the available water content at field scale using the data of soil texture (Hesse et al., 1986, Tabbagh et al., 2000). Recently, Buvat et al. (2014) showed that a specific apparent electrical resistivity profile is consistent with soil profile description, such as the presence of clay layer and the depth of soil.

In order to spatially predict the AWC using resistivity data, statistical methodology was developed to account for the following specificity of our context that are (1) a geographical area structured by vines planted in rows (2) a sample of resistivity data structured by this geometry: measurements on line spaced by several meters (3) a very sparse sample of AWC because of the intrusive measure not adapted to perennial plantation

Several standard statistical or geostatistical methods can be combined to adapt our specific design of data. Statistical regression models offer the direct ways to quantify the correlation between AWC and resistivity data while geo-statistical regression methods, with considering the spatial structure of data, are more often used to perform the spatial predictions in the presence of exhausted ancillary information (Bourennane et al., 2012, Hengl et al., 2004, Rivoirard, 2002). Many studies in environment performed regressions in a spatial context. Some of them used statistical tools that not account for the spatial structure: Celano et al. (2011) used ordinary regression models with link function and found a exponential link between soil resistivity and soil water content. To fit their data, Brillante et al. (2014) chose a non parametric model using splines in order to build a pedotransfer function between electrical resistivity and soil volume wetness. They noticed that the model had to be calibrate for each homogeneous layer of the area. Various geostatistical methods have been developed: McKinley et al. (2013) used Geographically Weighted Regres-

sion (GWR) to link the incidence of cancer with traces elements in soil. Rossi et al. (2013) used generalized linear model (GLM)to account for the spatial correlations of the residues. They also study the effect of tillage on electrical resistivity trough a spatial random field with a mean structure given by the two levels of the treatment.

With one ancillary variable used for interpolation, several methods are available: (1) kriging with external drift (KED); (2) regression Kriging (RK); and (3) collocated cokriging (CLCK). According to Rivoirard (2002), CLCK is theoretically the best linear estimator, accounting for coregionalization models. However, CLCK is generally not achievable when the ancillary variable is densely sampled. KED, as a particular formulation of universal kriging, should be preferred when the target variable is driven by the ancillary variable, moreover, it is more adapted than RK when the trend form is known but the parameters are unknown. Moreover, Regression kriging and Kriging with external drift are formally different initially, but they are, if not identical, very close mathematically speaking and lead to the same or almost the same results (Hengl et al., 2003). To account for the sparsity of the measured AWC, not adapted to spatial model, we also compare the spatial predictions using KED and classical statistical regression models, and discuss the possible sampling effect to the models prediction performances.

To perform the KED as well as linear regression models, the ancillary data need to be known at locations of AWC determination. As, it is not the case, the first step of the study concerns this prediction. Bourennane et al. (2012) used ordinary kriging (OK) to interpolate the resistivity data at AWC determination locations, based on the estimated variogram which described the whole-area spatial structure. However, due to the complex soil spatial structure within this vineyard plot, resistivity data has a high variability and has different spatial structures between sub-areas. Walter et al. (2001) pointed out that spatial trends and non-stationarity of the process may affect the accuracy of classical geostatistical approaches and showed that in such a case, the kriging methods with local variogram performed better than with whole-area variogram for spatial mappings.

In this work, we analyzed the complex spatial structure of resistivity data of the vineyard by local analysis. For the local area having a periodic variogram, we developed a one way median polish kriging method in order to erase this periodicity.

More precisely, the main objective of this paper is to find out a good model to fit the correlation between AWC and electrical resistivity data measured on one vineyard from the Burgundy region in France. To fulfill this, two sub objectives are: (1) characterize the local spatial structure of resistivity data and interpolate them at point of available water capacity measurement (2) parametrize the relationship between available water capacity and resistivity and use it to prediction.

The paper is organized as follows: Section 2 gives a description of the collected data. Section 3 presents the mathematical framework and methodologies used to build and validate the model. Section 4 motivates our choices and explain how the different methods are combined together. It is followed by a discussion.

4.2 Materials and methods

4.2.1 Data collection

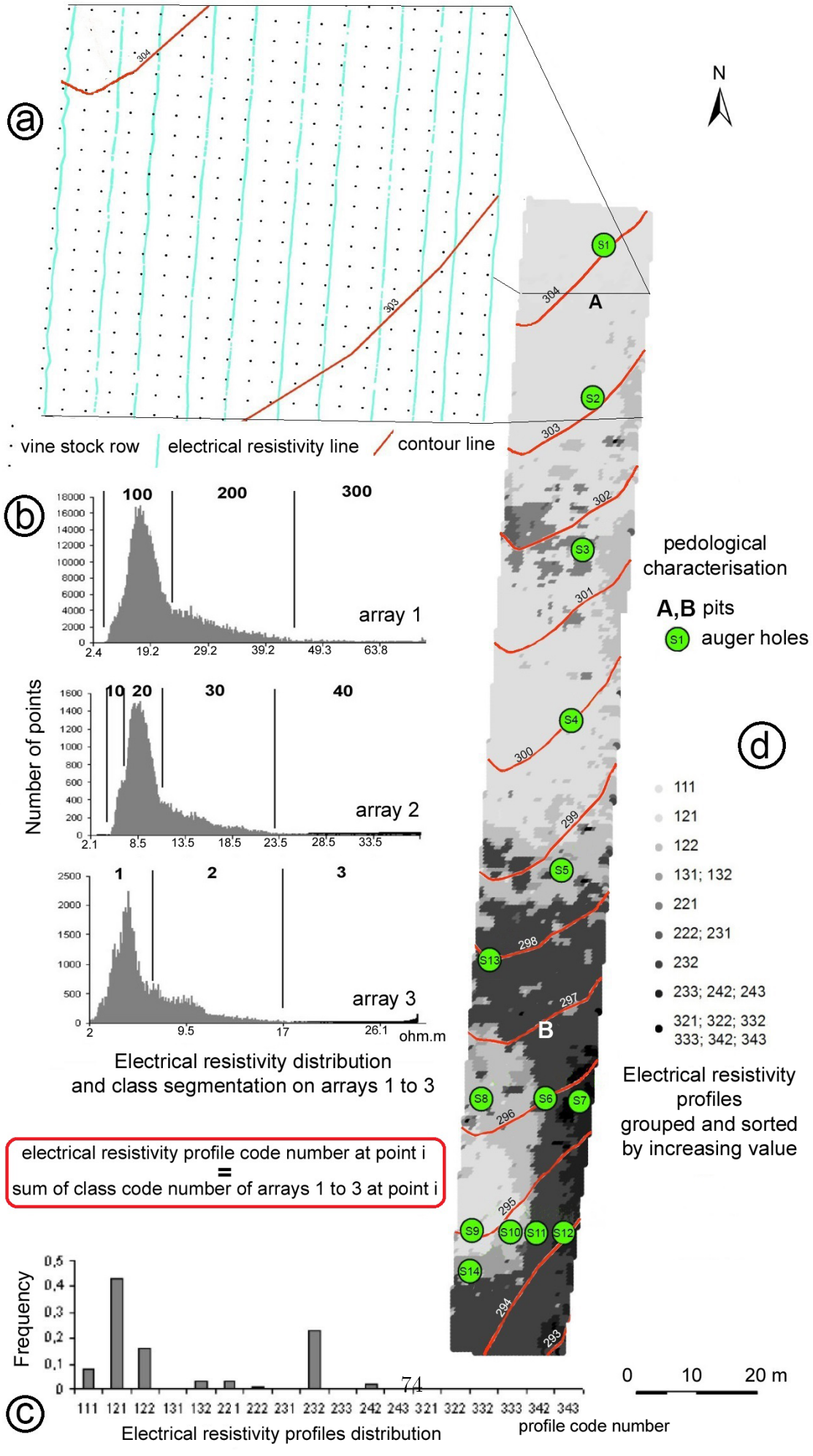
The study area was a 4000 m^2 vineyard plot (20 vine stock rows 1 m apart between rows and within rows) located at Saint Vallerin (Burgundy, France) (Figure 4.1). Two main soils types according Baize and Girard (2008) were found on the plot: thick heavy clay redoxic CALCISOL upslope and shallower stony clay CALCISOL further down the slope of 1 %. The electrical resistivity survey was done using Multi-depth Continuous Electrical Profiling device (MuCEP) (Samouëlian et al., 2005). This device allowed measuring electrical resistivity for three depth of investigations: 0-0.5 m (array 1), 0-1 m (array 2) and 0-1.7 m (array 3) (Figure 4.1b). The measurements were designed in lines along the plantation of vine rows (Figure 4.1a). They were densely distributed within lines (every inch) and lines are 2 m spaced. Thus more than 67000 measurements were done over a 4000 m^2 area on 11 lines oriented in direction of the slope.

AWC was calculated using the pedotransfer class method proposed by Bruand et al. (2002) applied on soil profile description and soils analysis of two pits, one on each type of soil. It was also estimated at 14 other locations, by means of auger vertical sample collections. The sampling plan of AWC was done according to the map of resistivity, along perpendicular transects of the variation direction (Figure 4.1): one group of sampling collected from upstream to downstream for slow transitions (s1-5, s13) and the other group (s6-12, s14) were collected on the same level curve with rapid changes. These last sampling was more closely distributed than the first group.

The resistivity data were measured at three depths, gave three sets of data strongly correlated. Consequently, we need to extract the best covariate for the regression model. For this Bourennane et al. (2012) used a geostatistical filtering for resistivity measurements at three depths to extract the principal component which keeps the large-scale tendency and removes the small variation scale. In our case, the studied area is much smaller and the sample of AWC is not regular so that the small-scale variation can not be ignored. In our case, we choose to use the resistivity data measured at the second lane for the spatial prediction of AWC motivated by two reasons: (1) in an exploratory purpose, we performed an ordinary kriging for each of the three variables to get the estimated values at target locations and compared their correlations with AWC and resistivity data measured at the second lane (1 m) showed stronger correlation, (2) a simple principal component analysis have been applied to these three variables (Figure 4.2) and the resistivity measured at the second lane was the most correlated variable with the first principal component.

4.2.2 Mathematical framework

Supposing that we have target variable Z_1 and the ancillary variable Z_2 valued in a spatial field $\mathcal{D}$, their spatial observations can be represented by $\mathbb{Z}_1 = (Z_1(s_1), ..., Z_1(s_m))'$, $\mathbb{Z}_2 = (Z_2(s_{m+1}), ..., Z_2(s_{m+n}))'$, where $\{s_1, s_2, ..., s_{m+n}\} \in \mathcal{D}$ correspond to their observed locations respectively, and $m << n$. In our study, $m = 14$ and $n = 67000$.


a
N
304
303
S1
S2
A
vine stock row
electrical resistivity line
contour line
b
100
200
300
array 1
10
20
30
40
array 2
1
2
3
array 3
Number of points
ohm.m
Electrical resistivity distribution
and class segmentation on arrays 1 to 3
electrical resistivity profile code number at point i
=
sum of class code number of arrays 1 to 3 at point i
c
Frequency
111 121 122 131 132 221 222 231 232 233 242 243 321 322 332 333 342 343
Electrical resistivity profiles distribution
profile code number
d
pedological
characterisation
A,B pits
S1 auger holes
302
S3
301
S4
300
299
S5
298
S13
297
B
S8
S6
S7
296
295
S9
S10
S11
S12
S14
294
293
111
121
122
131; 132
221
222; 231
232
233; 242; 243
321; 322; 332
333; 342; 343
Electrical resistivity
profiles
grouped and sorted
by increasing value
0
10
20 m

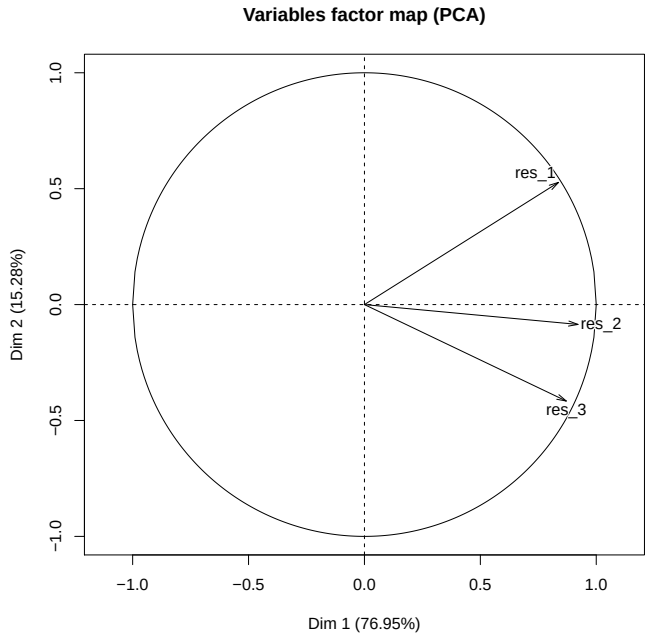


Figure 4.2: Principle component analysis of resistivity data at three arrays.

They can be considered as partial samplings of the realization of random processes $Z_1(s), Z_2(s), s \in \mathcal{D}$. Our aim is to propose a prediction for Z_1 at every points of $\mathcal{D}$.

We present hereafter the two kriging models used at different step of our methodology developed in section 4.3.

4.2.3 Kriging models

4.2.3.1 Ordinary kriging

Ordiexcuter la krigeage ordianirenary kriging (OK) is the most widely used kriging method (Wackernagel, 2003). The principle of OK is to predict variable Z for an unsampled location s_0 using n neighborhood sampling points and combined them linearly with weights ω_i, $Z(s)$ is assumed as an intrinsic random function (that the expectation of the increments are zero and the variance of the increment depends only on lag), and $Z^*(s_0)$ represents the predicted value at s_0:

$$Z^*(s_0) = \sum_{i=1}^{n} \omega_i Z(s_i) \tag{4.1}$$

To obtain an unbiased estimator, the weights have to be constrained to sum up to one: $\sum_{i=1}^{n} \omega_i = 1$. The weights are calculated using an estimated variogram by minimizing the estimation variance under the constraint. Thus we have a following objective function to minimize:

$$\phi_{OK} = \sigma(Z(s_0))^2 - \mu_{OK}(\sum_{i=1}^{n} \omega_i - 1)$$

where μ_{OK} is the Lagrange parameter for the constraints of the weights. In more detail, the weights are calculated by OK system:

$$\begin{cases} \sum_{j=1}^{n} \omega_i = 1. \\ \sum_{j=1}^{n} \omega_j \gamma(s_i - s_j) + \mu_{OK} = \gamma(s_i - s_0) \quad \text{for } i = 1, ..., n \end{cases}$$

where $\gamma(s_i - s_j)$ is the vector of calculated variogram values between interpolated points s_i and s_j (Wackernagel, 2003).

4.2.3.2 Kriging with external drift

Kriging with external drift (KED) is a popular geostatistical method (Goovaerts, 1997, Wackernagel, 2003). It allows the ancillary information to be used to account for the spatial variation of the target variable $Z_1(s)$'s local mean. The ancillary variable $f(s)$ should be known at desired points, and is considered as deterministic (Rivoirard, 2002). Assuming the linear relation between the variables, we have

$$\begin{aligned} E[Z_1^*(s_0)] &= \sum_{i=1}^{n} \omega_i E[Z_1(s_i)] = a + b \sum_{i=1}^{n} \omega_i f(s_i). \\ E[Z_1^*(s_0)] &= E(Z_1(s_0)) = a + b f(s_0) \end{aligned} \tag{4.2}$$

This implies that the weights should be consistent with an exact interpolation of s_0

$$f(s_0) = \sum_{i=1}^{n} \omega_i f(s_i). \tag{4.3}$$

The objective function ϕ to be minimized in this kriging system consists of the estimation variance σ_E^2 and two constraints.

$$\phi = \sigma_E^2 - \mu_0 (\sum_{i=1}^{n} \omega_i - 1) - \mu_1 (\sum_{i=1}^{n} \omega_i f(s_i) - f(s_0))$$

where μ_0, μ_1 are the Lagrange parameters of the constraints. Thus the supplementary universality condition (the regression), concerning the external drift variable measured exhaustively in the spatial domain is incorporated into the kriging system (Goovaerts, 1997).

In this study, the deterministic ancillary variable $f(s)$ at required locations can be obtained by estimations and observations of random variable Z_2 with a link function g: $f(s) = g^{-1}(Z_2(s))$. The link function allows us to model more flexible relationships that are linear between functions of the variables and we suppose here that $E[Z_1(s) - a - bf(s)|Z_2(s)] = 0$.

4.3 Theory and Calculation

The prediction procedure have been performed in two steps applying different statistical methods specific of the spatial statistics or not.

4.3.1 Predict ancillary variable on the whole area

We interpolate ancillary variable Z_2 at target variable sampled locations $\{s_1, s_2, ..., s_m\}$ using observations $(Z_2(s_{m+1}), ..., Z_2(s_{m+n}))$. Unlike other interpolation models, kriging models have the advantage that they integrate representation of the average spatial variability by estimating the variogram, and give us a best linear unbiased estimator (Wackernagel, 2003). However, our data do not have a global stationarity to ensure a good behavior of the kriging on whole-area directly. To account for the grid structure of the data before kriging, a local analysis (interpolation) is used to improve the assessment of Z_2 due to the heterogeneous spatial structures which often leads to the non-stationarity of increments (Walter et al., 2001). Furthermore, the number of predicting locations m is sufficiently small (m=14 here), local kriging with local estimated variogram for each location is applicable. Thus, for every $s_i, i = 1, ..., m$, the local area chosen for the analysis covers the surrounding points within 5-meters. $D_i = \{s_j, |s_i - s_j| < 5\}$.

4.3.1.1 Denoising

The purpose of this step is to detect the possible outliers using h-scatter plot. H-scatter plot is a bivariate plot of all the possible pairs $(Z(s_i), Z(s_j))$, such that $|s_i - s_j| \in H$ and $H = (a, b]$ with $0 < a < b$. Thin cloud on the h-scatter plot indicates strong relation between the values at the separation, while a fat cloud implicates a weak relations. By removing the data presenting the outlying values on h-scatter plot (the data that produced the points outside the cloud), we can reduce the effect of outliers and improve the robustness estimation of the variogram. We visualized h-scatter plots of Z_2 at each local area, and outliers possibly caused by measure errors were removed to improve the variogram estimation.

4.3.2 Median Polish and kriging

The purpose of this step is to account for the grid structure of the data before kriging. To describe the irregular gridded spatial data (two-way array) in which the grid spacings do not have to be equal in either the horizontal direction or the vertical direction, Cressie (1993) speaked of a mean structure obtained by additive decomposition of the row and column effect:

$$E(Z_2(s_i)) = u(s_i) = a + r_k + c_l, \quad s_i = (x_l, y_k) \tag{4.4}$$

where a is a global effect and r_k and c_l are respectively the row and column mean effects. In order to avoid bias and the influence of the extreme values, a specific approach called Median Polish, has been proposed to estimate the additive effects given above using Median theory (Cressie, 1993). Median Polish proceeds by repeated extraction of the row and column medians until convergence, with respect to a stopping criterion chosen by the investigator. The basic idea of the median polish algorithm is as follows:

- Compute the median of each row and record the value to the side of the row. Subtract the row median from each observation in that particular row.

- Compute the median of the row medians, and record the value as the grand effect. Subtract this grand effect from each of the row medians, and record the values as the row effect.
- Compute the median of each column and record the value beneath the column. Subtract the column median from each observation in that particular column.
- Compute the median of the column medians, and add the value to the current grand effect. Subtract this grand effect from each of the column medians, and record the values as the column effect.
- Repeat steps 1-4 and add the new grand effect, row and column effects to the current ones at each iteration until no changes occur with the row medians.

It gives estimators of a, r_k, c_l, which we write as $\tilde{a}, \tilde{r_k}, \tilde{c_l}$, so that the original spatial data can be expressed as :

$$Z_2(s_i) = \tilde{a} + \tilde{r_k} + \tilde{c_l} + R(s_i) = \tilde{u}(s_i) + R(s_i) \tag{4.5}$$

In our area structured by the vines arranged by lines and resistivity measurements driven by these lines, resistivity measure lines at both side of the survey exhibit a strong heterogeneity while the values of resistivity data in the same line varied smoothly. According to this, a one-way median polish method was adopted to deal with this spatial structure. We chose the two lines beside the survey to study and suppose that these two-line resistivity data is composed by a stationary component and a column effect for each line which caused the horizontal heterogeneity.

With this modeling, the observations of Z_2 can be decomposed as:

$$E(Z_2(s_i)) = a + c_l, \qquad s_i = (x_l, \text{whatever } y) \in \mathcal{D}. \tag{4.6}$$

Consequently, the rows effects no longer need to be estimated. A simplified one-way Median Polish algorithm which subtracts only the column effects and global effects for each iteration applied here to give us the effect estimators $\hat{a}, \hat{c_l}$ for the $s_i = (x_i, y_i), i = 1, .., m$ located between the columns l and $l+1$ leading to the estimation of Z_2 at s_i:

$$Z_2(s_i) = \hat{a} + \hat{c_l} + R(s_i) = \hat{u}(s_i) + R(s_i), \qquad s_i = (x_l, \text{whatever } y) \in \mathcal{D}. \tag{4.7}$$

where $\hat{R}(s_i)$ is the Median-Polish residual.

We interpolate the trend $\hat{u}$ and the residuals $\hat{R}$ separately:

- The trend at $s = (x, y)$ is estimated with a simple linear interpolation. If x is the coordinate corresponding to the rows of measurements and that $x_l < x < x_{l+1}$,
$$\hat{u}(s) \equiv \hat{a} + \hat{c_l} + (\frac{x - x_l}{x_{l+1} - x_l})(\hat{c}_{l+1} - \hat{c_l})$$
- A variogram analysis and an ordinary kriging (see Section 4.2.3) are performed on the residuals $\hat{R}(s_i)$, leading to the following residuals:
$$\hat{R}(s) = \sum_{i=1}^{n} \lambda_i R(s_i) \tag{4.8}$$

4.3.3 Spatial estimates of target variable by ancillary variable

Finally, as we have obtained the estimated values of Z_2 at all the required locations, it only remains to estimate Z_1. There are mainly two ways of estimating the target value. One is to consider only the relationship between target and ancillary variable $g(Z_1), Z_2$ (where g is a link function) and to model it using standard statistical models. The other way is to apply universal kriging with external drift (Section 4.2.3). Here we performed a single linear model (SLR) between $log(Z_1(s_i))$ and $Z_2(s_i)$ as exhibited in 4.9. We also performed generalized linear models (GLM) (McCullagh, 1984, Nelder and Baker, 1972) with a log link of Gaussian family (4.10).

$$log(Z_1(s_i)) = aZ_2(s_i) + b + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \tag{4.9}$$

$$Z_1(s_i) = g^{-1}(aZ_2(s_i) + b) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \tag{4.10}$$

4.3.4 Validation criteria

The performance of interpolation methods can be evaluated using interpolation and validation sets. However, in our case, the number of the target variable observations m is too small to be divided into interpolation and validation sets. So we applied leave-one-out cross validation (LOOCV) of which we re estimate each of the m points using the other $(m-1)$ other points, without re estimate the variogram model for kriging models (Lejeune, 2006).

Let $\hat{z_1}(s_j)$ represents the predicted value, $z_1^*(s_j)$ the observed value and m the number of the validation points. Then for LOOCV, the true prediction accuracy can be evaluated by comparing estimated values with actual observations at validation points in order to assess systematic error, calculated as mean prediction error (MPE):

$$MPE = \frac{1}{m}\sum_{j=1}^{m}[\hat{z_1}(s_j) - z_1^*(s_j)] \tag{4.11}$$

and accuracy of prediction, calculated as root mean square prediction error (RMSPE):

$$RMSPE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}[\hat{z_1}(s_j) - z_1^*(s_j)]^2} \tag{4.12}$$

In order to compare accuracy of prediction between variables of different types, the RMSPE can be normalized by the total variation s_z, $\bar{z_1}^*$ denotes the mean of $\{Z_1(s_j), j = 1, ..., m\}$:

$$RMSPE_r = \frac{RMSPE}{s_z}, \qquad s_z = \sqrt{\frac{\sum_j^m (z_1^*(s_j) - \bar{z_1}^*)^2}{m-1}} \tag{4.13}$$

As suggested by (Hengl et al., 2004), we consider that a value of $RMSPE_r$ close to 40% means a fairly satisfactory accuracy of prediction . Otherwise, if the values get $> 71\%$, this means that the model accounted for less than 50% of variability at the validation points and the prediction is unsatisfactory.

The three methods presented above are proposed by (Hengl et al., 2004), we also calculate the correlation between the actual observations and estimated values of LOOCV, and take it as a criteria of goodness and fit.

4.4 Results and discussion

As shown in the map of Figure 4.1, the resistivity profiles were distributed into two parts : in the North part of the map, the three values of resistivity were in the lower classes of the distribution except for a little area in the West with median values of the three resistivity data. The South part was more contrasted : there was a circular area in the West with low values of resistivity but it was surrounded by a great zone where the three measured values of resistivity are in the highest classes. That explained the higher number of AWC measurements in this zone.

As we presented before, knowing the existence of physical correlation between the AWC and resistivity, we managed to parametrize this correlation and use it to predict AWC values. Thus the sampling plan AWC were especially (intentionally) measured according to the distribution of resistivity. Such sampling enables to comprehend the relation as much as possible with limited measurements and is considered to be more effective than the uniform sampling for the prediction because it needs less AWC samples.

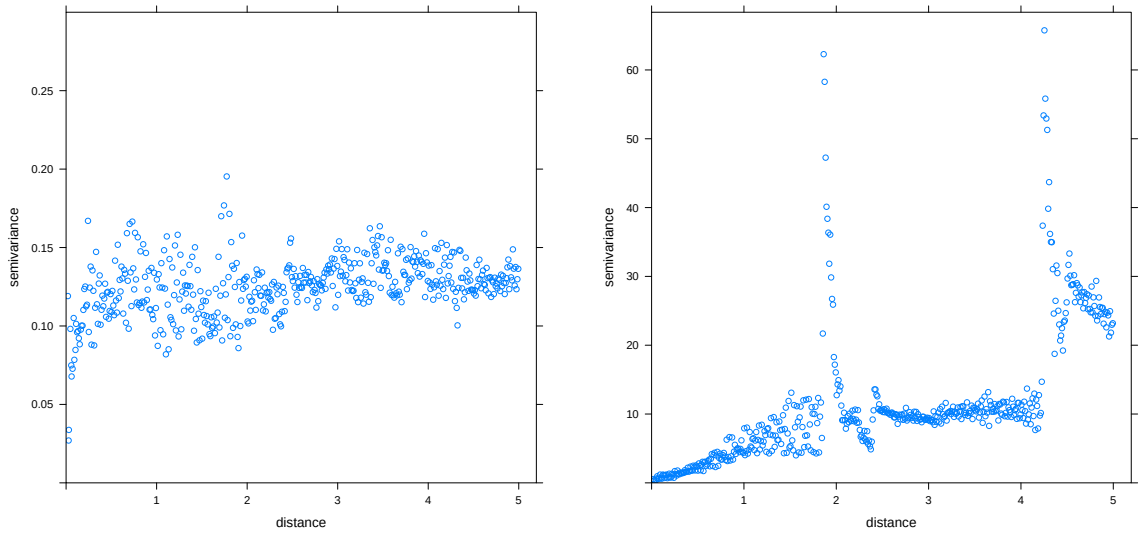
4.4.1 Statistical analysis

4.4.1.1 Interpolation of the ancillary variable Z_2

After denoising steps performed by examining h-scatter plots, the $m = 14$ local variograms of Z_2 were plotted for each target survey. Among them, 6 out of 14 surveys (5, 6, 7, 10, 11, 12) showed irregular variograms as the one in Figure 4.3b. There were two periodic peaks located on variograms each 2 meters. This irregularity indicated that the local analysis and h-scatter plot that partly reduce the spatial non-stationarity, were not sufficient for more complicated spatial structures. Moreover, the scales of variograms varied a lot between different surveys (Figure 4.3a, Figure 4.3b) motivating for local analysis.

The underlying cause of this special periodic variogram and the large difference between variogram scales could be found on local resistivity maps. Figure 4.4a and Figure 4.4b showed local resistivity maps of S1 and S6 respectively. The resistivity observations around S1 varied a little in a same scale (from 5 to 10) while the resistivity observations around S6 showed apparently two scales: higher resistivity (from 10 to 40) at right side and lower resistivity at left side (less than 10). The resistivity measure lines at both side of the survey exhibited a strong heterogeneity while the values of resistivity data in the same line varied smoothly. A one-way median polish method combined with ordinary kriging was applied as explained in section 4.3.2. We chose the two lines beside the survey to study and we supposed that these two-line resistivity data are composed by a stationary component and a column effect for each line which caused the horizontal heterogeneity.

Figure 4.4c and Figure 4.4d showed the variogram of the S6's local area before and after removing the column effect by median polish with irregularities removed.



(a) variogram of resisvitity data removing the outliers around s1

(b) Local variogram of resistivity data removing the outliers around s6

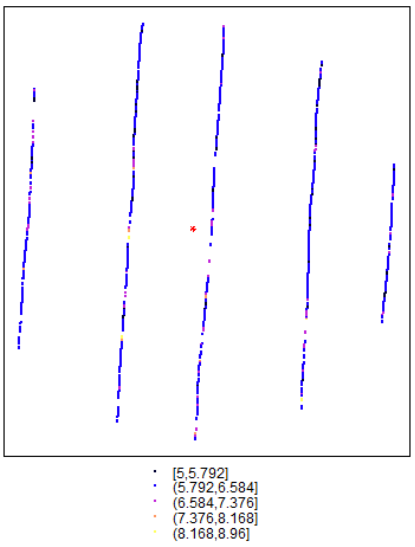
Figure 4.3: Local spatial analysis

4.4.1.2 Discussion for interpolation step

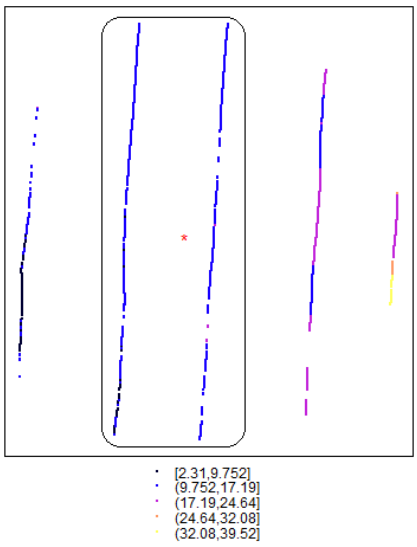
In this interpolation step, two types of local variograms of resistivity data observed around each AWC sampling are found. One type was spherical variograms from the upper area which is homogeneous; and the other one was the variograms exhibiting cyclic behaviors, the cycles corresponds to the distances between the geophysical lines.

It has to be noted that the periodic variogram does not indicate a periodic spatial structure. Actually, it corresponds to a zonal anisotropic structure: a spherical variogram structure for horizontal direction combined with a vertical 2 m spaced samplings highly uncorrelated (with high discrepancy)(see Figure 4.5). This variogram indicate a lateral soil variation at the same metric scale and in a perpendicular direction of prospection. Such local spatial structure was difficult to be characterized by classical anisotropic variogram: since the samples were spaced two meters, in a local region of 5m rayon, we could only get the empirical variogram at two lags, 2 m and 4 m.

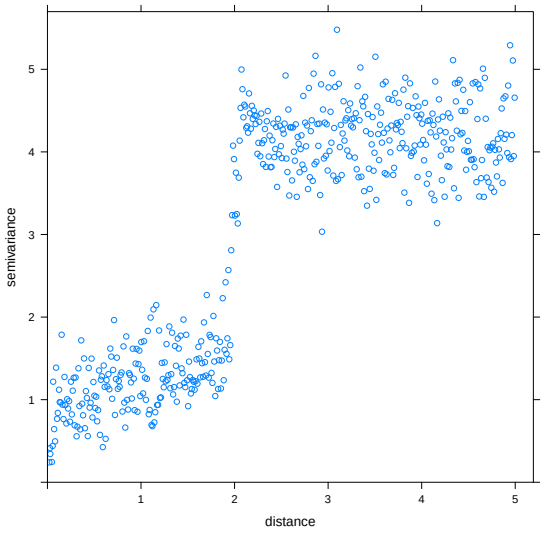
This anisotropic structure included all the descriptive terms in Zimmerman (1993) to define the zonal anisotropy: range, sill and nugget anisotropy. Zimmerman (1993) also highlighted that unequal sills may be evidence of non stationary, non vanishing spatial correlation, or measurement errors that are correlated or have unequal means. They further pointed out that if the experimental variograms in different directions have rather different sills, then it would not be prudent to plunge ahead with the modeling of the variogram, until some operations have been done such as removing the trend.



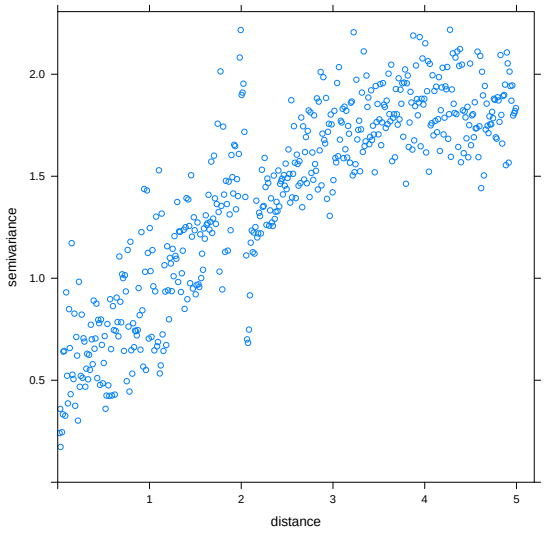
(a) resistivity map of S1's 5m neighbour



(b) resistivity map of S6's 5m neighbour



(c) variogram of resistivity data around s6 before removing the column effects



(d) variogram of resisvity data around s6 after removing the column effects

Figure 4.4: Median polish applied to periodic variogram.

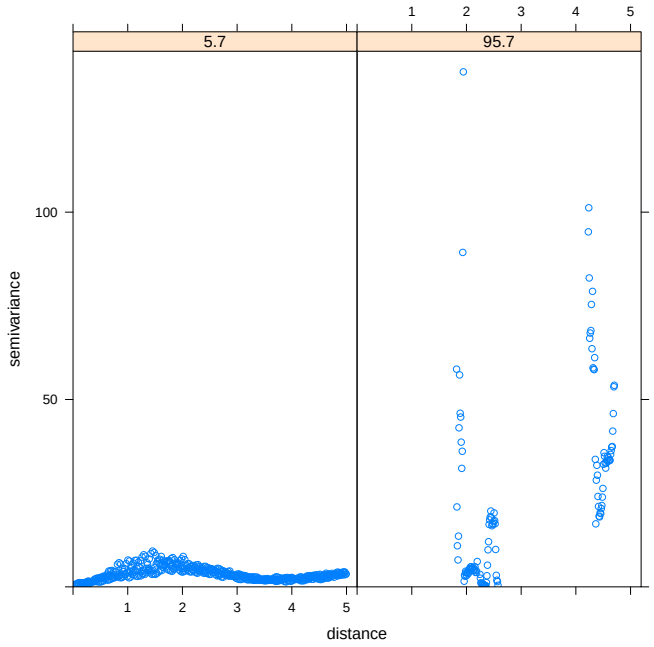


Figure 4.5: Anistropic variogram of resistivity data around S6. Left: direction N5.7, along the line, right 95.7.

	SLR	GLM(log.Gaussian)	KED
MPE	-0.636	0.53	-0.76
$RMSPE$	61.9	34.8	66.8
$RMSPE_r$	71%	39.8%	76.6%
correlation	0.74	0.82	0.63

Table 4.1: Criteria values for three models.

4.4.2 Spatial estimates of target variable by ancillary variable

After the resistivity data were estimated at each AWC determination locations $s_i, i = 1, .., m$, a scatter plot (Figure 4.6) have been plot to check their relation. The available water capacity values decrease while the resistivity data get higher. And the slope is steeper at the beginning than the end. We inferred a log link from this kind of plot by experience. The scatter plot between logarithm of AWC and the estimated resistivity shows an apparent linear relation, which further proves the link and legitimates the choice of the different models given by equations(4.9), (4.10). We also performed kriging with external drift.

The parameters of the regression are significant and negative in all the models and the results of validation are shown in Table 4.1: GLM model with log Gaussian distributed residuals fit best the data and both SLR and GLM (log.Gaussian) models have better performances than KED.
We suspect that linear regression models are more adapted than geostatistical model possibly because the target variable are too sparsely distributed to identify its spatial structure.

Figure 4.7 showed the estimated AWC values using GLM gaussian model for the

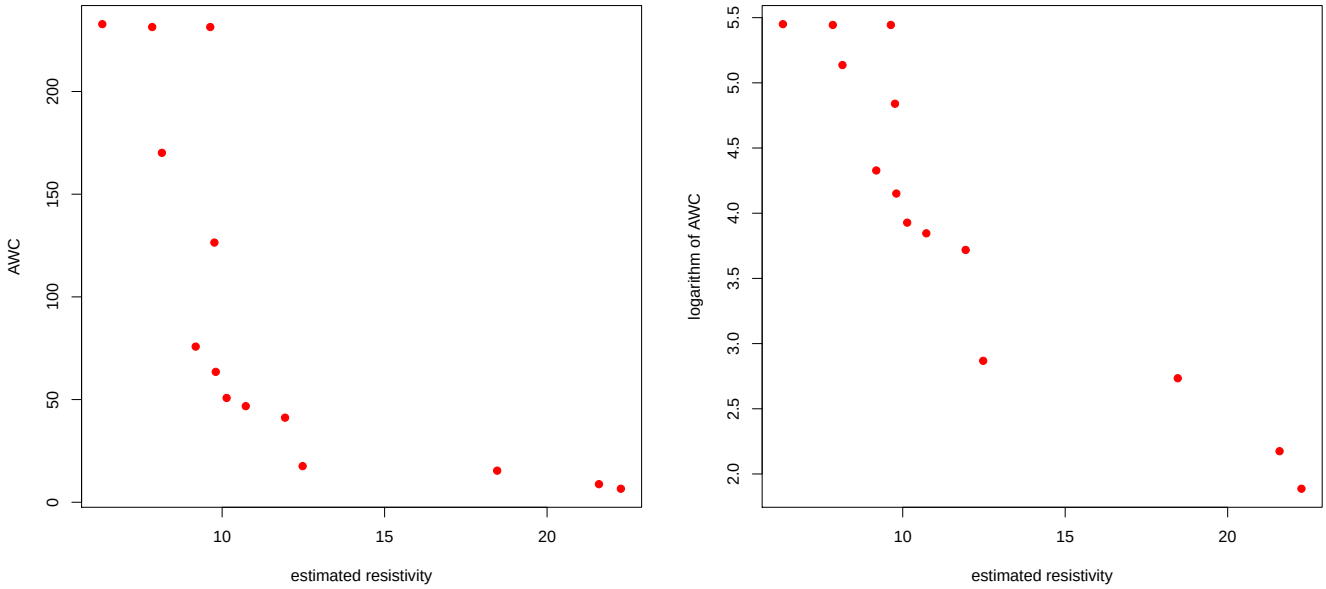


Figure 4.6: Plots of AWC functions (y, mm) and estimated resistivity(x, Ohm.m). Left AWC, right log(AWC).

last interpolation. Compared with the map of the resistivity profiles (Figure 4.1), we observed a negative correlation between resistivity and AWC already shown in plots of Figure 4.6.

4.5 Conclusion

In this paper, with few available AWC surveys, and very specific sampling of the ancillary variable oriented by the plantation in rows, we offered an effective approach to refine the estimation of available water capacity (AWC), using resistivity data, across the studied vineyard. The approach is a combination of several spatial statistical methods: h-scatter plot, local variogram analysis and median polish kriging methods have been used to characterize the local spatial structure and to perform the local kriging for resistivity data around the AWC surveys. Kriging with external drift, simple regression and generalized linear model have been used to fit the correlation between AWC and resistivity data with or without considering the spatial structure of AWC data. The results showed that there is a linear relation between $log(AWC)$ and resistivity, generalized linear model with log link fit the data best. Even with extremely few number of AWC observations, GLM model presented suitable performance from cross validation, which ensures the further prediction. Let us recall that the values of resistivity variable used in the regression models are not the observed values directly measured and are estimated from the kriging methods, this may cause a propagation of uncertainties. However the present study demonstrated that both statistical and geostatistical techniques can be combined together and used effectively to analyze AWC with resistivity data in an special situation because we need to get the estimations using the number of AWC surveys as few as possible.

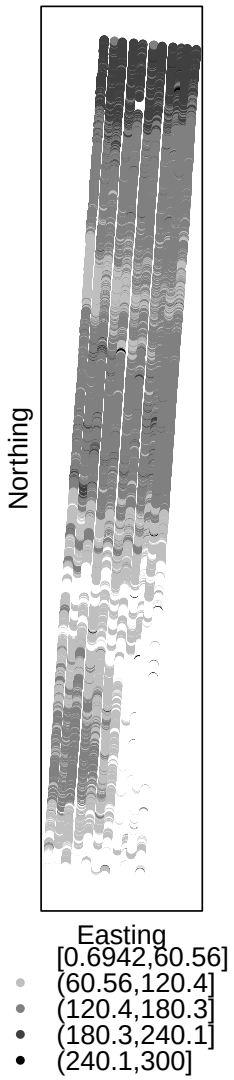


Figure 4.7: AWC values (mm) on the studied field, estimated by our method with GLM gaussian model for the next step of interpolation.

The one way median polish kriging method described in this article offered a direct way to subtract the unequal means of data in different directions. More than that, range and nugget anisotropy have, at the same time, been solved once applied the median polish algorithm. It is worked out especially well for the local line structure which commonly exists in agriculture domain as a typical large-data sampling approach. Here we declared that in practice, the soil spatial structure can be much more complex than theoretical assumptions, the application of median polish methods in this paper gave a good example to show that the integration of pedology information and statistical techniques can help us to find a quick and simple way to solve complicated problems. Moreover, this method can be adopted in the case of measurement of electric resistivity in already planted orchards or vineyards leading to a gap between two lines of resistivity.

Under the sampling plan of AWC drawn according to the resistivity map, it is not surprising that the statistical regression models have better performances than the geo-statistical regression model (KED)(Table 4.1). 14 AWC punctual determinations are far less than enough to characterize a spatial structure, as Cressie (1993) suggest at least 30 pairs of points for each distance lag to get a proper variogram.

The reason that we use two types of statistical regression models (SLR and GLM) to fit the data is that we did not want to give determined assumptions for the regression variances, since only 14 AWC data are available, little can tell from exploratory analysis. The GLM model assumed normally distributed homoscedastic residuals while the SLR model supposed that the the residuals of $log(AWC)$ (logarithm of AWC data) are normally distributed. We left the data to decide their own structures. Finally the results revealed that the GLM model is slightly better than SLR model, the assumption that the AWC data is composed with a spatial tendency and a Gaussian error is supported.

Our strategies succeed to get the estimations using limited AWC surveys. It's very important as for commercial agriculture field, to reduce the cost of expensive soil surveys and to avoid the destructive behaviors are very necessary.

The suitable performance of the prediction of AWC at a fine scale provided the opportunity to use it as a spatial covariate in spatio-temporal models at the scale of the vines stock. We will have to account for that the model used estimated data through kriging methods, which may cause a propagation of uncertainties. Predicted AWC data at vine scale give us the possibility to assess the vine water status using water balance modeling (Lebon et al., 2003). Spatial modeling of water regime of vine on the scale of a vineyard is a subject of current research by including data of ER and also plant data (i.e. surface of the canopy, trunk diameter, ratio measurement isotopic 13C / 12C on the sugars of the must at maturity) (Acevedo-Opazo et al., 2010, Gaudillère et al., 2002, Van Leeuwen et al., 2004).

4.6 Acknowledgment

This study was funded by Bordeaux Sciences Agro, the Regional Council of Aquitaine, the French Ministry of Agriculture and the Food-processing industry and Forest (programme CASDAR V907). This study was also supported by BIVB (Bureau interprofessionnel des vins de Bourgogne). We specially thank to geomatician Jean-Marc Brayer and two students Soufiane Ayachi, Jélesti Louamba for their helpful

preliminary works for this paper.

Prédiction spatiale pour les covariables liée à la malaide

4.7 Perspective pour la réserve utile

Après avoir quantifié le lien entre la réserve utile et la résistivité et estimé la réserve utile aux points de résistivité, nous allons prédire la réserve utile au pied de cep. Nous souhaitons utiliser la réserve utile comme un paramètre d'entrée pour le modèle de Lebon. Combinés avec les autres paramètres d'entrée temporels liés au climat, la sortie du modèle Lebon sera une variable spatio-temporelle qui mesure l'état de contrainte hydrique de la plante. Cette variable peut être utilisée dans le modèle spatio-temporel pour prédire et expliquer la variation spatio-temporelle de la maladie.

4.8 Prédiction spatiale pour les covariables écophysiologiques issues d'un échantillonnage géoréfencé

4.8.1 Introduction

L'objectif de cette section est de présenter les méthodes et les résultats de la prédiction spatiale pour deux paramètres retenus à inclure dans les modèles statistiques présentés dans chapitre 5. Deux parcelles, situées dans la région de Bordeaux, sont étudiées pour la modélisation. Cependant, les méthodes d'estimation de la réserve utile développées dans la partie 1 de ce chapitre ne conviennent pas ici car, entre autres, nous n'avons pas de variable auxiliaire. Il a été difficile d'estimer correctement la réserve utile à partir des points de sondage, à cause de la texture sableuse et caillouteuse du sol (région des Graves). Pour cette raison, nous avons choisi deux indicateurs supplémentaires liés à la plante : l'azote total et le $\delta^{13}C$ mesurés sur le jus de raisin. Le premier nous donne des informations sur l'état azoté de la plante, qui est relié à la vigueur de la plante (Van Leeuwen et al., 2009). L'état azoté de la plante est très relié aux caractéristiques du sol. Le $\delta^{13}C$ correspond au ratio C12/C13. Une plante en état de déficit en eau induit une limitation de l'incorporation de l'isotope C13 du CO2 de l'atmosphère. Cela conduit à une plus faible quantité de composés carbonés riches en C13. Les sucres produits par la vigne dans des conditions de contrainte hydrique ont un rapport C12/C13 différent de ceux produits par la plante sans contrainte hydrique. Cet indicateur nous donne une information sur l'état hydrique de la plante au stade de maturité des raisins (Van Leeuwen et al., 2011). Ces deux indicateurs sont intéressants pour étudier l'hétérogénéité intra parcelle. Ils sont faciles à mettre en œuvre à moindre coût. Nous décrivons dans cette partie, le protocole d'échantillonnage, les méthodes de géostatistique et les résultats des analyses statistiques.

4.8.2 Méthodes

4.8.2.1 Échantillonnage et dosage

Le dosage de l'azote total du jus de raisin a été réalisé au laboratoire de la Chambre d'Agriculture de Gironde, en utilisant la méthode IRTF dite par spectrométrie à

infra rouge, transformée de Fourier. Pour le rapport isotopique C12/C13, les dosages par spectrométrie de masse ont été réalisés au laboratoire CNRS Biogeosciences 694 de l'Université Bourgogne.

Pour chaque vignoble, 30 placettes sont échantillonnées et chaque placette contient 3 ceps. 50 baies de raisin sont sélectionnés dans chaque placette pour estimer une valeur de l'azote ou de $\delta^{13}C$ en laboratoire. Ces placettes sont échantillonnées de manière homogène dans le vignoble.

Des Krigeage ordinaires sont directement utilisés pour interpoler ces échantillons aux positions des ceps. En effet, comparé aux échantillonnages de la réserve utile, le nombre d'échantillons d'azote et $\delta^{13}C$ est double. De plus ils sont distribués uniformément dans la parcelle et en regardant leur variogramme Figure 4.8, nous trouvons une structure Gaussienne assez régulière.

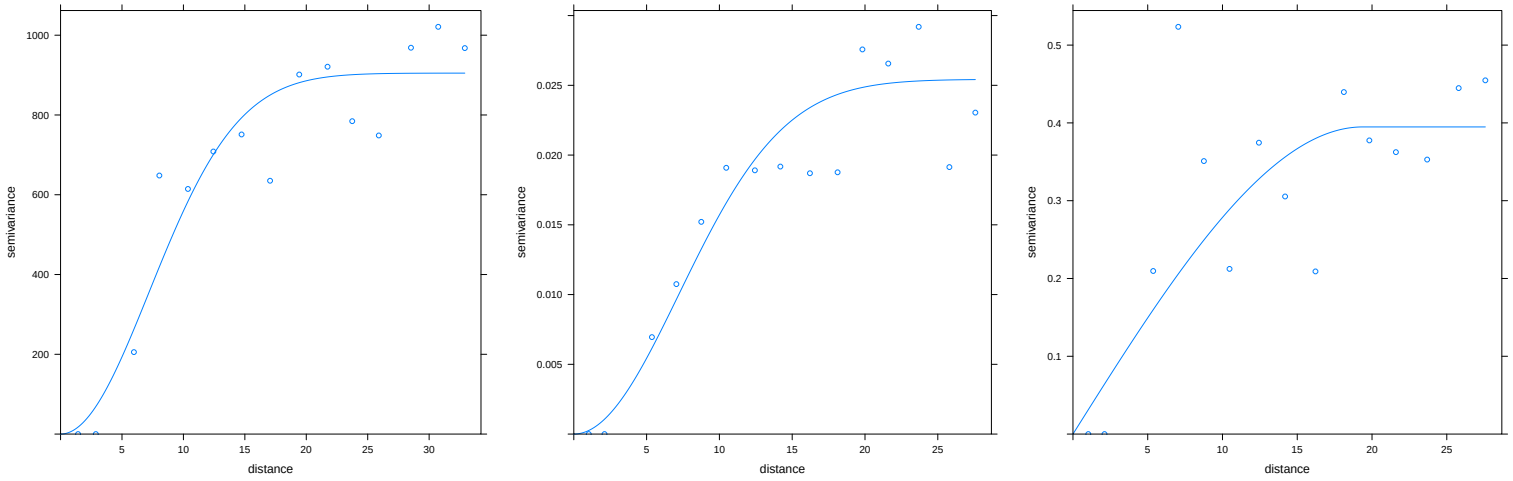


Figure 4.8: Variogrammes des échantillons. De gauche à droite : Azote total de la parcelle 12, la parcelle 8, $\delta^{13}C$ de la parcelle 8.

4.8.2.2 Quelques remarques

Pour la covariable azote, nous avons appliqué un krigeage ordinaire au logarithme de cette variable. Comme cette variable ne peut pas prendre de valeur négative, le krigeage ordinaire sur le logarithme permet d'éviter des estimations négatives par transformation inverse sur les estimations de $log(Azote)$. Cependant, la transformation inverse amène un biais (Cressie, 1993) car la transformation logarithme inverse induit une échelle exponentielle sur l'incertitude des estimations pour l'azote. Dans ce cas, une correction sur l'estimateur est fortement conseillée.

Malheureusement à cause de problèmes techniques, nous n'avons pas pu faire cette correction pour enlever le biais pour toutes les parcelles (des parcelles qu'on n'a pas montré dans la thèse) et pour la modélisation qui s'en suit. Cependant, nous avons comparé l'estimation biaisée et l'estimation corrigée pour une parcelle 8. Ces deux estimations sont assez proches. Nous montrons la différence entre ces deux estimations pour l'Azote à la parcelle 8 dans Figure 4.9. Nous trouvons que cette différence est faible compte-tenu de l'échelle de la variable qui varie de 20 à 200.

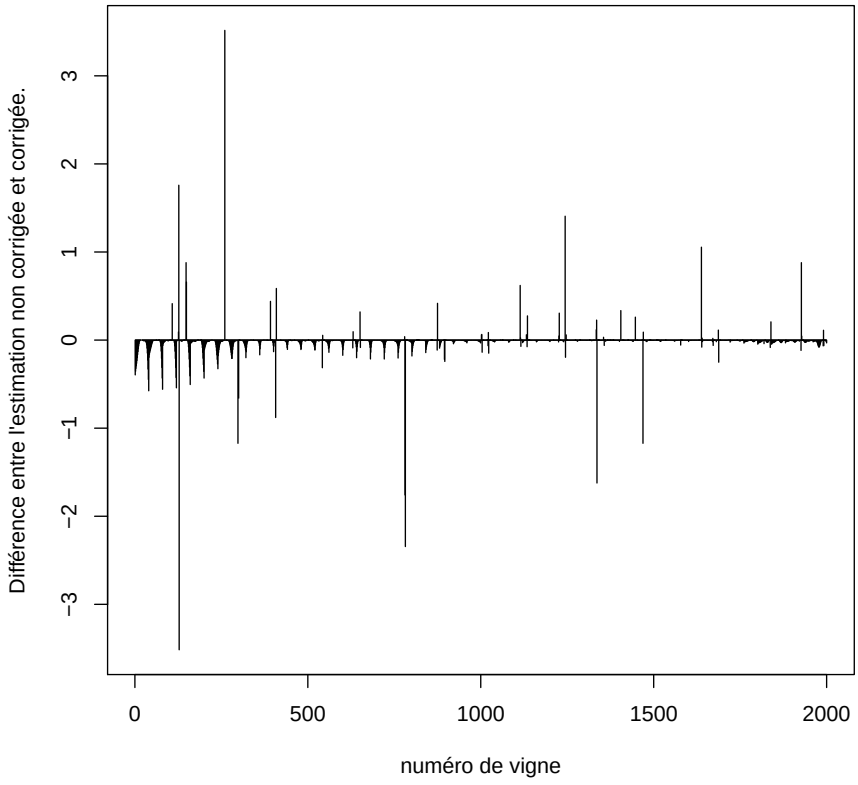


Figure 4.9: Différences entre l'estimation sans correction et l'estimation corrigée de la variable azote pour les ceps de la parcelle 8

4.8.3 Résultats

Nous présentons dans cette section les cartes d'échantillonnage et les cartes de prédiction pour l'azote de la parcelle 12 et l'azote et $\delta^{13}C$ de la parcelle 8.

Nous trouvons une plus forte variation pour l'azote en parcelle 12 qu'en parcelle 8, de plus, en parcelle 8 l'azote varie plus fortement que $\delta^{13}C$ Figure 4.10, 4.11.

Dans la parcelle 12, la distribution de l'azote montre une tendance à l'augmentation entre le Sud-est et le Nord-Ouest; pour la parcelle 8, la distribution de l'azote ne présente pas de direction privilégiée et la distribution de $\delta^{13}C$ est assez homogène.

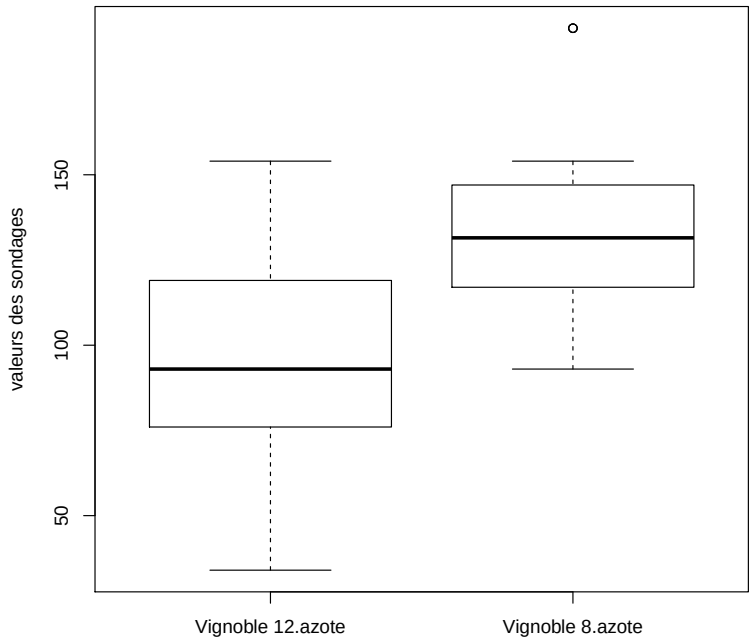


Figure 4.10: Boîte à moustaches des échantillons d'azote, de gauche à droite : les parcelles 12 et 8.

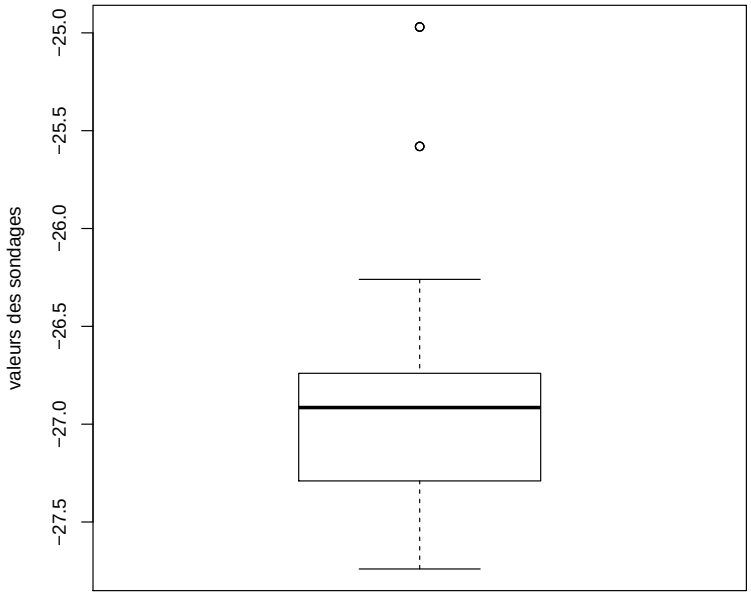


Figure 4.11: Boîte à moustaches des échantillons de $\delta^{13}C$ de la parcelle 8

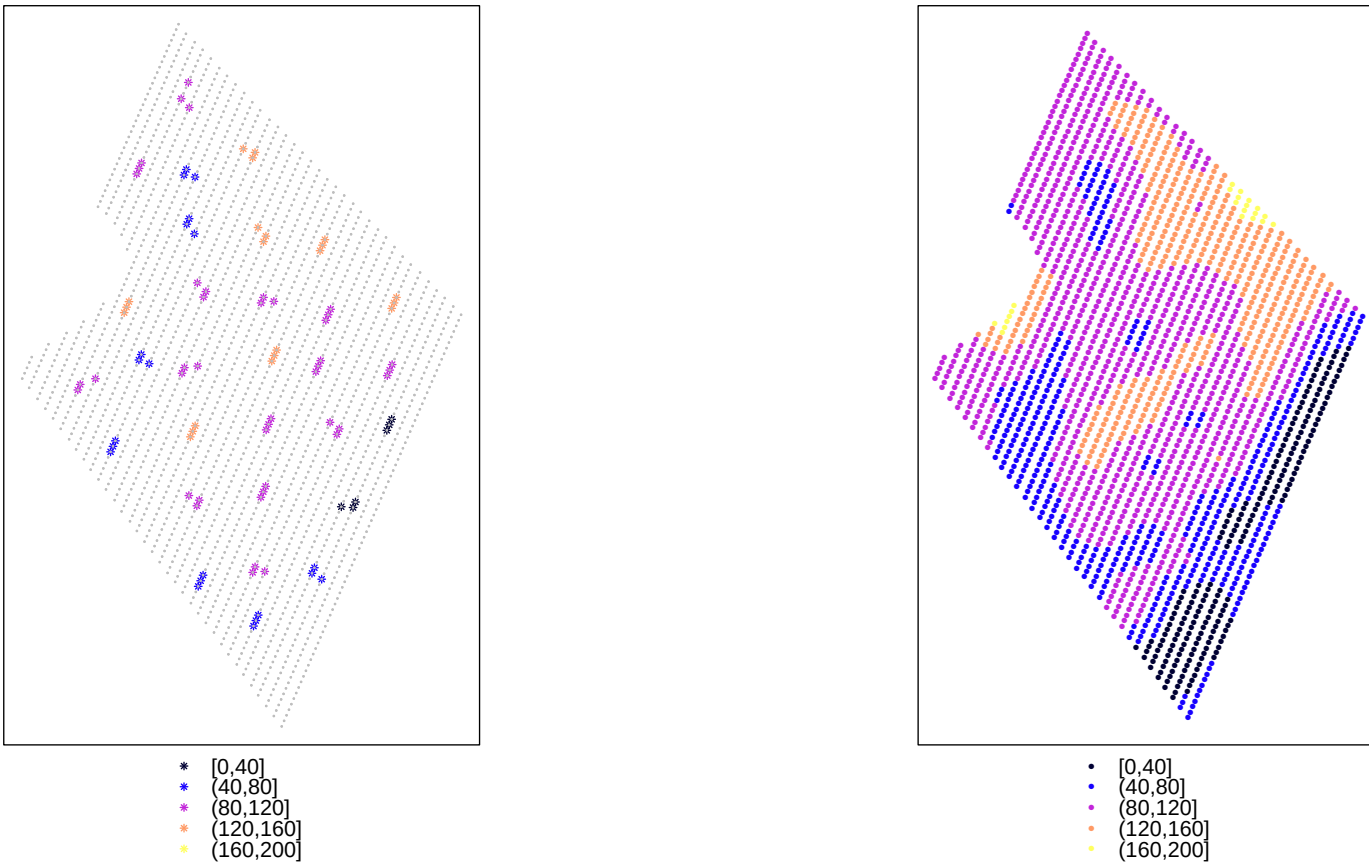


Figure 4.12: Parcelle 12. A gauche, géolocalisations des échantillons d'azote, à droite, carte des prédictions.

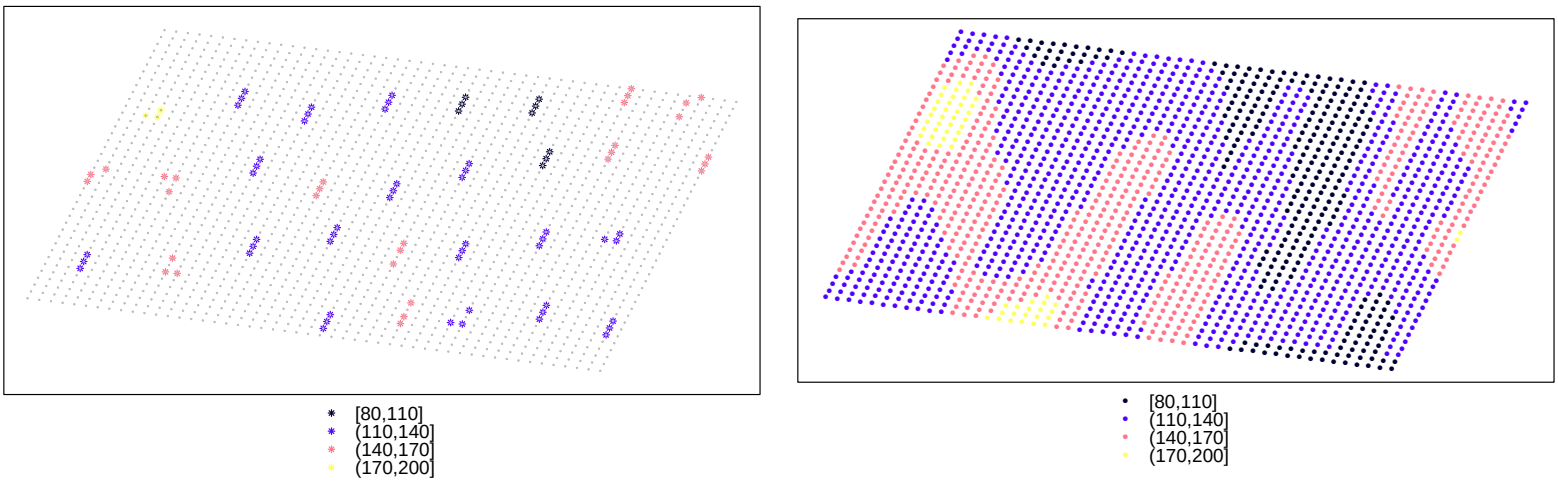


Figure 4.13: Parcelle 8. A gauche, géolocalisations des échantillons d'azote, à droite, carte des prédictions.

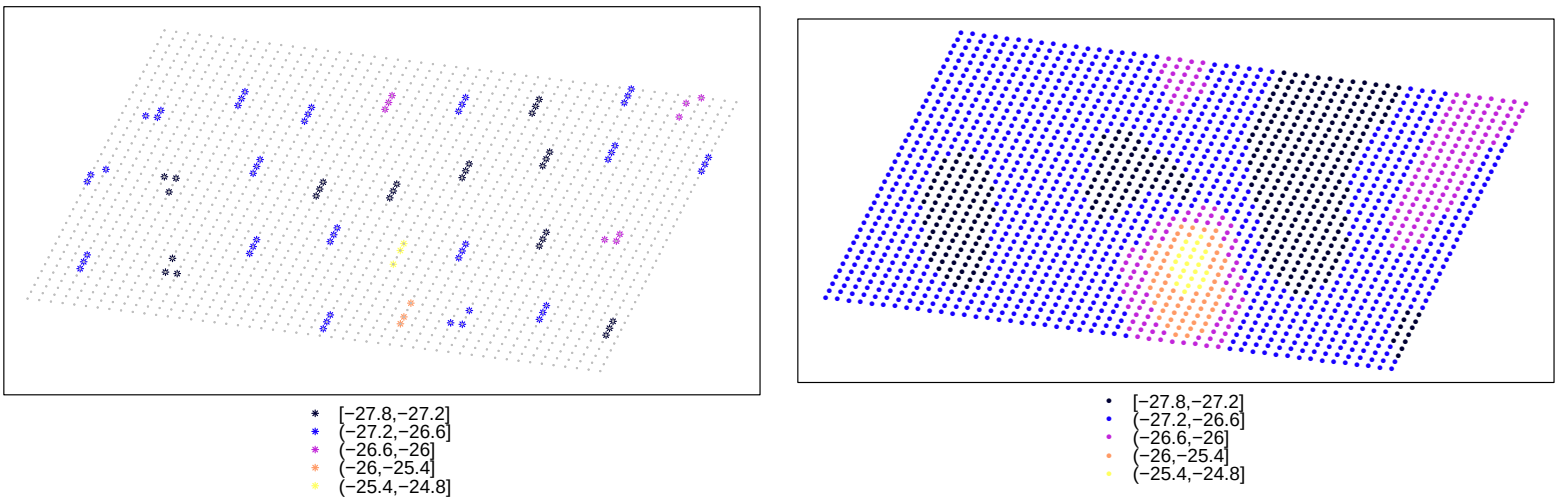


Figure 4.14: Parcelle 8. A gauche, géolocalisations des échantillons de $\delta^{13}C$, à droite, carte des prédictions.

Chapter 5

Modèles spatio-temporels de la dynamique de l'esca de la vigne

Introduction

L'esca est une maladie du bois complexe dont les nécroses internes sont reliées à des symptômes foliaires externes erratiques. A l'échelle de la plante, nous pouvons distinguer la première expression de la ré-expression des symptômes foliaires au cours de la chronologie des ceps infectés. A l'échelle de la parcelle, le taux de croissance de la maladie varie beaucoup suivant les parcelles, cela pourrait être dû à différents facteurs liés aux pratiques agronomiques, à l'environnement et aux champignons impliqués.

Dans la littérature, les motifs spatiaux de l'esca ont été analysés à l'aide de différentes méthodes statistiques (Surico et al., 2000a, Pollastro et al., 2000, Reisenzein et al., 2000, Surico et al., 2000b, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Stefanini et al., 2000, Zanzotto et al., 2013), ainsi que dans le Chapitre 3 Li et al. (2015). Les résultats montrent que la distribution des ceps symptomatiques à l'esca varie selon les parcelles d'aléatoire à fortement agrégée. Dans Li et al. (2015) (Chap 3), nous avons proposé que la complexité de la distribution de l'esca, qui présente différents niveaux d'agrégation, soulève la question de l'influence des facteurs environnementaux tels que le sol sur la distribution de l'esca. Jusqu'à présent, aucune étude n'a été réalisée pour examiner le rôle de l'environnement sur l'expression foliaire de l'esca à l'échelle de la parcelle et seulement deux études examinent la propagation secondaire en combinant les échelles spatiales et temporelles. Zanzotto et al. (2013) ont suggéré que la propagation de l'inoculum secondaire est à l'origine du motif agrégé. Ils ont étudié la manifestation des symptômes de l'esca dans une parcelle nouvellement plantée et comparé différents modèles avec differentes structures ajustées à leurs données en utilisant des modèles spatio-temporels bayésiens. Leurs résultats ne corroborent pas ceux de Li et al. (2015), puisque nous concluons dans cet article à une faible, voire inexistante, seconde propagation à partir des ceps symptomatiques. De plus, Zanzotto et al. (2013) n'ont pas intégré les covariables et ont seulement étudié la première expression de l'esca dans leur modèle. En utilisant un modèle statistique paramétrique, Stefanini et al. (2000) ont trouvé que la probabilité d'occurence de l'esca à l'échelle de la vigne augmente avec la variable « sanitary state in the close vicinity ».

Pour mieux comprendre les facteurs qui influencent la dynamique complexe de la maladie, nous précisons les composantes spatiales et spatio-temporelles à l'aide d'un modèle bayésien spatio-temporel hiérarchique. L'implémentation d'un tel modèle dont nous avons discuté au chapitre 2 nous permet de modéliser les données spatio-temporelles à partir d'un tel système, en ne prenant pas seulement en compte les covariables mais également les dépendances spatio-temporelles entre les ceps symptomatiques. Nous pouvons alors tester différentes hypothèses sur le rôle de différents facteurs environnementaux et nous pouvons alors confirmer ou infirmer le rôle d'une seconde propagation locale en utilisant des régressions logistiques. Nous estimons ces modèles en utilisant une nouvelle approche: la méthode *Integrated Nested Laplace Approximations (INLA)*, du fait de ses bonnes propriétés algorithmiques et de sa capacité à traiter des modèles complexes (Rue et al., 2009, Cameletti et al., 2013).

Nous avons choisi deux facteurs spatiaux pour expliquer l'hétérogénéité spatiale de la distribution des ceps malades : un indicateur de stress hydrique de la plante, Delta C13 qui estime l'état de contrainte hydrique et l'azote total du jus de raisin qui représente la vigueur des plants (chapitre 4). Deux facteurs climatiques sont également sélectionnés en utilisant une étude préliminaire basée sur les études de prévalence et d'incidence de plusieurs parcelles de la région de Bordeaux. Nous avons développé plusieurs modèles pour reproduire les données binaires d'occurrence et première occurrence des symptômes foliaires, avec pour objectifs de : (1) déterminer l'effet des prédicteurs décrits précédemment sur la propagation de l'esca, (2) étudier la dépendance spatiale ou spatio-temporelle entre vignes malades, (3) analyser l'effet de la vigne déjà malade à proximité sur le risque d'occurrence d'esca, (4) tester l'effet de rang sur la maladie.

Dans ce chapitre, nous allons tout d'abord décrire nos données et présenter les différents modèles dans la section 2. Ensuite, nous donnons une courte description des méthodes d'inférence de notre modèle. Dans la section 3, nous présentons les estimations et comparaisons de ces modèles. Enfin, nous discutons les résultats.

Résumé

Dans ce chapitre, pour mieux comprendre les facteurs qui influent sur la dynamique complexe d'esca et pour vérifier les hypothèses proposées dans le chapitre 3 pour la propagation secondaire de la maladie, nous avons travaillé sur les données spatio-temporelles de première occurrence et occurrence d'esca dans deux parcelles bordelaises enregistrées sur 10 ans (2004-2013). Les données sont ajustées par des modèles spatio-temporels hiérarchiques qui précisent les composantes spatiales et spatio-temporelles. Ce type de modèle est estimé par la méthode *Integrated Nested Laplace Approximations (INLA)*, du fait de la qualité des ses propriétés algorithmiques et de sa capacité à traiter des modèles complexes. Nous avons dérivé plusieurs modèles en intégrant la régression sur les facteurs environnementaux et/ou sur les indicateurs de voisins qui sont précédemment malades, avec un bruit spatialement ou spatial-temporellement dépendant, afin de poursuivre les objectifs suivants : (1) déterminer l'effet des prédicteurs décrits précédemment sur la propagation de l'esca (2) étudier la dépendance spatiale ou spatio-temporelle entre les vignes malades, (3) analyser l'effet de la vigne déjà malade dans un voisinage proche sur le risque d'occurrence d'esca, (4) tester l'effet de rang sur la maladie. Les comparaisons entre

les différents modèles de régression sur les facteurs environnementaux et les indicateurs de voisins déjà malades pour la première occurrence d'esca montrent que la dépendance spatiale existe seulement entre les ceps nouvellement malades. Les ceps voisins précédemment malades n'ont pas une influence significative pour la première expression de la maladie et il n'y a pas d'effet privilègié selon le rang. Ces résultats ont confirmé l'hypothèse proposée dans le chapitre 3 sur la non propagation secondaire de l'esca à partir de ceps symptomatiques. Notre résultat montre aussi que l'azote total du jus de raisin a un effet significatif pour le développement de l'esca. Cela suggère que dans les zones de la parcelle où les vignes sont plus vigoureuses le risque d'exprimer les sympômes foliaires est plus élévé. De plus, la covariable climat liée à la température, a montré un effet significativement négatif pour l'expression d'esca. Notre étude a également montré que les modèles spatialement corrélés et spatio-temporellement corrélés ajustaient mieux les données de première occurrences et occurrences d'esca. Cela suggère l'existence d'une dépendance spatiale soit entre des ceps symptomatiques, soit entre ceux nouvellement symptomatiques. De plus, nous avons trouvé une forte dépendance temporelle c'est-à-dire qu'un cep ayant exprimé des symptômes au temp t a une forte probabilité de réexprimer les symptômes au temp $t + 1$.

Modelling spatiotemporal dynamics of esca grapevine disease at vineyard scale

Shuxian Li[a,b] , **Frédéric Fabre**[a], **Lucia Guérin-Dubrana**[a,b] and **Anne Gégout-Petit**[c]

[a]Université de Bordeaux, ISVV, UMR-1065 INRA
[b]Bordeaux Sciences Agro, Gradignan, France
[c]Institut Elie Cartan, Université de Lorraine, INRIA BIGS Nancy Grand-Est, Nancy, France

Abstract

In this chapter, in order to better understand how the factors influence the complex dynamic of esca and to verify the hypothesis proposed in chapter 3 on the secondary spread of the disease, we studied on spatio-temporal data of the first occurrence and occurrence of foliar symptoms in two vineyards at Bordeaux region. The data were recorded over 10 years, from 2004 to 2013. The hierarchical logistic regression models were used to fit the data.

Such models were estimated by *INLA (Integrated Nested Laplace Approximations)* approach due to its good computational property and capability to deal with the complex models.

We derived several models by integrating regressions on the environmental factors and the indicators of neighboring vines which are already diseased, with a latent process spatially or spatio-temporally auto-correlated in order to respond to these following objectives : (1) to determine the effect of the previous described predictors on esca spread, (2) to study the spatial and the spatio-temporal dependence between diseased vines, (3) to analyze the effect of the previously-diseased vine in a close neighborhood on the risk of esca occurrence, (4) to test the row effect on the disease.

The comparisons between the spatial logistic regression models (residuals are spatially auto-correlated) together on environmental factors and former neighbors indicators for the first occurrence of Esca show that the spatial dependence exists only between newly diseased vines and neighboring vines previously diseased do not have a significant influence for the first occurence the disease and there is no privilege effect by row. These results confirmed hypothesis proposed in Chapter 3 on non secondary spread of esca from symptomatic vines. Our result also shows that the total nitrogen grape juice has a significant effect on the development of Esca. This suggests that in areas of vineyard where the vines are more vigorous, the risk to express foliar symptom is higher. In addition, the covariate climate related to temperature, showed a significantly negative effect on the expression of Esca.

Our study also selected models spatially correlated and spatio-temporally correlated which better adjusted data first occurrences and occurrences of esca symptoms respectively. This implies a spatial dependence exists between symptomatic vines and between those newly symptomatic. Moreover, we found a strong temporal dependence as a vine expressed foliar symptoms at time t has a high probability of re-expressed symptom at time $t + 1$.

5.1 Introduction

In the case of complex plant diseases, with etiology not fully elucidated, stochastic spatio-temporal modelling approach may be useful to identify key factors implied in the spread of the disease and to highlight some of the underlying ecological processes (?Gibson and Austin, 1996). For instance, fungus and virus plant diseases have been the subjects of numerous studies to analyze temporal space structure and to model their dynamics.

In viticulture, the esca disease remains a such complex disease or sometimes called a complex of several diseases(Mugnai et al., 1999, Surico et al., 2006). At least, two or three pathogenic fungus, acting in succession or in combination are involved (Larignon and Dubos, 1997, Surico et al., 2006). Each fungus produces different types of necroses in woody part of trunk and arms of the vine (Mugnai et al., 1999). Internal cryptic necroses are related to erratic external foliar symptom.

At vine scale, we can distinguish the first expression and the re-expression of foliar symptoms during the chronology of infected vines. The first foliar expression is related, at least in part, to an unknown length of time that the vine has been colonized with fungi. And the foliar re-expression occurs once or twice in a period of 3 to 4 years before the death of a vine(Guérin-Dubrana et al., 2013).

At vineyard scale, a linear increase of cumulative incidence of esca was observed over time(Surico et al., 2006, Li et al., 2015). The rate of disease increase greatly varied according vineyards, this should be related to various factors link to agronomic practices, environment and implied fungus. The rate of foliar expression varied yearly according to the climate conditions in spring and summer (Marchi et al., 2006).

Spatial pattern of esca was analyzed using different statistical methods also at vineyard scale, (Surico et al., 2000a, Pollastro et al., 2000, Reisenzein et al., 2000, Surico et al., 2000b, Edwards et al., 2001, Redondo et al., 2001, Sofia et al., 2006, Stefanini et al., 2000, Zanzotto et al., 2013). According to the vineyards, the distribution of esca expressed vines varied from random to strongly aggregated patterns.

Based on the data set collected from 15 vineyards of Bordeaux region and by applying several non-parametric tests, based on join count statistics and permutation methods, Li et al. (2015)(Chap 3) also found a large range of spatial patterns among vineyards from random to strongly structured. In four vineyards, the complex esca distribution indicating different levels of clustering raises questions about the environmental factors such as soil in driving esca distribution. Surico et al. (2000b) related the heterogeneity of esca distribution to the variation of soil water content. They observed a low number of diseased vines in areas with a low soil content. At our knowledge, the spatial heterogeneity within vineyard with regard to the spatial structure of esca has never been explored using spatial statistical tools.

To include the time scale in the temporal analysis should help to differentiate the components of esca dynamics. Using non-parametric analysis, applied on spatio-temporal data, Li et al. (2015) (Chapter 3) explored the contagious process at short scale by testing the spatial relationship between the previously- and newly-symptomatic vines. They concluded on a slight or no, secondary spread from symptomatic vines. Their results did not corroborate with those of Zanzotto et al. (2013) and Stefanini et al. (2000). Using hierarchical Bayesian spatio-temporal mod-

els, the results of Zanzotto et al. (2013) suggested a secondary disease spread along row. By applied a parametric statistical model, Stefanini et al. (2000) found that the probability of esca occurrence at vine scale increased with the variable "sanitary state in the close vicinity". To our knowledge, these modelling studies are the only ones which explored the secondary spread of esca by using data from one vineyard for each study. No modelling study has been done before to both explore the secondary spread and the role of environmental factors on esca dynamics. The implement of Bayesian space-time hierarchical model, discussed in chapter 2, gives us the opportunities to test competing hypotheses about environmental factors and contagious processes affecting the short-scale dynamics. It also permits to explore the effect of spatial heterogeneity of environment on esca spread and to account for the spatio-temporal dependencies between symptomatic vines.

It assumes that a pair of individuals close in space and time are more correlated than the distant ones. In particular, in chapter 2, we have proposed dependence structures easy to implement and appropriate to our disease data. At spatial scale, we propose a Markov random field which limit the spatial dependence in a certain neighborhood and in temporal scale, an auto-regressive of order 1 which takes into account a one-year memory of process. The auto-regressive order could also be extended to $k > 1$ to study more temporal properties, however, due to the fact that we do not have many years of data, we comment such results with caution. It has to be noticed that this hierarchical structure explicits the spatial and temporal behaviors.

Such models will be estimated by a novel approach: *Integrated Nested Laplace Approximations (INLA)*, due to its good computational property and capability to deal with the complex models (Rue et al., 2009, Cameletti et al., 2013).

To determine the impact of spatial heterogeneity of the environment on the distribution of vines affected by esca, the factors have been chosen from two assumptions. The first one is that water status of the vine determines the susceptibility of the plant and / or the development of the disease. In that case, we want to study if the risk that a vine expressing esca is higher when its water status is not constrained. The second assumption concerned the effect of vine vigor on esca development : the underlying assumption being that the vine susceptibility to esca is greater in more vigorous plants. The water stress condition was estimated at each vine location from regular sampling measurements of a water stress indicator, $\delta^{13}C$ (see Chapter 4). This indicator corresponds to the isotopic ratio 12C / 13C of carbohydrates, sugars measured on the vine grape juice mature (Van Leeuwen et al., 2011). To map the status of vine vigor, we used data from regular point sampling of the available nitrogen content of grape juice (Van Leeuwen et al., 2011). This indicator reflects the nitrogen status of the vine, connected to a level of vigor. Two climatic factors have also been selected from a preliminary study using the data from Bordeaux region.

From a logistic regression model, including a latent process, we derived several models to fit binary data of foliar symptom occurrence or first occurrence in order to respond to these following objectives : (1) to determine the effect of the previous described predictors on esca spread, (2) to study the spatial and the spatio-temporal dependence between diseased vines, (3) to analyze the effect of the previously-diseased vine in a close neighborhood on the risk of esca occurrence, (4) to test the row effect on the disease.

According to the definition of the diseased vine neighborhood, we studied the assumption of a density-dependent transmission. In that case the density of the disease was defined by the number of diseased neighbor cases and their severity based on the number of expression and re-expression in the past. To take into account the geometry of the vineyard, we also studied the effects of diseased vine neighbors situated on the same row and out of row on the probability of first occurrence of esca. We also tested the row effect on the dynamics of esca : does the probability of esca occurrence vary according the along-row location in the vineyard ?

In this chapter, we will first describe our data and present the different models in section 2, then we give a brief description of the method of inference of our models. In section 3 presents the estimations and comparisons of these models and we give the discussion of the results and perspectives in the last section.

5.2 Materials and methods

5.2.1 Monitored vineyards and data disease collection

Two of the fifteen commercial vineyards in the Bordeaux region, monitored for esca disease for 10 consecutive years, from 2004 to 2013 were selected for this study. By a preliminary study exploring spatio-temporal patterns of esca for 15 vineyards (Chapter 3), two vineyards, called Vineyard 8 and Vineyard 12, were selected because they showed aggregated structures at small scale.

Vineyard 8 and Vineyard 12 were, respectively, planted in Martillac and Castres with the cultivar Cabernet Sauvignon (Vitis vinifera L.), and were trained in accordance with the Guyot method. Within plot, the distance was 1 m, 1 m for Vineyard 8 and 1.2 m, 1.4 m for Vineyard 12, between row and between vines within each row respectively.

The number of living vines, monitored at the beginning of the observation period, was 2000 and 2281, for Vineyard 8 and Vineyard 12 respectively. As the vineyard plots did not form a regular lattice, the location of each vine was estimated using records with a global positioning system (GPS) device.

From 2004 to 2013, at the end of August, all of the contiguous vines from the both vineyards were individually surveyed for esca foliar expression included "tiger-stripe" patterns (chronic form) and the wilting of some branches (acute form). Newly-planted, and dead or missing plants were also recorded. The yearly prevalence of esca symptomatic vines, defined as the ratio between the total number of vines exhibiting esca symptom at one year on the number of living vines, was calculated, as well as the annual incidence of esca using the newly cases of foliar symptomatic vines from 2005. The newly case of esca is defined by a vine that expressed foliar symptoms for the first time. The georeferenced individual vine data was used.

5.2.1.1 Climatic data and annual covariates

The key meteorological parameters (daily air temperature and daily rainfall) were provided from the Météo France database. They are produced by spatial interpolation using the SAFRAN analysis system. From the parameters measured at ground level, they are interpolated on an 8×8 km^2 grid covering France (Quintana-Seguí

et al., 2008). Two grid data (N^0 7502 and 7503 respectively for Vineyard 8 and Vineyard 12) were used to calculate the climatic annual covariates.

A preliminary statistical study showed that some climate variables were linked to the incidence and prevalence of esca (computed for different vineyards of Bordeaux region). Two of those were selected for covariates of models: the sum of total rainfall for a year (RRtotal) and the number of days with minimum temperature inferior or equal to zero degree Celsius during spring (from 1st of March to the 31th of May).

5.2.1.2 Spatial covariates

To investigate the impact of vine water and nitrogen status on the dynamics of esca, and to account for the spatial variability intra-vineyard, we used two spatial indicators estimated at vine scale: the $\delta^{13}C$, that indicated vine water status during grape ripening (Van Leeuwen et al., 2011, Gaudillère et al., 2002) and the total nitrogen of must (grape juice extract) (Van Leeuwen et al., 2009). This last one was used to monitor vine nitrogen status within the vineyard and was related to the vine vigor. Both covariates were estimated at each vine location by spatial extrapolation using kriging method (see chapter 4).

We called these two covariates, Nitrogen and $\delta^{13}C$, the environmental covariates also, as in a sense, these plant indicators reflect an environmental status.

5.2.2 Spatio-temporal modelling

A key aspect of accounting for the spatial and temporal dynamics of esca disease is to distinguish the impacts of environmental or physiological factors and sanitory state of neighborhood. Transitions from asymptomatic state to symptomatic state could be associated with environmental conditions that have impacts on physiological state of vine. Transitions can also be associated with presence of already infected vines in a local neighborhood. A range of models based on logistic regression with a latent process was set up to determine the probability of esca occurrence and of first occurrence in vineyards. They were defined by their conditional law knowing the covariates and the latent process. Such models with the hierarchical structures have been discussed in Cressie and Wikle (2011) to model spatio-temporal data (see also Chapter 2, section 4.3).

5.2.2.1 Variables of interest

Taking into account the discontinuity of symptom appearances on vines, we distinguished the first occurrence of foliar symptom corresponding to the first stage of visible disease from the foliar expression of esca at a given year. We investigated the effect of factors on the occurrence of symptom as well as the effect of factors on the first occurrence .

Consider a lattice of grapevines in which the spatial location of the vine labeled by i is given by its coordinates (u_i, v_i). Let $Y_{i,t}$ be the binary random variable indicating the presence-absence of esca foliar symptoms associated with vine i at year t and $y_{i,t}$ be its realization.

To regard esca-status of a vine i at time t, we define $Z_{i,t}$ as follows:

$$Z_{i,t} = \begin{cases} 0 & \text{if } Y_{it'} = 0 \text{ for all } t' \leq t \\ 1 & \text{if } Y_{it'} = 1 \text{ for any } t' \leq t \end{cases}$$

Note that $Z_{i,t}$ is fully known if we know the history of $Y_{i,.}$ until t. The purpose of this chapter is to model the probability to express symptom whatever the past, that the probability of $(Y_{i,t} = 1)$ and also to model the transition of $Z_{i,t}$ from 0 to 1 that is the probability to express symptom for the first time.

5.2.2.2 Modelling the occurrence of esca foliar symptom

We proposed a model for the occurrence of esca foliar symptom. The model is determined by the marginal conditional law of each $Y_{i,t}$ knowing the vector of covariates $\mathbf{X}_{it}$ and a latent auto-regressive (AR) process of order 1.

It is defined by the following elements:

$$\begin{aligned} Y_{i,t}|c_{i,t} &\sim Bernoulli(c_{i,t}), && (5.1) \\ c_{it} &= \mathbb{P}(Y_{i,t} = 1|\xi_{i,t}, \mathbf{X}_{i,t}), \\ logit(c_{i,t}) &= \mathbf{X}_{i,t}\boldsymbol{\beta} + \xi_{i,t}, && (5.2) \\ \xi_{i,t} &= \rho_1 \xi_{i,(t-1)} + \omega_{i,t}, && (5.3) \end{aligned}$$

where conditionally on $(\xi_{i,t}, \mathbf{X}_{i,t})$, $Y_{i,t}$ is independent of the other processes $(Y_{j,t})_{\{0 \leq t\}}$ and of its past $(Y_{i,t'})_{\{t'<t\}}$.

It means that $Y_{i,t}$ depends conditionally on its own covariates $\mathbf{X}_{it}$ and the latent process ξ_{it} where $\mathbf{X}_{i,t} = (X_{i,t,1}, ..., X_{i,t,p})$ denotes a vector containing the values of the p measured covariates, including an intercept. $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is the $p-$length vector of regression coefficients estimated. The spatio-temporal process $\omega_{i,t}$ has a zero-mean Gaussian distribution, and it is assumed to be temporally independent and is characterized by the spatio-temporal covariance function:

$$Cov(\omega_{i,t}, \omega_{j,t'}) = \begin{cases} 0 & \text{if } t \neq t' \\ \sigma^2_\omega C(h) & \text{if } t = t' \end{cases}$$

where $C(h)$ is of Matérn class:

$$C(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^\nu K_\nu(\kappa h) \quad (5.4)$$

with unknown parameters $\kappa > 0$ for scale related to the range γ and $\nu > 0$, a parameter presents the order in the modified Bessel function K_ν. We ususally keep ν fixed and estimate the range γ, it is estimated by $\gamma = \frac{\sqrt{8\nu}}{\kappa}$ and corresponds to the distance where the spatial correlation is close to 0.1(Cameletti et al., 2013).

Other variants can be proposed. For instance, the latent process ξ_t can be also be modeled as an AR process of order 2 as follows to measure the 2-year dependence:

$$\xi_{it} = \rho_1 \xi_{i(t-1)} + \rho_2 \xi_{i(t-2)} + \omega_{it}, \quad (5.5)$$

or a latent process without temporal dependence:

$$\xi_{it} = \omega_{it}. \quad (5.6)$$

This last model assume a temporal independence, i.e. $\rho = 0$, with outcome does not depend on the previous years and with the spatial field can vary from year to year. The model corresponding to Equation 5.6 is denoted spatial replicate (SR) model.

To find the model that fit the data at best, we compared the models with different types of spatio-temporal dependence.

If the spatial heterogeneity among rows exhibited in the by-row distributed data can not be explained by the covariates and the temporal dependence on the previous states, a row-specified random effect can be integrated to the model as follows:

$$logit(c_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} + \xi_{it} + R_{r|i\in n_r} \tag{5.7}$$

where $i \in \{n_r\}$ means that location i belongs to the r_{th} row. Morever, the R_r are independent and $R_r \sim N(0, \sigma_R^2)$.

5.2.2.3 Modelling the first occurrence of esca foliar symptom

In this section, we want to model the probability of foliar expression for asymptomatic vine (risk of first expression). The variable indicator of a first infection is a Bernoulli:

$$Z_{i,t}|p_{i,t}, Z_{i,t-1} \begin{cases} \sim Bernoulli(p_{i,t}) & \text{if} \quad Z_{i,t-1} = 0 \\ = 1 & \text{if} \quad Z_{i,t-1} = 1 \end{cases}$$

and the following model is proposed for $p_{i,t}$:

$$M0 : logit(p_{i,t}) = \mathbf{X}_{it}\boldsymbol{\beta} + \omega_{i,t} \tag{5.8}$$

where ω_{it} follows a zero-mean Gaussain distribution defined as in Equation 5.2.2.2.

To further study the influence of previously-infected neighboring vines on the risk of first esca expression, the model was extended by incorporating different neighbor terms. In this paper, we consider a neighborhood structure with neighboring order 3 referred to the results of Zanzotto et al. (2013) who found that the optimal neighboring order to consider in a logistic regression model is 3. Moreover, from neighbor tests in chapter 3, for one third of the studied vineyards, the newly-symptomatic vines significantly related to the previous vines within neighbor order 3. From a practical point of view, it is easy to define the neighboring vines according to their coordinates (u_i, v_i) by taking into account the non-equal distance between two adjacent rows (Δu) and two adjacent vines along the same row (Δv). The set N_i of neighbors of vine i is then defined by:

$$N_i = \{j | j \neq i, \quad \frac{(u_i - u_j)^2}{\Delta u^2} + \frac{(v_i - v_j)^2}{\Delta v^2} \leq 9\} \tag{5.9}$$

Different definitions of neighbors were determined according to the foliar esca expression at time t or $< t$ and also to their location, around the vine i, along the same row than vine i or not.

$$
\begin{aligned}
N^0_{i,t} &= \{(j,t)|j \in N_i, Z_{j,t-1} = 1\} \\
Nr^0_{i,t} &= \{(j,t)|(j,t) \in N^0_{i,t}, v_i = v_j\} \quad Nor^0_{i,t} = \{(j,t)|(j,t) \in N^0_{i,t}, v_i \neq v_j\} \\
N^1_{i,t} &= \{(j,t)|j \in N_i, Z_{j,t-1} = 1, Z_{j,t-2} = 0\} \\
N^2_{i,t} &= \{(j,t)|j \in N_i, Z_{j,t-1} = 1, Z_{j,t-2} = 1\}
\end{aligned} \tag{5.10}
$$

$N^0_{i,t}$ corresponds to the neighbors which had previously expressed foliar symptom at least one time at $t-1$ or before. In that case, two neighbor divisions are determined: $Nr^0_{i,t}$ and $Nor^0_{i,t}$ corresponded to neighbors located on the same row than vine i and those located in different row than vine i, respectively. $N^1_{i,t}$ and $N^2_{i,t}$ correspond to the neighbors which had previously expressed foliar symptoms respectively, only at $t-1$ and, at least, another time at $t-2$ or before. Let us note that $N^0_{i,t} = Nr^0_{i,t} \cup Nor^0_{i,t}$ and $N^0_{i,t} = N^1_{i,t} \cup N^2_{i,t}$. We then calculated the proportion of each neighbor division among neighbors to construct the associated covariates:

$$
\begin{aligned}
PN^0_{i,t} &= \frac{|N^0_{i,t}|}{|N_i|} \quad PNr^0_{i,t} = \frac{|Nr^0_{i,t}|}{|N_i|} \quad PNor^0_{i,t} = \frac{|Nor^0_{i,t}|}{|N_i|} \\
PN^1_{i,t} &= \frac{|N^1_{i,t}|}{|N_i|} \quad PN^2_{i,t} = \frac{|N^2_{i,t}|}{|N_i|}
\end{aligned} \tag{5.11}
$$

where if A is a set, $|A|$ stands for its cardinality.

The derived covariates allows us to take the following irregularities of the vineyard into account:

- plants at the edge
- plants with missing neighbors

By incorporating the neighborhood covariates described above, four models were constructed to study the effects of previously diseased vines in vicinity on the risk of first expression:

$$M1 : logit(p_{i,t}) = \mathbf{X}_{it}\boldsymbol{\beta} + \eta PN^0_{i,t} + \omega_{i,t} \tag{5.12}$$

$$M2 : logit(p_{i,t}) = \mathbf{X}_{it}\boldsymbol{\beta} + \eta_1 PN^1_{i,t} + \omega_{i,t} \tag{5.13}$$

$$M3 : logit(p_{i,t}) = \mathbf{X}_{it}\boldsymbol{\beta} + \eta_1 PN^1_{i,t} + \eta_2 PN^2_{i,t} + \omega_{i,t} \tag{5.14}$$

$$M4 : logit(p_{i,t}) = \mathbf{X}_{it}\boldsymbol{\beta} + \eta_r PNr^0_{i,t} + \eta_{or} PNor^0_{i,t} + \omega_{i,t} \tag{5.15}$$

At last, integration of a random effect helps to explain the heterogeneity among rows, similarly as presented before, we have:

$$logit(p_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} + \omega_{it} + R_{r|i \in n_r} \tag{5.16}$$

The different models are parametrized by $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\eta}, \rho, \sigma^2, \sigma^2_R, \kappa\}$, where $\boldsymbol{\beta}$ stands for the p-vector of coefficients of the regression on the vector of environmental covariates $\mathbf{X}_{i,t}$, $\boldsymbol{\eta}$ for the regression on the different neighborings, $\boldsymbol{\rho}$ for the temporal auto-regression coefficients, κ for the parameter of the covariance of the noise. The dimension of $\boldsymbol{\theta}$ varies along the models.

5.2.3 Inference

Common approach to draw inference of such hierarchical structures is to estimate the parameters from their marginal posterior distributions using Bayesian methods (Banerjee et al., 2014). MCMC (Markov Chain Monte Carlo) simulations are often used to approximate the posterior distributions. However, in this paper, due to a large data-set, about 2000 sites in a field over 10 years, the use of the packages such as *Openbugs* to run this model using MCMC becomes computationally impractical.

For the current inference, we applied the INLA (*Integrated Nested Laplace Approximations*) method to draw approximate Bayesian inferences. INLA is a computationally effective algorithm that produces fast and accurate approximations to posterior distributions, and thus an attractive alternative to MCMC. INLA have been first proposed by Rue et al. (2009), it meets a big success in various application domains after combined with the SPDE (Stochastic Partial Differential Equations) approach (Lindgren et al., 2011). These applications domains include pathology (Jousimo et al., 2014), environment (Shaddick and Zidek, 2014) and so on.

5.2.3.1 Bayesian Inference

To infer the value of $\boldsymbol{\theta}$. If $\mathbf{y} = (y_t)_{(1 \leq y \leq T)}$, the posterior distribution in a Bayesian framework is as follows:

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) &\propto \pi(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\theta})\pi(\boldsymbol{\xi}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= (\prod_{t=1}^{T} \pi(\mathbf{y}_t|\boldsymbol{\xi}_t, \boldsymbol{\theta}))(\pi(\boldsymbol{\xi}_1|\boldsymbol{\theta}) \prod_{t=2}^{T} \pi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}, \boldsymbol{\theta}))\pi(\boldsymbol{\theta}) \end{aligned} \tag{5.17}$$

5.2.3.2 INLA and SPDE approaches

INLA approach is based on the following approximation $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ of the marginal posterior of $\boldsymbol{\theta}$ (Rue et al., 2009):

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\xi}|\boldsymbol{\theta}, \mathbf{y})}|_{\xi = \xi^*(\boldsymbol{\theta})} \tag{5.18}$$

where $\tilde{\pi}_G(\boldsymbol{\xi}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{\xi}$, and $\xi^*(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{\xi}$, for a given $\boldsymbol{\theta}$.

The SPDE (Stochastic Partial Differential Equations) approach introduced by Lindgren et al. (2011) aims to find a GMRF (Gaussian Markov Random Field) model that best represents the current Gaussian field with Matérn covariance since GMRF model has good computational properties thanks to the sparse precision matrix.

Basically the SPDE approach uses a finite element representation to define the Matérn field as a linear combination of basis functions defined on a triangulation of the domain $\mathcal{D}$. This consists in subdividing $\mathcal{D}$ into a set of non-intersecting triangles meeting in at most a common edge or corner.

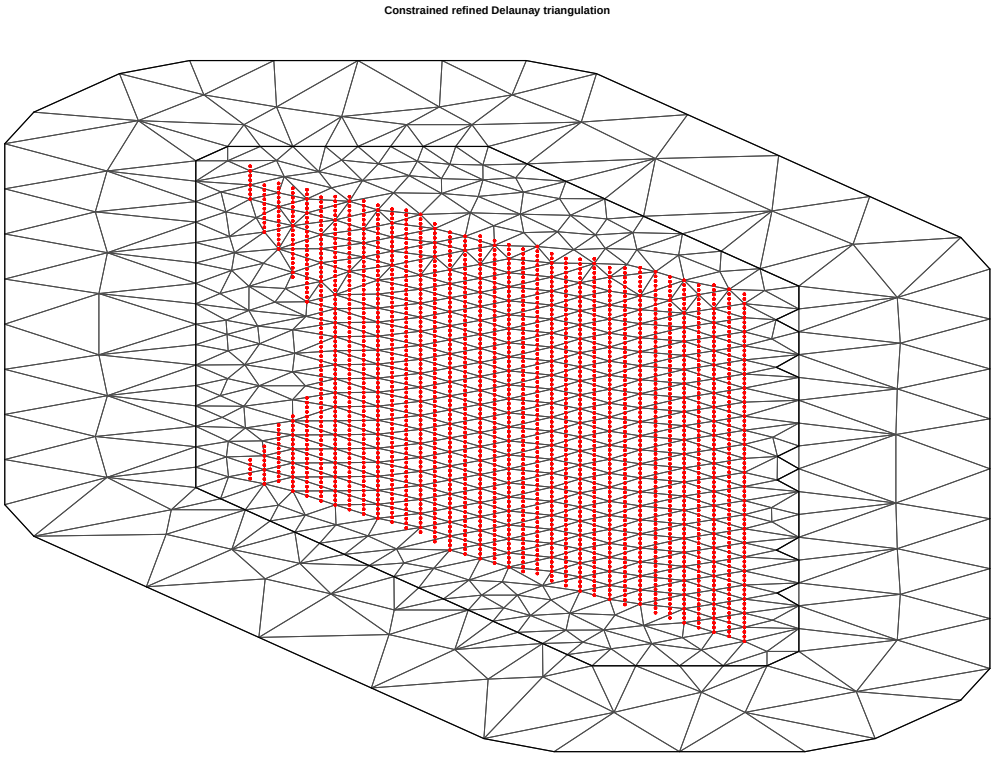


Figure 5.1: A mesh constructed for the spatio-temporal expression model of Vineyard 12 with observation locations (red dots) and estimation locations (connected by black edges).

Suppose that ω is a Matérn Gaussian field, given the triangulation presented above, $\omega(s)$, the variable of the spatial unit s is represented as follows:

$$\omega(s) = \sum_{l=1}^{L} \psi_l(s)\lambda_l \tag{5.19}$$

where L is the total number of vertices, $\{\psi_l(s)\}$ are the basis functions and $\{\lambda_l\}$ are Gaussian distributed weights.
The studied domain is a lattice rather regular and this approach allows to approximate the local irregular structures such as the area with missing plants and the edge effects. Moreover, a sparse triangulation can largely reduce the number of the spatial units to estimate but may hide small-scale variation. Even in this way, we lose the accuracy of the model, a compromise must be done between a gain of accuracy and a gain of practicality.

For each vineyard, we used the package R-INLA (*http://www.r-inla.org/*) to construct the meshes and found $m = 500, ..., 700$ nodes to be a reasonable choice depending on the vineyard (see the mesh constructed for Vineyard 12 in Figure 5.1). We added external nodes outside the observation domain to avoid possible boundary effects near the edges. with the choice of 700 nodes, the time spent to run the model including a latent AR and auto-correlated error term was 3 to 4 hours (about 10 hours when the model also including a row-specified random).

5.2.4 Criterion for model selection

The performances of the large range of models were evaluated and were compared using two criteria. The first criteria DIC (deviance information criteria) (Spiegelhalter et al., 2002), frequently used for Bayesian model comparison, was applied to compare the regression models. As it tends to under-penalize the complex models (Plummer, 2008, Riebler and Held, 2009), such as hierarchical models, the models with mixed effects were selected using cross-validatory predictive checks. This criteria appreciates the quantities:

$$CPO_i = \pi(y_i^{obs}|y_{-i}) \tag{5.20}$$

Where y_{-i} denotes the observations y with the $i-$th component omitted. CPO_i (Cross-validatory Predictive Ordinate) expresses the posterior probability of observing the value of y_i when the model is fitted to all data except y_i. A high value means a better fit of the model to y_i, while a low value suggests that y_i is an outlier and an influential observation.
Based on the CPO values, we can calculate the mean of logarithmic score

$$logscore = -\frac{1}{n}\sum_{i}^{n} log(CPO_i) \tag{5.21}$$

A smaller value indicates a better prediction quality of the model.

5.3 Results

In this section, the modelling results were presented according to two variables of responses: the occurrence and first-occurrence of esca foliar symptoms at year t.

For both of them, we began by presenting the temporal evolution of prevalence and incidence of esca, followed by the presentation of the results of different models. At first we examined the large-scale-structure explained by the covariates, and then we studied the temporal dependence by either auto-regressive model or regressions on different pre-defined neighborhood indicators. At last we integrated a row specified random effect into the models to see if the distribution of the disease is better measured by taking into account the heterogeneity of the rows.

Table 5.1 presents the different models tested to fit the data. EXP is the prefix for models concerning the occurrence and FST the one for the first occurrence of foliar symptoms.

Table 5.1: Summaries of models used in this article

response	objective	$logit(c_{i,t})$	abbreviation
Symptoms occurrence Y_{it}	covariates	$X_{itj}\beta_j \quad j = 1, ..., p$	EXP_REG_one
		$\mathbf{X}_{it}\boldsymbol{\beta}$	EXP_REG_all
	Spatio-temp. dependence	$\mathbf{X}_{it}\boldsymbol{\beta} + \omega_{i,t}$	EXP_SR
		$\mathbf{X}_{it}\boldsymbol{\beta} + \rho_1\xi_{i,t-1} + \omega_{i,t}$	EXP_ST_AR1
		$\mathbf{X}_{it}\boldsymbol{\beta} + \rho_1\xi_{it-1} + \rho_2\xi_{it-2} + \omega_{it}$	EXP_ST_AR2
	along-row heterogeinity	$\mathbf{X}_{it}\boldsymbol{\beta} + \xi_{it} + R_{r\|i\in n_r}$	EXP_ST_AR1_row
response	objective	$logit(p_{i,t})$	abbreviation
First symptoms occurrence Z_{it}	covariates	$X_{itj}\beta_j \quad j = 1, ..., p$	FST_REG_one
		$\mathbf{X}_{it}\boldsymbol{\beta}$	FST_REG_all
		$\mathbf{X}_{it}\boldsymbol{\beta} + \omega_{i,t}$	FST_M0
	disease spread	$\mathbf{X}_{it}\boldsymbol{\beta} + \eta PN^0_{i,t} + \omega_{i,t}$	FST_M1
		$\mathbf{X}_{it}\boldsymbol{\beta} + \eta_1 PN^1_{i,t} + \omega_{i,t}$	FST_M2
		$\mathbf{X}_{it}\boldsymbol{\beta} + \eta_1 PN^1_{i,t} + \eta_2 PN^2_{i,t} + \omega_{i,t}$	FST_M3
		$\mathbf{X}_{it}\boldsymbol{\beta} + \eta_r PNr^0_{i,t} + \eta_{or} PNor^0_{i,t} + \omega_{i,t}$	FST_M4
	along-row heterogeinity	$\mathbf{X}_{it}\boldsymbol{\beta} + \omega_{i,t} + R_{r\|i\in n_r}$	FST_M0_row

5.3.1 Temporal evolution of prevalence and incidence of esca within both vineyards

The dynamics of prevalence and incidence in both vineyards greatly differed Figure 5.2. During the recording period, in Vineyard 8, esca prevalence varied from 1.2 % to 14.2 %. It showed a first increase until 2007, then became stable and greatly increased in 2012 with the highest rate. The incidence followed a similar curve with, however, lower percentage varying from 1 % to 4.4 %. In comparison, the vineyard 12 showed higher prevalence, varying from 6.1 to 14.2 % during the period. Two peaks of prevalence were recorded in 2007 and in 2012, contrasting with two decreases of prevalence in 2005 and 2011.

The annual total rainfall estimated in each area of the vineyards showed similar variation trends, with a maximum value reaching 992 mm. The minimal values of 587 and 610 mm of rainfall , respectively in 2005 and 2011 corresponding to the

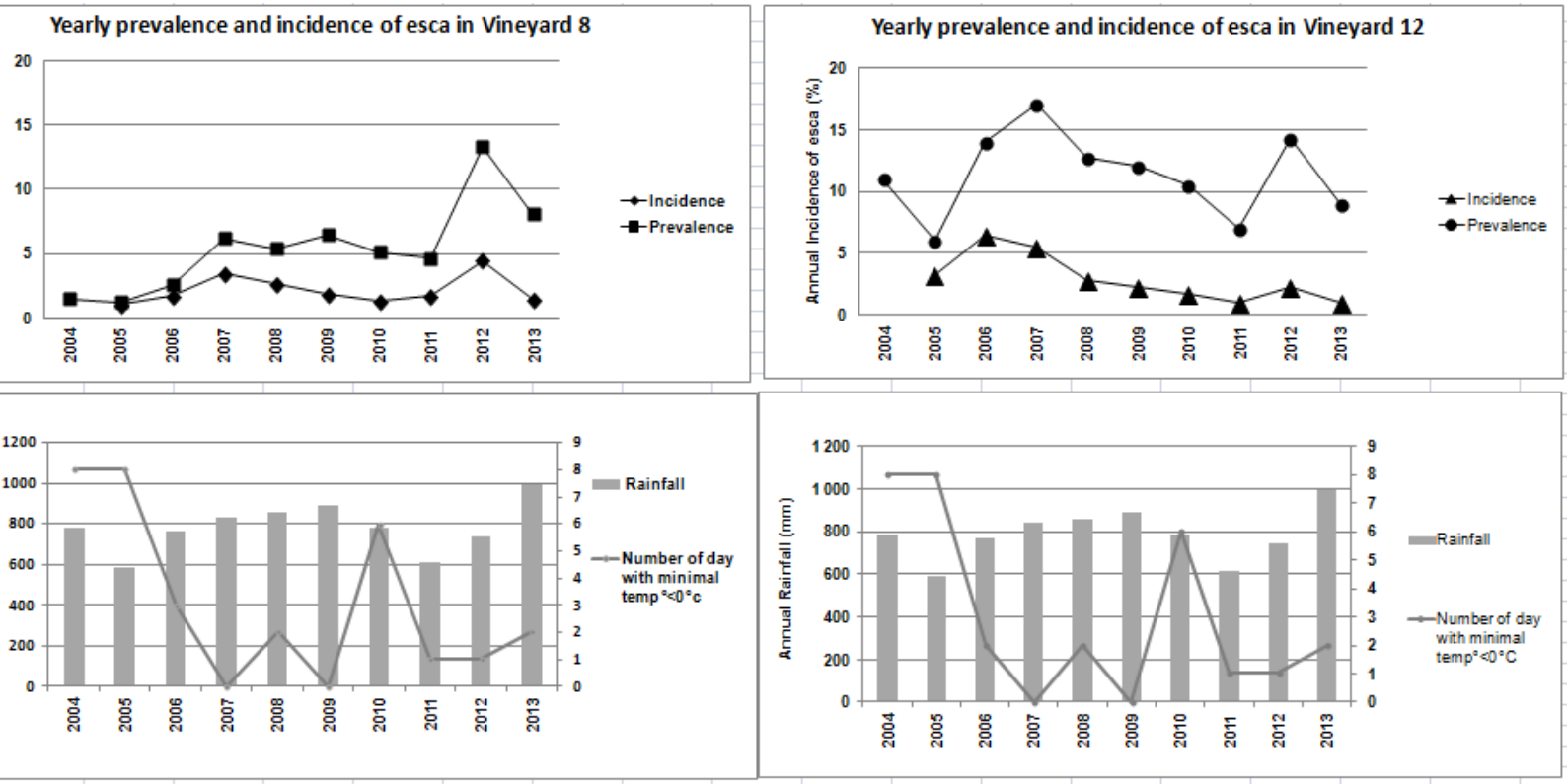


Figure 5.2: Top: annual prevalence and incidence of esca within Vineyards 8 (left) and 12 (right); bottom: annual rainfall and number of day with minimal temperature $\leq 0^oC$ for Vineyards 8 (left) and 12 (right). The climate variations are closed in these two vineyards.

lowest prevalence in these year in Vineyard 12. The number of days with minimal temperature inferior or equal to 0^o varied from 0 to 9, according to the year. A similar trend was shown for the both vineyards. The three years, 2004, 2005 and 2010 corresponded with the highest values.

5.3.2 Spatio-temporal structure explained by the covariates

First of all, we examined the effects of the covariates for both occurrence and first occurrence of foliar symptoms. Results are given in Tables 5.2 and 5.3. The significant effects of each covariate are obtained by examining if the empirical confidence interval covers 0 and their signs are decided by comparing the empirical confidence interval with 0.

5.3.2.1 Effects of covariates on the probability of symptom occurrence for a given year

For Vineyard 12, the three covariates tested, Total Rainfall (RRtotal) , Number of day with minimum temperature inferior to 0^oC (NBprep) and total nitrogen content of must (Nitrogen), had a significant effect for all the models of foliar expression: simple regression with a single covariate (EXP_REG_one), multiple regression with all the covariates (EXP_REG_all), Spatial replicated model with multiple regression (EXP_SR) and a hierarchical regression model with a latent AR(1) process (EXP_ST_AR1).

The covariates NBprep and Nitrogen had significant positives effects, it meant that the probability of esca symptom occurrence increased with the increase of the annual precipitation rate. We found that for the variable nitrogen content of musts (Nitrogen), the risk of esca symptom occurrence increased with the increase of nitrogen content of vines. In contrast with the covariate NBprep, a negative effect was obtained on the probability of esca symptom occurrence: an increase of this covariate is related to a decreased risk of expression of esca. For this vineyard, the influence of these covariates did not change after the integration of a latent spatial replicates process or a latent spatial temporal autoregressive process .

In the case of the Vineyard 8, four covariates (RRtotal, NBprep, Nitrogen and $\delta^{13}C$) were used for modelling, but the number and/or the type of significant covariates varied according to the tested model. Only the covariate relative to the temperature (NBprep) showed a significant effect for all the tested models. As shown for the Vineyard 12, an increase of this indicator was related to a decrease of the risk of esca foliar expression. The covariate RRtotal only revealed a significant positif effect with the mono covariate simple regression model. In opposite, the variable $\delta^{13}C$ showed a negative significant effect when it was associated with the other covariates NBprep and Nitrogen. That meant that a stronger water constraint of vine was associated with a decrease of the risk of esca foliar disease. The multivariate regression model selected three covariates: NBprep, Nitrogen and $\delta^{13}C$.

When the model included a latent spatial process or an spatio-temporal auto regressive process, only the variate NBprep was selected.

The spatio-temporal distribution of esca symptom occurrence in Vineyard 12 can be explained by the environmental factors integrated in the model: RRtotal, NBprep and Nitrogen. While in Vineyard 8, the environmental factors can only

explain the variation of symptom occurrence in the model without spatio-temporal dependence.

Among the models mentioned before to test the effect of factors on the probability of symptom occurrence at one year, for both Vineyard 12 and Vineyard 8, the multiple regression model (REG_all) had better DIC than the simple regression models with only-one covariate (REG_one) Table 5.4. This implies that the expression of the foliar symptoms should be the result of the combination of several environmental factors.

Table 5.2: Effects of covariates on the probability of esca expression. Tested covariates, RRtotal : sum of yearly rainfall, NBprep : number of days with the minimum temperature inferior or equal to 0^oC in spring, Nitrogen: nitrogen content of musts, $\delta^{13}C$. Tested models : one-covariate regression (EXP_REG_one) for each covariate, multi-covariate (EXP_REG_all), covariates & spatial replicate (EXP_SR) , covariates & latent spatio-temporal auto-regressive EXP_ST_AR1. The significance, noticed + or−, was determined by the empirical confidence interval. +: positive significant effect on the probability of symptom occurrence. −: negative significant effect on the probability of symptom occurrence. Ns: non-significant effect of the covariate.

(a) Vineyard 12

$EXP\_$	REG_one	REG_all	SR	ST_AR1
RRtotal	+	+	+	+
NBprep	-	-	-	-
Nitrogen	+	+	+	+
$\delta^{13}C$	//	//	//	//

(b) Vineyard 8

$EXP\_$	REG_one	REG_all	SR	ST_AR1
RRtotal	+	Ns	Ns	Ns
NBprep	-	-	-	-
Nitrogen	+	+	Ns	Ns
$\delta^{13}C$	Ns	-	Ns	Ns

Table 5.3: Effects of covariates on the probability of first esca expression. Tested covariates, RRtotal : sum of yearly precipitations, NBprep : number of days with the minimum temperature inferior or equal to 0^oC in spring, Nitrogen: nitrogen content of musts, $\delta^{13}C$. Tested models : one-covariate regression (FST_REG_one) for each covariate, multi-covariate (FST_REG_all), covariates & spatial auto-correlation (FST_M0). The significance, noticed + or−, was determined by the empirical confidence interval. +: positive significant effect on the probability of symptom occurrence. −: negative significant effect on the probability of symptom occurrence. Ns: non-significant effect of the covariate.

(a) Vineyard 12

$FST\_$	REG_one	REG_all	$M0$
RRtotal	Ns	Ns	Ns
NBprep	-	-	-
Nitrogen	+	+	+
$\delta^{13}C$	//	//	//

(b) Vineyard 8

$FST\_$	REG_one	REG_all	$M0$
RRtotal	+	Ns	Ns
NBprep	-	-	-
Nitrogen	+	+	Ns
$\delta^{13}C$	Ns	-	Ns

5.3.2.2 Effects of covariates on the probability of first symptom occurrence

The results given in Table 5.3 for the first occurrence of symptom were similar to the Table 5.2, those of occurrence of symptom. For the both vineyards, most of the pre-selected environmental factors, the symptom occurrence and first symptom occurrence had the same signs of significance. The only exception was *RRtotal*, the total rainfall of the year, which was significant for expression but showed a non-significant effect on the probability of first symptom occurrence in Vineyard 12. The risk of expressing esca symptom increase with the increase of the total rainfall of a year, while the risk of the first esca symptom occurrence did not seem to be influenced by this covariate.

However, comparing the DIC values among models for the probability of first occurrence of symptom, in both Vineyard 12 and Vineyard 8, the multiple regression model (REG_all) had better DIC than the simple regression models with only-one covariate (REG_one) as shown in Table 5.5. That meant an additive effect of the different environmental factors on the risk of first symptom occurrence and suggested the effect of different environmental factors on the first esca foliar symptom occurrence.

Table 5.4: DIC and CPO (log-score) values for different models of symptoms occurrence, with the posterior estimates (quantiles) of the auto-regressive coefficient vector ρ, ρ_1 et ρ_2 for EXP_ST_AR1 and EXP_ST_AR2 models.

(a) Vineyard 12

Models		DIC	CPO (log-score)	0.5 quantile, [0.025 quant., 0.975 quant.]
EXP_REG_one	RRtotal	12926		
	NBprep	12917		
	Nitrogen	12864		
EXP_REG_all		12813	0.349	
EXP_SR		11967	0.326	
EXP_ST_AR1		11849	0.322	$\rho = 0.943, [0.917, 0.960]$
EXP_ST_AR2		11786	0.320	$\rho_1 = 0.983, [0.977, 0.988]$ $\rho_2 = -0.978, [-0.999, -0.876]$

(b) Vineyard 8

Models		DIC	CPO (log-score)	0.5 quantile, [0.025 quant., 0.975 quant.]
EXP_REG_one	RRtotal	7192		
	NBprep	7058		
	Nitrogen	7199		
	$\delta^{13}C$	7234		
EXP_REG_all		7014	0.200	
EXP_SR		6438	0.183	
EXP_ST_AR1		6229	0.1767	$\rho = 0.992, [0.970, 0.998]$
EXP_ST_AR2		6233	0.17778	$\rho_1 = 0.985, [0.943, 0.996]$ $\rho_2 = 0.406, [0.031, 0.688]$

Table 5.5: DIC and CPO (log-score) values for different models of symptoms first-occurrence. Data used from 2004-2013.

(a) Vineyard 12

models		DIC	CPO (log-score)
FST_REG_one	RRtotal	4795	
	NBprep	4785	
	Nitrogen	4779	
FST_REG_all		4772	0.183
FST_M0		4686.43	0.1801234

(b) Vineyard 8

models		DIC	CPO (log-score)
FST_REG_one	RRtotal	3542	
	NBprep	3500	
	Nitrogen	3526	
	$\delta^{13}C$	3545	
FST_REG_all		3480	0.121
FST_M0		3400.99	0.1180957

5.3.3 Spatio-temporal dependence for occurrence of foliar symptoms

Table 5.4 summarized the DIC-values and CPO (log-score)-values for the various constructed models to compare.

Compared with the multiple regression model (EXP_REG_all) for expression data, the spatial replicate model EXP_SR the spatio-temporal (ST) models with latent auto-regressive process (EXP_ST_AR1, EXP_ST_AR2) had both better(lower) DIC and CPO (log-score) values, which meant the integration of the spatial-only and spatio-temporal auto-correlation largely improved the modelling of the disease distribution. Among them, the ST auto-regressive models (EXP_ST_AR1, EXP_ST_AR2) have better DIC and CPO (log-score) values than SR model EXP_SR, this implied that the spatio-temporal dependence model better fitted the data and consequently better explained the spatio-temporal dynamics of esca.

The high value of the AR(1) temporal coefficient ρ corresponded to a high probability of a vine to express foliar symptom at year t when it had already expressed symptom at year $t-1$. As a comparison, a model with latent AR(2) process was also used to fit the data of the vineyards Vineyard 12 and Vineyard 8. On one hand, the performances of the models were very closed to the model with latent AR(1) process; on the other hand, the value of the second coefficient of AR(2) which measured the temporal dependence on year before the previous year, was much smaller than the value of the first coefficient which measured the temporal dependence of previous year. Moreover, it is very interesting to see that for Vineyard 12, the latent AR(2) process was cyclic because $\rho_1^2 + 4\rho_2 < 0$, and with the values, in our case the average period of cycle was about 4 years, while for Vineyard 8, the latent process was stationary.

Although some attractive properties were observed by the latent AR(2) process, the models were not largely improved. For cyclic behaviors, we need data observed at long term to verify this finding. On short-term, only the strong dependence on the previous year were confirmed by our analysis.

5.3.4 The spread of disease: models for first symptoms occurrence

In order to test the effect of short-scale neighborhood structure on the probability of first esca foliar symptom occurrence, we constructed four models $M1$ to $M4$, with different neighbor indicators integrated in the model. These neighbor indicators varied according to the infected years and locations of the diseased neighboring vines.

First of all, we compared the model performances between FST_REG_all and FST_M0 to see if the integration of spatial auto-correlation among the first symptoms occurrences better fitted the model. The results showed that for both two vineyards, FST_M0 had lower DIC and CPO (log-score) values than FST_REG_all, this implied that the first symptoms occurrences are spatially dependent.

The model $M0$, corresponding to the regression only on the covariates with those integrated a linear link on the previous neighborhood pattern, was compared with the models $M1$ to $M4$, to study the effect of previously diseased vine in the close neighborhood on the probability of the first esca occurrence.

These models were tested in the following orders to answer the epidemiological questions:

- $M0$: A model without regression on neighbouring indicators as a null model to be compared with the others
- $M1\&M0$: Is the risk for a vine to express the esca symptoms for the first time related with the formal neighboring diseased vines which had expressed at least once foliar symptoms?
- $M2\&M0$: Is the risk for a vine to express the esca symptoms for the first time related only with the neighboring vines recently expressed symptoms at year t-1?
- $M1\&M3$: Is the risk for a vine to express the esca symptoms for the first time related with formal neighboring diseased vines. Then is this risk more related with the neighboring vines that express the symptoms at year t-1 than with the neighboring vines that had already expressed at least once symptoms before year t-1?
- $M1\&M4$: Is the risk for a vine to express the esca symptoms for the first time related with formal neighboring diseased vines. Then is this risk more related with the formal neighboring diseased vines located on the same row than the formal neighboring diseased vines off the same row?

Table 5.5 summarized DIC and CPO (log-score) values for the multiple regression models with (FST_M0) or without (FST_REG_all) spatially auto-correlated

residuals. For both Vineyard 12 and Vineyard 8, models FST_M0 have both better (lower) DIC and CPO (log-score) values than FST_REG_all models, consequently the model integrating covariates and auto-correlation better fitted the data, These results suggested a local similarity with new visible diseased vines always occurring together in small cluster.

Table 5.6: DIC and CPO (log-score) values and significance of neighborhood indicators for models $FST\_M0 - -FST\_M4$ of Vineyard 12, with the signs of significance of the neighborhood indicators' coefficients. Data used from 2007-2013.

model	DIC	CPO (log-score)	linked neigbouring patterns	signs of significance
FST_M0	3072	0.1637		
FST_M1	3036	0.1619	$N^0_{i,t}$	$\eta_0 = (-)$
FST_M2	3075	0.1638	$N^1_{i,t}$	$\eta_1 = (+)$
FST_M3	3042	0.1622	$(N^1_{i,t}, N^2_{i,t})$	$(\eta_1, \eta_2) = (0, -)$
FST_M4	3076	0.1638	$(Nr^0_{i,t}, Nor^0_{i,t})$	$(\eta_r, \eta_{or}) = (0, +)$

Table 5.7: DIC and CPO (log-score) values and significance of neighborhood indicators' coefficients for models $FST\_M0 - -FST\_M4$ of Vineyard 8, with the signs of significance of the neighborhood indicators' coefficients.

model	DIC	CPO (log-score)	linked neigbouring patterns	signs of significance
FST_M0	2862.36	0.1337303		
FST_M1	2866.47	0.1339238	$N^0_{i,t}$	$\eta_0 = (0)$
FST_M2	2861.44	0.1337	$N^1_{i,t}$	$\eta_1 = (0)$
FST_M3	2865.89	0.1339072	$(N^1_{i,t}, N^2_{i,t})$	$(\eta_1, \eta_2) = (0, -)$
FST_M4	2861.82	0.1337201	$(Nr^0_{i,t}, Nor^0_{i,t})$	$(\eta_r, \eta_{or}) = (0, +)$

The results of $FST\_M0 - FST\_M4$ were presented in Tables 5.6 and 5.7 for Vineyard 12 and Vineyard 8. A surprising result is that the effects of $PN^0_{i,t}$ and $PN^2_{i,t}$ in $M1$ and $M3$ are negatives. In that case, the probability that the newly expressing vine occur at a available distant non-infected location is higher than close to formal diseased vines. With regards to FST_M1 (regression on recently infected neighbors) and FST_M4 (split of the neighbors into on-the-same-row and off-row subgroups), they have no better DIC and CPO (log-score) values. Moreover, the parameter of on-the-same-row neighbors is not significant while the parameter of off-row neighbors is significant. This off-row effect can be verified with the map of esca vines.

For Vineyard 8, only FST_M2 and FST_M4 have slightly better DIC than FST_M0, however, all the parameters of neighborhood indicators are not significant. These results indicated no strong effect of the past neighborhood on the first occurrence of esca symptoms.

5.3.5 Row random effect

The row effect meant that a spatio-temporal distribution of the disease may be related to the row. The probability of the first symptom and symptom occurrence may increase or decrease according to the row in the vineyard.

As the vines were planted in rows in an vineyard, we may consider a row effect in the disease distribution: the vines planted in a same row may have larger or smaller probability to be infected or to express the symptoms. A row-specified random effect should explain this variability if the row effect occurred.

Table 5.8: Vineyard 12, spatial field parameter estimations of logistic auto-regressive model of order 1 for expression data: EXP_ST_AR1, DIC=11849, CPO (log-score)=0.322

	mean	sd	0.025quant	0.5quant	0.975quant
ρ	0,942	0,011	0,917	0,943	0,961
σ^2_ω	3,889	1,271	2,353	3,607	6,353
γ	2,399	0,506	1,456	2,389	3,409

Table 5.9: Vineyard 12, spatial field parameter estimations of logistic auto-regressive model with row random effect for expression data: $EXP\_ST\_AR1\_row$, DIC=11812, CPO (log-score)=0.3214438

	mean	sd	0.025quant	0.5quant	0.975quant
ρ	0,939	0,012	0,914	0,94	0,959
σ^2_ω	5,851	2,727	2,914	5,122	11,192
σ^2_R	0,061	0,031	0,022	0,054	0,141
γ	1,896	0,479	1,021	1,883	2,860

First, based on the hierarchical regression model with AR(1) model ($EXP_$ST$_$AR1), we integrate the row-specified random effect to see if the disease distribution of along-row planted vines exhibit the among-row heterogeneity. For both vineyards, Vineyard 12 and Vineyard 8, the DIC and CPO (log-score) are better after integrating this random effect (Tables 5.9 and 5.11). Moreover, the estimated spatial range

Table 5.10: Vineyard 8, spatial field parameter estimations (mean, standard error, quantiles) of logistic auto-regressive model of order 1 for occurrence of symptoms: EXP_ST_AR1, DIC=6229, CPO (log-score)=0.1767

	mean	sd	0.025quant	0.5quant	0.975quant
ρ	0,9901	0,0076	0,9702	0,992	0,9981
σ^2_ω	16,9718	11,2952	5,2377	13,925	38,8781
γ	42,9922	14,1685	22,1022	40,5639	77,3553

Table 5.11: Vineyard 8, spatial field parameter estimations of logistic auto-regressive model with row random effect for expression data: $EXP\_ST\_AR1\_row$, DIC=6201, CPO (log-score)=0.177

	mean	sd	0.025quant	0.5quant	0.975quant
ρ	0,989	0,008	0,969	0,991	0,998
σ^2_ω	14,853	9,396	4,831	12,397	33,088
σ^2_R	0,109	0,054	0,036	0,099	0,243
γ	42,885	13,69	22,417	40,637	75,836

(γ, the distance limit between vines to have spatial dependencies) of Vineyard 12 and Vineyard 8 have been reduces after adding the random effect (Tables 5.8, 5.9). However, the σ^2_ω increases after integrating the row random effect. Let us also note that the ratio between σ^2_R and σ^2_ω is so weak, this led to a negligible effect of the row compared with the one of ω (Tables 5.9 and 5.11).

Table 5.12: Estimations of spatial field parameters of model FST_M0 with row-specified random effect: FST_M0_row for Vineyard 12, DIC=4686.27, CPO (log-score)=0.1801168

	mean	sd	0.025quant	0.5quant	0.975quant
σ^2_ω	0,3895	0,1128	0,2337	0,3735	0,5988
σ^2_R	0,0001	0,00016	1,4396e-05	7,4626e-05	0,0006
γ	10,5083	3,7498	4,9256	9,9011	19,5300

Table 5.13: Estimations of spatial field parameters of model FST_M0 with row-specified random effect: FST_M0_row for Vineyard 8, DIC=3400.89, CPO (log-score)=0.118092

	mean	sd	0.025quant	0.5quant	0.975quant
σ^2_ω	0,5728	0,2407	0,2714	0,5253	1,0329
σ^2_R	0,0001	0,0001	1,5e-05	7,5e-05	0,0005
γ	19,5875	8,6446	7,9616	17,7752	41,4592

Next for first symptom occurrence data, the model FST_M0 with row specified random effect neither improved the model, nor contributed notable variance compared to the σ^2_ω (Tables 5.12 and 5.13).

5.3.6 Concluding models

After analyzing and comparing all of the models, the models that best fitted the data for each vineyard were selected. Tables 5.14 and 5.15 give the estimated values of the parameter and their quantiles.

It has to be noted that due to short time series of data, the model EXP_ST_AR2, of which the latent process is characterized by a AR(2) process, should be interpreted with caution.

Excluding that model, for both vineyards, the best model for occurrence of symptom is EXP_ST_AR1: a logistic regression model with a latent autoregressive process characterized by a Gaussian Markov random field (GMRF). With respect to first-occurrence data, the best model for both vineyards is FST_M0 if we don't consider the models with regression on neighborhood indicators: it is a logistic regression model with a spatial auto-correlated term characterized by a GMRF.

Model EXP_ST_AR1 fitted better the symptom occurrence in Vineyard 12 than in Vineyard 8: all the covariates are significant and the estimation of spatial field, $\{\sigma^2_\omega, \gamma\}$ had reasonable estimated values. While in Vineyard 8, only one covariate is significant and the estimation of σ^2_ω is very high, as well as the estimation of γ.

Model FST_M0 fitted the first symptom occurrence in both vineyards, as mentioned before, a few co-variates are significant in such model. However, the spatial field is properly estimated: first symptom occurrence in Vineyard 8 and Vineyard 12 was characterized by a large-scale spatial structure with $\gamma > 10$.

Table 5.14: Vineyard 12, parameter estimations (mean, standard error, quantiles) of logistic auto-regressive model of order 1 for occurrence of symptoms: EXP_ST_AR1

	mean	sd	0.025quant	0.5quant	0.975quant
Intercept	-2,425	0,088	-2,602	-2,424	-2,256
RRtotal	0,123	0,031	0,061	0,123	0,185
NBprep	-0,172	0,032	-0,236	-0,172	-0,109
Nitrogen	0,304	0,079	0,150	0,304	0,460
ρ	0,942	0,011	0,917	0,943	0,961
σ^2_ω	3,889	1,271	2,353	3,607	6,353
γ	2,399	0,506	1,456	2,389	3,409

Table 5.15: Vineyard 12, parameter estimations (mean, standard error, quantiles) of first occurrence symptoms model FST_M0, logistic regression model with a spatial auto-correlated term characterized by a GMRF

	mean	sd	0.025quant	0.5quant	0.975quant
Intercept	-3,1784	0,1275	-3,4427	-3,1757	-2,929
RRtotal	-0,0455	0,0516	-0,1467	-0,0456	0,0557
NBprep	-0,2096	0,0603	-0,3286	-0,2093	-0,0918
Nitrogen	0,1695	0,086	-0,0022	0,1702	0,337
σ^2_ω	0,3911	0,1133	0,234	0,3753	0,6012
γ	10,5061	3,7771	4,9051	9,8862	19,6147

Table 5.16: Vineyard 8, parameter estimations (mean, standard error, quantiles) of logistic auto-regressive model of order 1 for occurrence of symptoms: EXP_ST_AR1

	mean	sd	0.025quant	0.5quant	0.975quant
Intercept	-3,89	2,9266	-10,2399	-3,8482	2,3431
RRtotal	0,0804	0,1011	-0,1201	0,0798	0,2838
NBprep	-0,3105	0,1166	-0,531	-0,3144	-0,0673
Nitrogen	-0,0321	0,1546	-0,3342	-0,0329	0,2739
$\delta^{13}C$	-0,0167	0,2505	-0,509	-0,0171	0,4773
ρ	0,9901	0,0076	0,9702	0,992	0,9981
σ^2_ω	16,9718	11,2952	5,2377	13,925	38,8781
γ	42,9922	14,1685	22,1022	40,5639	77,3553

Table 5.17: Vineyard 8, parameter estimations (mean, standard error, quantiles) of first occurrence symptoms model FST_M0, logistic regression model with a spatial auto-correlated term characterized by a GMRF

	mean	sd	0.025quant	0.5quant	0.975quant
Intercept	-11,6499	7,6678	-26,6399	-11,6987	3,6339
RRtotal	-0,1128	0,0627	-0,2353	-0,1131	0,0108
NBprep	-0,3559	0,1059	-0,5695	-0,354	-0,1534
Nitrogen	0,0068	0,0054	-0,004	0,0069	0,0171
$\delta^{13}C$	-0,2645	0,2786	-0,8111	-0,2656	0,289
σ^2_ω	0,5191	0,2118	0,2508	0,4787	0,923
γ	17,1636	7,8944	6,6056	15,5024	37,1647

5.4 Discussion

In this study, we showed that the application of logistic regression modelling with a latent spatial or spatial-temporal auto-correlated process is a useful tool for quantifying temporal and spatial patterns of the partially-known esca disease. Several competing hypotheses have been tested, in particular about the role of the environmental factors in the probability of occurrence of esca foliar symptoms. We also tested the spatio-temporal dependence on this probability and the role of neighborhood of infected vines in the risk of first esca occurrence.

In the following sections, we analyze all the models together in order to capture the effects of influential covariates on the esca expression or first expression.

5.4.1 Environmental selected factors in modelling of esca foliar symptom occurrence

The model that fitted the best the data included environmental factors and a latent process of spatio-temporal dependence, characterized by a temporal auto-regressive model with spatially auto-correlated residuals. Among the two environmental selected factors, the total rainfall of a year (RRtotal) partly explained the occurrence of esca in one of the studied vineyards (Vineyard 12). These results corroborated those of Marchi et al. (2006). They showed that the incidence of esca expression in a given year was related to the amount of precipitation in the spring and summer period. The fungal species responsible of internal necroses can produce phytotoxic compounds that are supposed to cause visible foliar symptoms (Bruno and Sparapano, 2006b). Based on our results, probability of foliar symptom occurrence increases with rainfall of one year might due to more active sap circulation favoring the transport and the accumulation of the phytotoxins in the leaves, as suggested by Marchi et al. (2006).

For Vineyard 8, the multiple regression model did not select the covariate RRtotal. This covariate has been selected on the basis of correlation study using data from different vineyards in Bordeaux region. For Vineyard 8, the best correlation was found with the total of rainfall during spring and the beginning of summer (April, May, June, July) (results not shown).

Moreover, among the plant parameters selected by the multiple regression model, $\delta^{13}C$, which describes the level of plant water constraint during the maturity stage (in Summer), was negatively significant. This means that a decrease of plant water constraint increase the probability of esca foliar occurrence. The decrease of plant water constraint is related to soil factors and climatic factors including rainfall. The selection of $\delta^{13}C$ factor was consistent with the selection of the total rainfall for other vineyards. However, we need to keep in mind that the indicator $\delta^{13}C$ is a global indicator of vine water stress during the ripening period, and consequently it does not account for the previous moisture conditions during spring. It was chosen for its facility of implementation and its low cost to interpolate vine hydric status at vineyard scale. It could be possible to measure this indicator at different period, for instance, to relate the water status in Spring or in early Summer to the risk of esca occurrence. Other indirect method could also used biophysical models (Lebon et al., 2003), requiring the spatialization of soil water balance (see chapter 4).

The other significant factor selected in multiple regression for the both vineyards was the total nitrogen content of must. An increase of nitrogen in must was associated with an increase of the probability of esca foliar occurrence. The total nitrogen content of must gives information about the nitrogen status of plant, consequently, about the level of plant vigor (Van Leeuwen et al., 2011). It suggested that the area in vineyard with vigorous vines may have high probability to express the esca symptoms. This inference corroborated to the preliminary study that compared different vineyards (Destrac-Irvine et al., 2007) with different vigor status. This is the first time that the hypothesis is verified by statistical analysis at a intra plot scale. Vine nitrogen status, as vine water status, are soil-dependent. They varied with soil parameters such as soil structure, soil organic matter, organic matter C/N and organic mater mineralization speed. These variation can occur over a short distance as a few meters. Consequently, the heterogeneity at intra vineyard may be observed.

The climatic factor NBprep was the only covariate selected in the model with a latent spatio-temporal auto-correlated process, which fitted the best the data. A decrease of the number of days with minimum temperature inferior or equal to 0^oC in Spring seems to be related to an increase of esca occurrence. We should take this result with caution as the number of day only varied from 0 to 9 over a period of three months and lead questions on the biological significance of this effect. This parameter could be the effect of climatic conditions over Spring which may influence esca occurrence en Summer time. Further study is required to better determine the relation between with Spring climatic conditions and esca.

About selected factors related to environment or plant, similar results were found for the first esca foliar symptom occurrence in Vineyard 8 as for the esca foliar symptom occurrence. By contrast, for Vineyard 12, the covariate RRtotal was not selected by models for first esca foliar symptom occurrence while the covariates NBprep and Nitrogen were selected.

5.4.2 Selected models for the probability of esca foliar symptom

In both plots, the models fitted best the spatio-temporal data were similar: the logistic regression model with a latent spatio-temporal dependent process, characterized by an autoregressive process with auto-correlated residuals: EXP_ST_AR1 model for Vineyard 8 and EXP_ST_AR2 (Vineyard 12) that integrated respectively, one or three environmental covariates.

However, as mentioned before, we have to interpret the model EXP_ST_AR2 with caution due to the lack of time series. To get more faithful results, the final model we chose for Vineyard 12 is the model with AR(1) process: EXP_ST_AR1.

Comparing the modelling results of two studied plots, we observed two types of disease structure: the estimation of γ, which measured the maximum distance of spatial dependence between vines, were 42.9 m (1.9, respectively) for Vineyard 8 (Vineyard 12). In the case of Vineyard 8: this distance can reflect a spatial structure of esca vines very sparse. As the annual prevalence of esca in Vineyard 8 at the beginning survey is very low (Figure 5.2), corresponds a early stage of the epidemic, the estimated sparse structure indicates a random spatial structure of esca

at the beginning of the epidemic. In contrast, in Vineyard 12, the estimation of γ is much shorter (1.9 m) than in Vineyard 8, reflects a small-scale spatial structure, the cases of esca are aggregated at small distance. Moreover, the rate of prevalence at the beginning of the survey is higher too. These results corroborated those of Zanzotto et al. (2013) who analyzed the spatio-temporal dynamic of esca from the beginning of the epidemic: over the first 7 years of survey, the disease seems to be randomly distributed and then aggregated.

The AR(1) coefficient ρ measured the temporal dependence of status of esca (symptomatic or not) at year t on year $t-1$. The estimations of ρ are very closed to 1 for both two plots, this shows that the probability of esca expression at year t strongly depend on the vine status at the year $t-1$. Once a vine express the foliar symptom at one given year, the probability to re-express at the following year is very high. By applying logistic modelling to the temporal data of transition status of esca disease, Guérin-Dubrana et al. (2013) also found that the probability to re express disease symptom was high.

From the estimations of AR(2) coefficients in the model EXP_ST_AR2, a cyclic behavior about 4 years was found for Vineyard 12. The curve of annual prevalence of esca in Vineyard 12 may also reflect this cyclicity: 3 pics could be found at 2004, 20007 and 2012 in Figure 5.2. Although this result need further studies on the data with more years to ascertain, it is still meaningful to help the scientists to relate the disease with some cyclic factors.

5.4.3 Selected models for the probability of first esca foliar symptom

For first esca foliar symptom occurrence, the models fitted best the data for two vineyards are different.

For Vineyard 12, the model fitted best the data is FST_M1, it includes the regression not only on the environmental covariates but also on the neighbor indicator which counts a local neighborhood of vines expressed symptoms at least one time before($PN^0_{i,t}$). This model slightly outperformed the model $M0$ without the neighbor indicator. In this model, the neighbor indicator plays a significant negative effect on the the probability of the first occurrence. Such negative effect is also observed in model FST_M3 for the neighbor indicator who counts the neighboring vines expressed symptoms for more than two years ($PN^2_{i,t}$).

These negative effects may be explained by the fact that the vine located near the diseased vines is less vulnerable than the vine located near the asymptomatic vines. Another possible reason that why the two indicators $PN^0_{i,t}$ and $PN^2_{i,t}$ could more likely have negative effects for the first occurrence of symptom may due to a statistical problem. In fact, these two indicators, calculate the proportions of previously diseased vines in the neighborhood, are increasing functions with respect to t. Then a positive effect of such indicators corresponds to a increasing influence of previously diseases neighbors on the probability of first occurrences, this implies that a marginal expectation of Z_{it} would be monotonous at the end. Obviously, it is not the case for esca, of which the annual prevalence always has the non-monotonous and strong variation.

Moreover, the model M4 performed the regression on the neighbor indicators

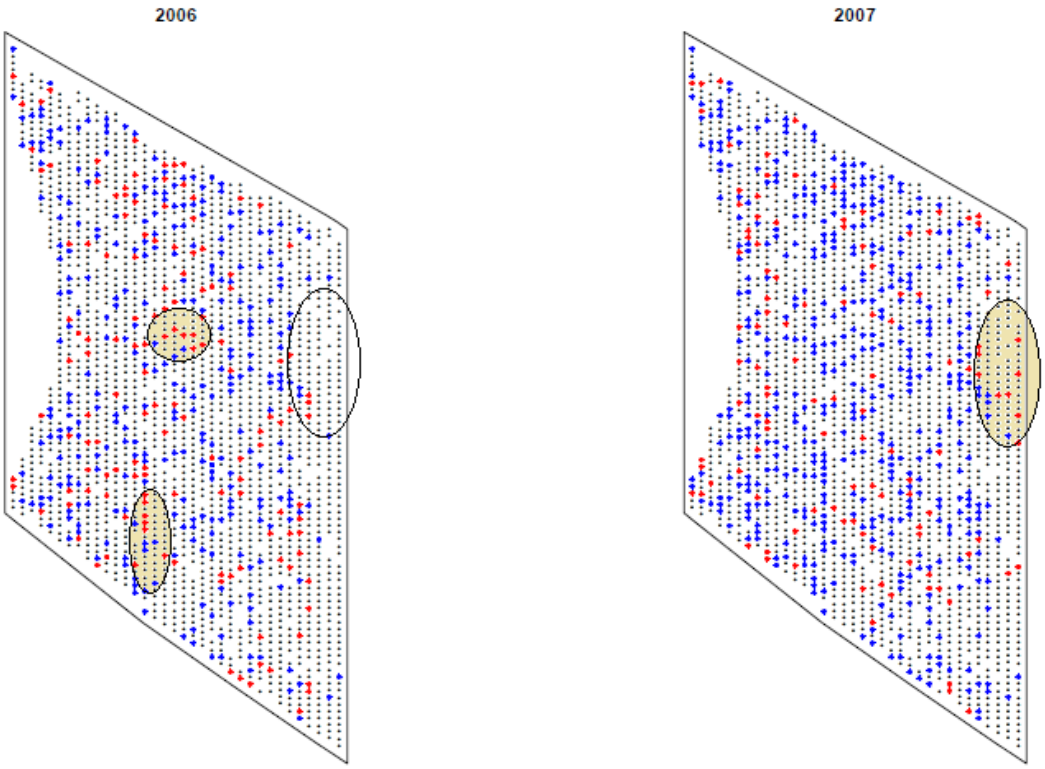


Figure 5.3: A disease map of Vineyard 12 at year 2006 and 2007. The red points represents the vines with first occurrences of esca and the blue points represents the vines previously express the symptoms. We see that the expression of symptoms tends to occur at a available distant non-infected location is higher than close to ancient diseased vines in yellow zones.

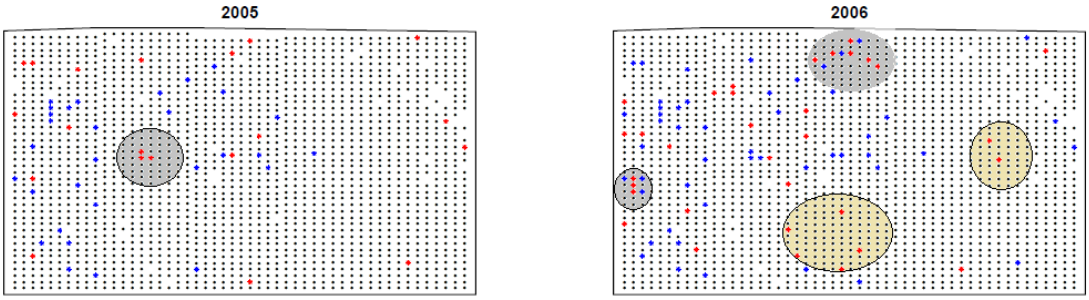


Figure 5.4: A disease map of Vineyard 12 at year 2005 and 2006. The red points represents the vines with first occurrences of esca and the blue points represents the vines previously express the symptoms. We see that the expression of symptoms tends to occur at a available distant non-infected location is higher than close to ancient diseased vines in yellow zones and in zones grey, the red points occurred besides the blue points in other rows rather in the same row.

which divided the previously neighboring diseased into on-same-row and off-row groups. The estimations showed no significant effect of the neighbors previously symptomatic vines situated on the row and an positive significant effect of off-row. These effects could be visually verified directly on the map Figure 5.4 and also be confirmed by the exploratory analysis that we performed in Chapter 3. These results are opposite to Zanzotto et al. (2013).

For Vineyard 8, as the performances of models FST_M0 to $M4$ with different predefined neighbor indicators are very closed and none of the indicators was significant in the model, we finally retained the FST_M0 model: a logistic regression model with environmental factors and a auto-correlated residuals, without considering the neighbor indicators.

We conclude that the esca spatio-temporal development should be dominated by the environmental factors and may be influenced by neighboring previously-infected vines.

5.4.4 Statistical point of view

Despite very helpful modelling results to better understand the dynamics of esca, some methodological limits need to be pointed out.

As the records of the disease only started in 2004, the data are truncated leading to a discrepancy between observation data and true first symptom expression, mainly for the first five years. In further study, uncertainly should be considered in modelling.

In Vineyard 8, the spatial covariates are significant in multiple regression models of symptom occurrence, but they are no longer significant in the models integrating the spatial auto-correlation. A possible reason of this is that as the spatial covariates (Nitrogen and $\delta^{13}C$) at vine scale are estimated using kriging models, the estimations are themselves auto-correlated. This auto-correlation of the covariates could hide

the autocorrelation of the symptom occurrence data. Further work should be done to measure the influence of using a kriging result as a covariate in a auto-correlated model.

As we have already mentioned before, the non-increasing indicators $PN^0_{i,t}$ and $PN^2_{i,t}$ integrated in the model may cause estimation and interpretation problems for spatio-temporal structures, particularly temporal, predominated by large-scale trend. In Chapter 6, we will present the centered parametrization on auto-regressive neighbor terms to resolve this problem. As the estimation of this centered parametrization can be very complex and its theory haven't been verified yet, we did not adopt this parametrization in this chapter.

At last, we modeled the simultaneous dependence at year t between occurrence of first expression by using a approximate GMRF. An alternative way to model this dependence is to achieve the regression on the current neighbors (Z_t, Y_t) directly. We will introduce such models in Chap 6.

5.4.5 Agronomical point of view

Results from hierarchical logistic regression models applied to spatio-temporal data confirmed the results of (Li et al., 2015) (Chapter 3) on the low capacity of local short spread from symptomatic vine. The temporal and spatial distribution is rather driven by environmental factors and probably other factors related to the availability inoculum not studied and difficult to achieve.

Our results are quite consistent with those of Surico et al. (2000b), Cortesi et al. (2000). They did not retain the ability of the disease to spread along the row on the basis of spatial patterns and/or genetic and epidemiological study. Our results, as their ones, did not corroborate those of Zanzotto et al. (2013) and Stefanini et al. (2000).

If these first results are confirmed, they open outlooks for risk prediction of esca at vineyard scale. Available information on the heterogeneity of the vine nitrogen status as well as vine water status may be used to determine risk zone for esca and to manage vine to avoid or decrease esca disease development in the context of precision agriculture. However, further studies, using additional data from different vineyards need to be conducted to confirm these results. Moreover, climatic parameter integrated pluriannuel data may be deepen to select any ones that better explain the temporal development of esca. For instance, the role of the rainfall acting either on the pathogenic fungi, on the host physiology and their interactions need to be better explored.

Chapter 6

Une nouvelle approche pour modéliser la dynamique spatio-temporelle de l'esca de la vigne : modèle de régression auto-logistique spatio-temporel deux fois centré

Introduction

Dans ce chapitre, nous développons un nouveau modèle autologistique pour l'analyse des données spatiales avec covariables. Dans le chapitre 5 nous avons utilisé des modèles logistiques pour la modélisation de variables binaires (expression ou première expression des symptômes) sur un réseau formant ainsi un champ spatial. Ces modèles traitaient de la probabilité de réaliser l'évènement en un site i au temps t en fonction de covariables "exogènes" mais aussi de covariables dépendant de la valeur du champ les années précédentes. Un terme de bruit permettait de modéliser l'autocorrélation entre les sites. Le terme de régression sur les valeurs du champ avant t est vu de manière "causale" en quelque sorte. Les modèles autologistiques diffèrent des modèles précédents en ce sens qu'il décrivent la distribution de la variable binaire sur tout le champ par les distributions conditionnelles de réaliser l'évènement en un point i à l'instant t en fonction des valeurs du champs au voisinage à l'instant t (Equation 6.1), c'est la définition d'un auto modèle. Bien sûr, il faut des conditions pour que ces lois conditionnelles se recollent en une unique distribution jointe. Les modèles autologistiques et leurs généralisations sont adaptés à l'étude de données environnementales (Gumpertz et al., 1997, Huffer and Wu, 1998) et par conséquent aux données de maladie de l'esca. C'est d'ailleurs un modèle autologistique spatio-temporel que Zanzotto et al. (2013) utilisent pour étudier la dynamique de l'esca de la vigne. Notre étude poussée de la littérature des modèles logistiques (Caragea and Kaiser, 2009, Wang and Zheng, 2013) nous a amené à considérer la difficulté d'interprétation d'un terme de régression sur une variable qui croît au cours du temps et de la nécessité de centrer ce terme de régression comme le propose (Caragea and Kaiser, 2009). Cette opération de centrage est motivée par l'interprétation des coefficients du modèle que nous voulons pertinente. De plus, si on régresse à la fois sur la valeur du voisinage au temps t et au temps $t-1$, il est alors nécessaire de centrer deux fois le modèle. Dans ce chapitre, nous justifions et présentons un nouveau modèle autologistique doublement centré. Ce modèle est bien sûr destiné à étudier la distribution des cas de symptômes d'esca dans une parcelle. Cependant, le chapitre reste introductif sur la méthode car il faut encore développer des méthodes d'estimation de ce modèle adapté à la taille des parcelles.

Résumé

Le modèle auto-logistique est un modèle de champ de Markov aléatoire pour des données spatiales binaires. Il a été introduit par Besag (1974). Il peut prendre en compte à la fois la dépendance statistique entre les données et l'effet de covariables potentielles et il est très adapté pour modéliser diverses données écologiques comme la présence ou l'absence d'une certaine maladie de la plante sur un réseau. Dans le cadre spatial uniquement, Caragea and Kaiser (2009) ont proposé un modèle auto-logistique dépendant uniquement des voisinages, puis ils ont développé une paramétrisation centrée avec régression sur les covariables pour ce type de modèle, dans le but de surmonter les difficultés d'interprétation des paramètres sous différents niveaux de dépendance spatiale.

Au cours des dix dernières années, de tels modèles ont été étendus pour prendre en compte la dimension temporelle, puisque les analyses spatio-temporelles ont

montré leur utilité et gagné en popularité. En se basant sur les spécifications spatio-temporelles du modèle auto-logistique généralisé par Zhu et al. (2008), Zheng and Zhu (2008), Wang and Zheng (2013) ont développé un modèle spatio-temporel centré pour analyser un réseau binaire de données temporelles.

Cependant, la spécification temporelle de ce modèle auto-logistique dépend non seulement du passé mais aussi du futur. Cette spécification n'est pas adaptée à la modélisation de maladie de la plante puisque les principaux défis de la modélisation de maladie sont de prédire et comprendre la maladie plutôt que de lisser l'évolution temporelle.

Dans notre étude, nous nous intéressons au modèle spatio-temporel auto-logistique qui ne dépend que du passé et nous en proposons une paramétrisation centrée en deux étapes. Les études de simulations montrent que le modèle à une étape ne peut pas reproduire la structure temporelle des données quand les dépendances spatiale et temporelle sont toutes les deux importantes, tandis que le modèle à deux étapes est en accord avec la structure des données et la structure temporelle à grande échelle. Les résultats des estimations pour des réseaux simulés sur plusieurs années ont été obtenus par l'algorithme espérance-maximisation pour la pseudo-vraisemblance, en deux étapes. Les résultats montrent que sous la paramétrisation centrée en deux étapes, l'estimation des paramètres temporels (ρ_2) et spatiaux (ρ_1) par régression sont précises, tandis que sous la paramétrisation centrée en une étape, les estimations de ρ_1 et ρ_2 sont toujours contradictoires.

Two-step centered spatio-temporal auto-logistic regression model

Shuxian Li

Abstract

The auto-logistic model is a Markov random field model for spatial binary data. It was introduced by Besag (1974) who gave some conditions to define the model through its conditional marginals. It can account for both statistical dependence among the data and for the effects of potential covariates and so is very suitable to model various ecological data such as presence-absence of a certain plant disease observed on a lattice. In the spatial framework, Caragea and Kaiser (2009) developed a centered parametrization for such model with regression on covariates, in order to overcome the interpretation difficulties of parameters across varying level of spatial dependance.

In past ten years, such models were extended to integrate the temporal dimension since the spatio-temporal (ST) analysis. ST models have shown their ability to model the data from environment and they drawn rapidly increasing attention. Based on the ST specification for auto-logistic model generalized by Zhu et al. (2008), Zheng and Zhu (2008), Wang and Zheng (2013) developed a centered spatio-temporal model for analyzing the binary lattice data over time.

However, the temporal specification of this centered ST auto-logistic model depends not only on the past but also on the future. This specification is not adapted for the plant disease modelling as the main challenge of disease modelling is to predict and understand the disease rather than smoothing the temporal evolution.

In our study, we focus on ST auto-logistic model that only depends on the past and proposed a two-step-centered parametrization version of it. The simulation study showed that the one-step model can not reflect the temporal data structure when both spatial and temporal dependance are strong, while for the two-step model, there is an adequate agreement between the data structure and the temporal large-scale structure. The results of estimation for simulated lattices over years were performed by expectation-maximization(EM) pseudo-likelihood in two stages. They showed that under the two-step centered parametrization, the inference for parameters of both temporal (ρ_2) and spatial (ρ_1) regressions are accurate, while under one-step centered parametrization the inference of ρ_1 and ρ_2 are always conflicting.

Additional keywords: spatial-temporal modelling; large-scale, small-scale model structure; binary response

6.1 Introduction

Since spatial and spatio-temporal binary data are commonly existing in nature, the models on such data had drawn large interests of scientists from various fields such as ecology, epidemiology and image analysis, during past years.

40 years ago, Besag (1974) firstly proposed an auto-logistic model for spatial binary data, assuming a simple dependence on surrounding neighbors. This model was proved a very useful model and then was extended in order to integrate the regression on the covariates (Gumpertz et al., 1997, Huffer and Wu, 1998).

Recently, Zhu et al. (2008) and Zheng and Zhu (2008) generalized the auto-logistic regression models to account for covariates, spatial dependence, and temporal dependence simultaneously for binary data that are measured repeatedly over time on a spatial lattice.

However, the non-centered parametrization of the auto-logistic regression models present parameter interpretation difficulties across varying levels of statistical dependence. This problem has been first pointed out by Caragea and Kaiser (2009) who proposed a centered parametrization for spatial auto-logistic regression model to overcome this difficulty. Hughes et al. (2011) further discussed the estimations and simulations of the auto-logistic model on lattice under the centered parametrization. Wang and Zheng (2013) developed a centered spatio-temporal auto-logistic regression model depending on the past and future and compared several estimation methods.

In this paper, we propose a two-step centered auto-logistic model which depends only on the past and will show its advantage over the existing centered parametrization for the auto-logistic models only depending on the past. Since the joint distribution of such models are very complex, we present the expectation-maximization pseudo-likelihood estimation in two steps. The estimation is performed on a 20×20 simulated lattice and we find that this method is fairly accurate.

The paper is organized as follows: first of all, we will present the main researches in the history for the auto-logistic model. After that, in Section 2 we will present the two-step centered auto-logistic model. Then we show several simulations to compare the spatio-temporal auto-logistic model depending on past under one-step and two-step centered parametrization in Section 3. Afterward we present the algorithm of maximization pseudo-likelihood estimation in two steps and show the estimations results on a simulated lattice in Section 4. At last Section, we give perspectives of this study.

6.1.1 Introduction to spatial auto-logistic model

Let us use $[Z]$ to denote the distribution of random variable Z. Let $\mathbf{Z}$ be the random field of interest, $\{Z_i : i = 1, ..., n\}$ represents the binary variables on the lattice. The distribution $[Z]$ is given by:

$$[Z_i|Z_j, j \neq i] \sim Binary(p_i),$$

where $p_i = \mathbb{P}(Z_i = 1|Z_j, j \neq i)$ and $i = 1, ..., n$.

In addition, we assume that corresponding to each location i is a set of other locations N_i considered to be neighbours of i. A Markov random field then results from assuming that

$$[Z_i|Z_j, i \neq j] = [Z_i|Z_j, j \in N_i] \quad \text{for } i = 1, ..., n \tag{6.1}$$

If spatial covariate $\mathbf{X} = \{\mathbf{X}_i, i = 1, ..., n\}$ are also considered, the conditional binary probability can be expressed in exponential family form:

$$\mathbb{P}(Z_i|Z_j, j \in N_i) = \frac{exp(Z_i A_i\{Z(N_i), \mathbf{X}_i\})}{1 + exp(A_i\{Z(N_i), \mathbf{X}_i\})}$$
$$= exp(Z_i A_i\{Z(N_i), \mathbf{X}_i\} - B_i\{Z(N_i), \mathbf{X}_i\})$$

where A_i is called a natural parameter function and B_i is a function of A_i, that is equal to $log(1 + exp(A_i\{Z(N_i), \mathbf{X}_i\}))$ for the auto-logistic model.

Besag (1974) showed that the natural parameter functions for this formulation must be of form:

$$A_i = \alpha_i + \sum_{j \in N_i} \eta_{ij} Z_j \tag{6.2}$$

with α_i a leading constant and η_{ij} being statistical dependence parameters that must satisfy certain restrictions for a joint distribution to exist.

For the traditional spatial auto-logistic model with the covariates, the natural parameter function is presented by (Caragea and Kaiser, 2009):

$$A_i\{Z(N_i), \mathbf{X}_i\} = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j \in N_i} \eta_{ij} Z_j \tag{6.3}$$

Caragea and Kaiser (2009) discussed the interpretation difficulties for traditional auto-logistic model Equation 6.3:
Let $p_i = \mathbb{P}(Z_i = 1|Z_j, j \in N_i, \mathbf{X}_i)$ so that the odds that $Z_i = 1$ in model Equation 6.3 is $p_i/(1-p_i)$. Let us denote $c_i = \frac{exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1+exp(\mathbf{X}_i^T \boldsymbol{\beta})}$ being the probability of occurrence under the spatial independence model, the odds that $Z_i = 1$ is $c_i/(1 - c_i)$. Then the log odds ratio for model Equation 6.3 relative to the independence model is:

$$log[\frac{p_i/(1-p_i)}{c_i/(1-c_i)}] = \sum_{j \in N_i} \eta_{ij} Z_j$$

This implies that the odds of $Z_i = 1$ in model Equation 6.3 relative to the independence model increases for any nonzero neighbors, and can never decrease. This is not reasonable if most of neighbors are zeros and could bias the realizations towards 1.

To overcome the interpretation difficulties, a centered spatial auto-logistic model to better interpretation was developed by (Caragea and Kaiser, 2009).

$$log(\frac{\mathbb{P}(Z_i = 1|Z_j, j \in N_i, \mathbf{X}_i)}{\mathbb{P}(Z_i = 0|Z_j, j \in N_i, \mathbf{X}_i)}) = \mathbf{X}_i^T \boldsymbol{\beta} + \rho \sum_{j \in N_i} (Z_j - logit^{-1}(\mathbf{X}_j^T \beta)) \tag{6.4}$$

They pointed out that model Equation 6.4 is similar to the parametrization customarily used for auto-Gaussian models. This parametrization consider the overall level of a process, possibly adjusted by influence of covariates, to be appropriately modeled as what is called the large-model component, while variances, covariances,

and other high-order portions of the data structure are accounted for by what is called the small-scale model component. (Cressie (1993), p.114). In addition, the parameters were estimated by maximum pseudo-likelihood.

Hughes et al. (2011) focused on the methods about the estimation of the centered auto-logistic model. They studied maximization pseudo-likelihood (PL) followed by parametric bootstrap, Monte Carlo maximum likelihood (ML) and MCMC Bayesian approaches to inference and describe ways to optimize the efficiency of these algorithms. They also compared the performance of the three approaches in a thorough simulation study and found that pseudo-likelihood inference, which is far easier to understand and to implement than the MCML and Bayesian approaches, is both statistically and computationally efficient for data sets that are large enough to allow valid inference.

6.1.2 Introduction to spatio-temporal auto-logistic model

Now let us use $\mathbf{Z}_t$ to denote a random field at year t, $\{Z_{it}, i = 1, ..., n, t \in \mathbb{Z}\}$ represent the random binary variable on site $s_i = (u_i, v_i), i = 1, ..., n$ at year $t, t \in \mathbb{Z}$.

Zhu et al. (2005) generalized the auto-logistic regression models to account for covariates, spatial dependence, and temporal dependence simultaneously. The model specifies the joint distribution of $\{\mathbf{Z}_t : t \in \mathbb{Z}\}$ by a family of conditional distributions:

$$\begin{aligned}&\mathbb{P}(\mathbf{Z}_{t1}, ..., \mathbf{Z}_{t2} | \mathbf{Z}_t; t \neq t_1, ..., t_2) \\ &\propto exp\left\{\sum_{t'=t_1}^{t_2}\left[\sum_{i=1}^{n}\sum_{k=0}^{p}\theta_k X_{k,i,t'}Z_{i,t'} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\in N_i}\theta_{p+1}Z_{i,t'}Z_{j,t'}\right] + \sum_{t'=t_1}^{t_2+1}\sum_{i=1}^{n}\theta_{p+2}Z_{i,t'}Z_{i,t'-1}\right\}.\end{aligned} \tag{6.5}$$

for all $t_1, t_2 \in Z$ such that $t_1 < t_2$, where $X_{k,i,t} = X_k(i,t)$ is the k th covariate at site i and time t. Note that the specification is consistent for all $t_1 < t_2$, and the joint distribution of $\{\mathbf{Z}_t : t \in \mathbb{Z}\}$ can be shown to exist by Theorem 2.1.1 of Guyon (1995).

Let $N_{i,t} = \{(j,t) : j \in N_i\} \cup \{(i, t-1), (i, t+1)\}$ denote a neighborhood set for the ith site and the t th time point. Thus, the full conditional distribution of the model is

$$[Z_{i,t} | Z_{i',t'} : (i',t') \neq (i,t)] = [Z_{i,t} | Z_{i',t'} : (i',t') \in N_{i,t}]$$

$$\begin{aligned}&log(\frac{\mathbb{P}(Z_{i,t} = 1 | Z_{i',t'} : (i',t') \in N_{i,t})}{\mathbb{P}(Z_{i,t} = 0 | Z_{i',t'} : (i',t') \in N_{i,t})}) \\ &\qquad = \sum_{k=0}^{p}\theta_k X_{k,i,t} + \sum_{j\in N_i}\theta_{p+1}Z_{j,t'} + \theta_{p+2}(Z_{i,t-1} + Z_{i,t+1})\end{aligned} \tag{6.6}$$

Zheng and Zhu (2008) pointed out that one major drawback of Zhu et al. (2005) is that parameter estimation for the spatial-temporal auto-logistic regression model was based on maximum pseudo-likelihood, which is statistically inefficient. Zheng

and Zhu (2008) propose a fully Bayesian approach and compared it to the maximum pseudo-likelihood and MCMC maximum likelihood approaches.

Followed by that, Zhu et al. (2008) developed a spatio-temporal auto-logistic regression model which depends on only the past. Assuming $[\mathbf{Z_t}|\mathbf{Z}_{t'}, t' = t-1, t-2, ...] = [\mathbf{Z_t}|\mathbf{Z}_{t'}, t' = t-1, ..., t-S]$, for a given point t, follow a Markov random field,

$$[Z_{it}|Z_{jt}, j \neq i; \mathbf{Z}_{t'}, t' = t-1, ..., t-S] = [Z_{it}|Z_{jt}, j \in N_i; \mathbf{Z}_{t'}, t' = t-1, ..., t-S]$$

$$log(\frac{\mathbb{P}(Z_{it}=1|Z_{jt}, j \in N_i; \mathbf{Z}_{t'}, t' = t-1, ..., t-S)}{\mathbb{P}(Z_{it}=0|Z_{jt}, j \in N_i; \mathbf{Z}_{t'}, t' = t-1, ..., t-S)}) = \sum_{k=0}^{p} \theta_k X_{k,i,t} + \sum_{j \in N_i} \theta_{p+1} Z_{j,t} + \sum_{s=1}^{S} \theta_{p+1+s} Z_{i,t-s}$$

They drew statistical inference of this model via maximum likelihood.

Wang and Zheng (2013) developed the estimation methods for the model Equation 6.6 under the centered parametrization:

$$log(\frac{\mathbb{P}(Z_{i,t}=1|Z_{i',t'} : (i',t') \in N_{i,t}; \mathbf{X})}{\mathbb{P}(Z_{i,t}=0|Z_{i',t'} : (i',t') \in N_{i,t}; \mathbf{X})}) = \sum_{k=0}^{p} \theta_k X_{k,i,t} + \sum_{j \in N_i} \theta_{p+1} Z^*_{j,t} + \theta_{p+2}(Z^*_{i,t-1} + Z^*_{i,t+1}) \quad (6.7)$$

Where $Z^*_{i,t} = Z_{i,t} - \frac{exp(\sum_{k=0}^{p} \theta_k X_{k,i,t})}{1+exp(\sum_{k=0}^{p} \theta_k X_{k,i,t})}$.

They proposed expectation-maximization pseudo-likelihood and Monte Carlo expectation-maximization likelihood, as well as consider Bayesian inference to obtain the estimates of model parameters, and they found that Monte Carlo expectation-maximization likelihood is optimal considering the computational cost and the estimation accuracy. Further, they compared the statistical efficiency of the three approaches.

6.2 A two-step centered ST auto-logistic model

6.2.1 Model specification

In this paper, we propose a two-step centered ST auto-logistic model, specified my Markov field Markov chain (Gaetan and Guyon, 2008). We suppose that $\{\mathbf{Z}_t, t = 1, 2, ...\}$ is a Markov chain :

$$[\mathbf{Z}_t|\mathbf{Z}_{t-1}, ...] = [\mathbf{Z}_t|\mathbf{Z}_{t-1}]$$

we assume that $\mathbf{Z}_t$ is a Markov random field conditional on $\mathbf{Z}_{t-1}$ with neighbor structure N_i, thus

$$[Z_{i,t}|Z_{j,t}; j \neq i, \mathbf{Z}_{t-1}] = [Z_{i,t}|Z_{j,t} : j \in N_i, \mathbf{Z}_{t-1}]$$

$$log(\frac{\mathbb{P}(Z_{i,t}=1|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}{\mathbb{P}(Z_{i,t}=1|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_1\sum_{j\in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1}^{*} \tag{6.8}$$

where $Z_{i,t-1}^{*} = Z_{i,t-1} - \frac{exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})}{1=exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})}$ and $Z_{i,t}^{**} = Z_{i,t} - \frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1}^{*})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1}^{*})}$.

Note that this two step centered specification is similar with a hierarchical model with a latent auto-regressive model of order 1 :

$$log(\frac{\mathbb{P}(\mathbf{Z}_{i,t}=1|\xi_{i,t},\mathbf{X}_{i,t})}{\mathbb{P}(\mathbf{Z}_{i,t}=0|\xi_{i,t},\mathbf{X}_{i,t})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \xi_{i,t}$$

$$\xi_{i,t} = \rho_2\xi_{i,t-1} + \omega_{i,t}$$

With following correspondences:

$$Z_{i,t-1}^{*} = Z_{i,t-1} - \frac{exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})}{1+exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})} \approx \xi_{i,t-1}$$

$$\sum_{j\in N_i} Z_{j,t}^{**} = \sum_{j\in N_i}(Z_{j,t} - \frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1}^{*})}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1}^{*})}) \approx \omega_{i,t}$$

6.2.2 Model Interpretation

The main difference between one-step centered (Wang and Zheng, 2013) and two-step centered (describe above) auto-logistic model is that the spatial neighborhoods $Z_{j,t}$ s are centered differently: the former is centered with $\frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta})}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta})}$, and the latter is centered with $\frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1}^{*})}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1}^{*})}$. It has to be noted that the expectation $E(Z_{it}|\mathbf{Z}_{t-1})$ is $\frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1}^{*})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1}^{*})}$ due to the link function.

The construction of this two-step centered auto-logistic model is explained as follows:

At first step, assuming the spatial independence, we consider logistic auto-regressive model of order 1:

$$log(\frac{\mathbb{P}(Z_{i,t}=1|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}{\mathbb{P}(Z_{i,t}=0|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1}^{*}$$

where $Z_{i,t-1}^{*} = Z_{i,t-1} - \frac{exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})}{1+exp(\mathbf{X}_{i,t-1}^T\boldsymbol{\beta})}$. We center the $Z_{i,t-1}$ by removing the expectation $E(Z_{i,t-1})$ under the spatio-temporal independence to separate the large-scale structure and the small-scale temporal variation.

Next, we consider the spatial dependence. As explained in (Caragea and Kaiser, 2009), to get the interpretable parameters, the $\sum Z_{jt}$ have to be centered by removing the expectation under spatial dependence.

So two-step centered model is construct as follows centering with two expectations :

$$log(\frac{\mathbb{P}(Z_{i,t}=1|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}{\mathbb{P}(Z_{i,t}=0|Z_{j,t};j\neq i,\mathbf{Z}_{t-1},\mathbf{X})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_1\sum_{j\in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1}^{*}$$

The key issue of this two-step centralization is to separate the large- and small-scale structures temporally and spatially. The parameters of spatio-temporal dependence ρ_1, ρ_2 can be hierarchically interpreted for practical biological processes:

- Spatial dependence: $\rho_1 \geq 0$.
  Strong spatial dependence indicates a highly aggregated spatial structure, this gives a guide to identify and monitor the aggregated zones with high infection possibility.

- Temporal dependence: $0 \leq \rho_2 \leq 1$. ($\rho_2 < 0$ indicates a temporal evolution with high frequency at 2-year cycle, this is not adapted for most of the biological processes; $\rho_2 > 1$ indicates an explosive temporal evolution which is rare for plant disease.)
  Strong temporal dependence can be interpreted as a smooth temporal evolution. If the exterior effects are stable, the individuals have a tendency to keep their status. This may indicate no need to monitor two consecutive years if the exterior factors are similar between two years.

6.3 Comparative Simulation

6.3.1 Simulation objective

The idea of the two-step centered model is to make an agreement between large-scale model and data structure. Thus we compare the one- and two- step centered models, of which the marginal structure of data should reflect the large-scale structure.

The agreement between spatial large-scale model structures and marginal data structures have been already examined by (Caragea and Kaiser, 2009) for both centered and traditional spatial auto-logistic regression models. And they showed that the data structure of traditional auto-logistic regression model cannot reflect the large-scale structure, and this difficulty can be alleviated by centered parametrization.

In this paper, we focus on examining the temporal-only large-scale model structure and marginal data structure for one- and two-step centered spatio-temporal auto-logistic models since both these two models adopted the centered parametrization proposed by (Caragea and Kaiser, 2009) to get the agreement for the spatial model and data structures.

Therefore, to simulate a dynamic process of Markov random field, we consider a temporal large-scale structure with a temporal tendency without spatial covariates. For site i at year t, we define a large model structure with one temporal covariate: $A_{it} = \beta_0 + \beta_1 X_{it} + \rho_1 \sum_{j \in N_i} Z^{**}_{j,t} + \rho_2 Z^{*}_{i,t-1}$, where $X_{it} = X_t$ is a temporal covariate, that is spatial constant at year t. Thus the average large-scale model at year t is:

$$L_t = \frac{1}{n} \sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 X_t)}{1 + exp(\beta_0 + \beta_1 X_t)} = \frac{exp(\beta_0 + \beta_1 X_t)}{1 + exp(\beta_0 + \beta_1 X_t)}$$

To represent marginal data structure, the marginal data mean of the Markov random field at year t is calculated as:

$$D_t = \frac{1}{n}\sum_{i=1}^{n} Z_{it}^{sim}$$

The simulation studies presented here analyze the relationship between L_t and D_t.

6.3.2 Sampling Algorithms

Hughes et al. (2011) proposed to use perfect sampling to generate MRF samples. The advantage of the perfect sampling compared to Gibbs sampling is that we don't need the burn-in step, neither to decide the spacing numbers. It gives us the exact draws from a given target distribution when the algorithm completes. However, its algorithm running time is random even is still finite. We don't know at which moment the lower chain and the upper chain coalesce. So the number of repetition is random. Evidently in our case, it is long.

Here we use Gibbs sampler but start at a perfect simulation sample, we call it PGS sampling here. It is less time consuming than perfect sampling, and don't need to decide burn-in and spacing when compared with Gibbs sampling. The PGS sampling was often used to generate the simulations of auto-logistic model (Zhu et al., 2008, Zheng and Zhu, 2008, Wang and Zheng, 2013).

In this paper, we generate 4 PGS chains departed with different perfect simulation samples by parallel computing performed on a 4-cores laptop. On one hand, various perfect simulation departs can explore the possible multiple modes of the target distribution, on the other hand, it can increase the chances for the samplings to fall in regions of high probability mass. Each chain contains a departed perfect simulation sample and 10 Gibbs simulation samples.

6.3.3 Simulations

We focused on three types of large-scale model structures with constant, increasing and decreasing temporal tendency respectively. These model structures have been simulated with different values of (ρ_1, ρ_2), in order to evaluate the joint effects of (ρ_1, ρ_2) to the agreement between large-scale structure models and data structures.

Here we simulated the lattices at 20×20 sizes over 15 years. The simulations are presented in the following sections.

6.3.3.1 Intercept only

The model is given by $A_{it} = \beta_0 + \rho_1 \sum_{j\in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1}^{*}$, where $Z_{i,t-1}^{*} = Z_{i,t-1}$ since there is no temporal covariate in this model.

We give two groups of values settings for the parameters: $(\beta_0, \rho_1, \rho_2)$

- Low-level infection: Set $\beta_0 = -1.386$ and $(\rho_1, \rho_2) \in \{(0.1, 0.1), (0.9, 0.9), (0.9, 0.1), (0.1, 0.9)\}$, thus we have $L_t = \frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.2$, the initial field is generated by Bernoulli distribution with parameter 0.2.

- High-level infection: Set $\beta_0 = 1.386$ and $(\rho_1, \rho_2) \in \{(0.1, 0.1), (0.9, 0.9), (0.9, 0.1), (0.1, 0.9)\}$, thus we have $L_t = \frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.8$, the initial field is generated by Bernoulli distribution with parameter 0.8.

From Figure 6.1, we observed that when there is little spatial dependence ($\rho_1 = 0.1$), the data structure agrees with the large scale model structure for both one- and two-step centered auto-logistic models under either low-level-infection or high-level-infection case.
Moreover, when there is a a strong spatial dependence ($\rho_1 = 0.9$), we noticed that:

1. For low-level-infection simulations, the data structures are higher than the large-scale structure; and for the high-level-infection simulations, the data structures are lower than the large-scale structure.

2. When the temporal dependence is also strong ($\rho_2 = 0.9$), the data structure of one-step centered auto-logistic model is more away from the larg-scale model structure than the two-step centered model, and moreover, the distance between the data structure and model structure is steadily increased (for low-infection)/decreased (for high-infection) for the first eight years.

In summary, the difference between the one and two-step centered auto-logistic model for the simulation are closed except when both spatial and temporal dependence are strong. The data structure of one-step centered model more over-reflect the large-scale structure and more seriously, it reflect a increasing/decreasing trend which wrongly conclude the model structure and may confuse the further analysis. Next, we simulate the MRF data from a large-scale structures with increasing and decreasing temporal trend to further study the performances of one- and two-step centered models.

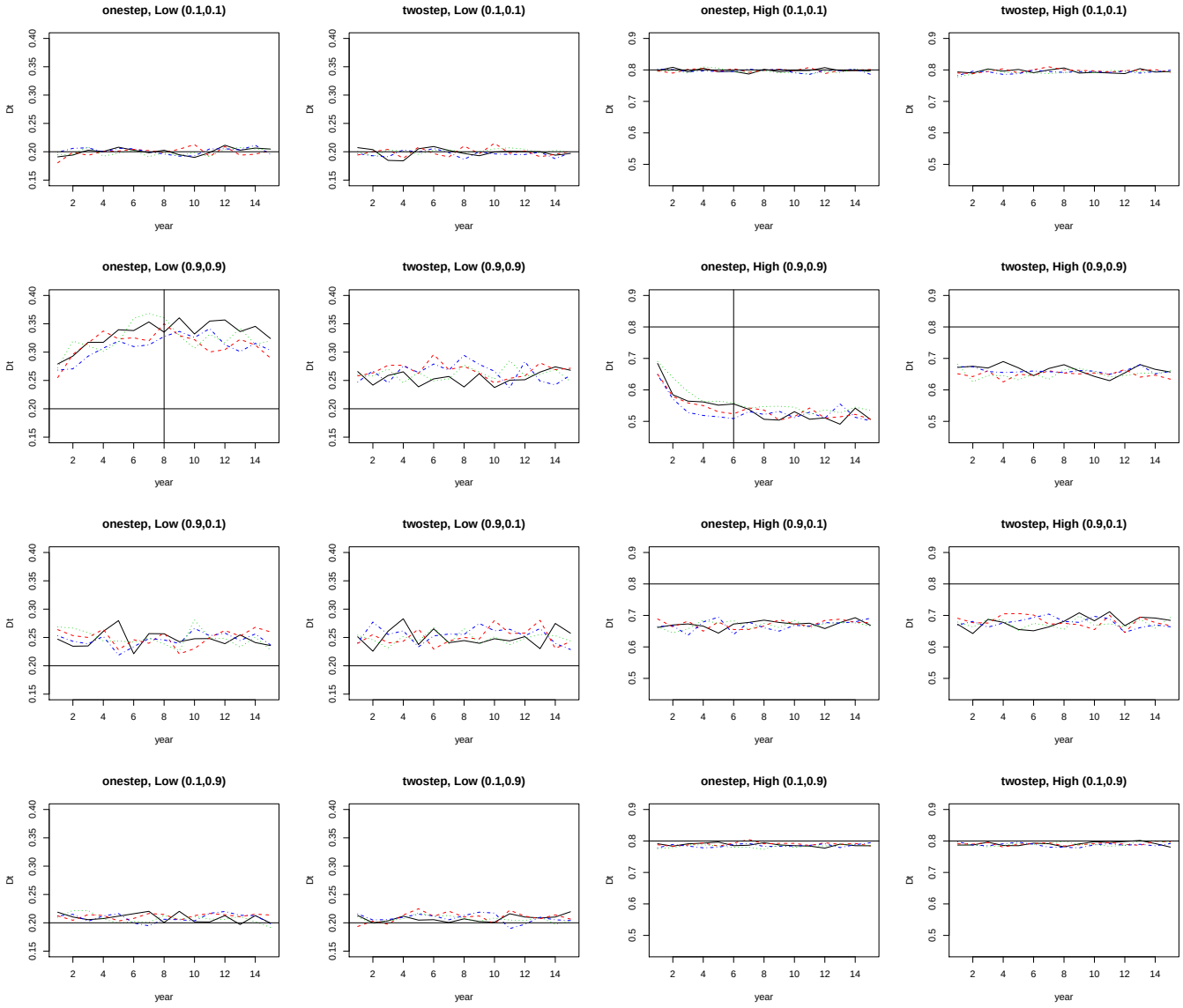


Figure 6.1: Comparison between large-scale model structure L_t (represented by black constant line) and mean of data structures D_t calculated by 4 PGS simulation chains (Each chain contain 11 simulations), they are represented by four curves with different colors and different line types. The small title of each figure specifies the model. For example, the first figure represent the simulations of one step-centered model with low infection($L_t = 0.2$), and (ρ_1, ρ_2) are set to (0.1,0.1)

6.3.3.2 Model with increase or decrease temporal trend

We consider large-scale structures with one temporal covariate: $\beta_0 + \beta_1 X_t$. X_t is defined with monotonic increasing or decreasing tendency.

- Increasing temporal trend
  Set $\beta_0 = -1.5, \beta_1 = 0.2, X_t = t, t = 1, ..., 15$, thus we have $L_t = \frac{exp(\beta_0+\beta_1 t)}{1+exp(\beta_0+\beta_1 t)}$, a monotonic increasing function.

- Decreasing temporal trend
  Set $\beta_0 = -1.5, \beta_1 = 0.2, X_t = 16 - t, t = 1, ..., 15$, thus we have $L_t = \frac{exp(\beta_0+\beta_1 t)}{1+exp(\beta_0+\beta_1 t)}$, a monotonic decreasing function.

From Figure 6.2, we observed that the large-scale model and data structures are not perfectly superposed for both one- and two-step centered logistic model. The slope of the large-scale model is more steep than the data structures. However, the data structures of the two-step centered model is closer to the large scale structure than the one-step centered model, and more important, they reflect a increasing/decreasing trend as required while the data structures of one-step centered model is much less variant.

6.4 Pseudo likelihood Estimation

Since the structure of two-step centered auto logistic model is more complicated than the one-step centered one, both monte carlo maximum likelihood (MCML) and Bayesian methods can be very heavy and sophisticated. For the moment, we estimate such models using Maximum pseudo-likelihood which is relatively easier for estimation. In this section, we give the details to estimate the two-step centered auto-logistic model using pseudo-likelihood.

6.4.1 Algorithms

Let us denote $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \rho_1, \rho_2\}$, the parameters to estimate, we applied the EMPL (Expectation maximization pseudo-likelihood), the principle is the same as described in Zheng and Zhu (2008), but with two iteration steps to accelerate the numerical algorithm/calculation.

The principles are as follows,

- Step 1: To obtain the estimation of $\boldsymbol{\theta_1} = \{\boldsymbol{\beta}, \rho_2\}$,denoted by $\{\tilde{\boldsymbol{\beta}}, \tilde{\rho_2}\}$, from model $log(\frac{\mathbb{P}(Z_{i,t}=1|\mathbf{Z}_{t-1})}{\mathbb{P}(Z_{i,t}=0|\mathbf{Z}_{t-1})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1}^*$

  1. Initialization: obtain $\hat{\boldsymbol{\beta}}$ from model $log(\frac{\mathbb{P}(Z_{i,t}=1)}{\mathbb{P}(Z_{i,t}=0)}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta}$, and start from $\boldsymbol{\theta_1}^0 = \{\hat{\boldsymbol{\beta}}, 1\}$
  2. Expectation step: Given $\boldsymbol{\theta_1}^{l-1} = \{\boldsymbol{\beta}^{l-1}, \rho_2^{l-1}\}$, compute centered responses Z_{it}^{*l-1} by subtracting $\frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}^{l-1})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}^{l-1})}$ from them.

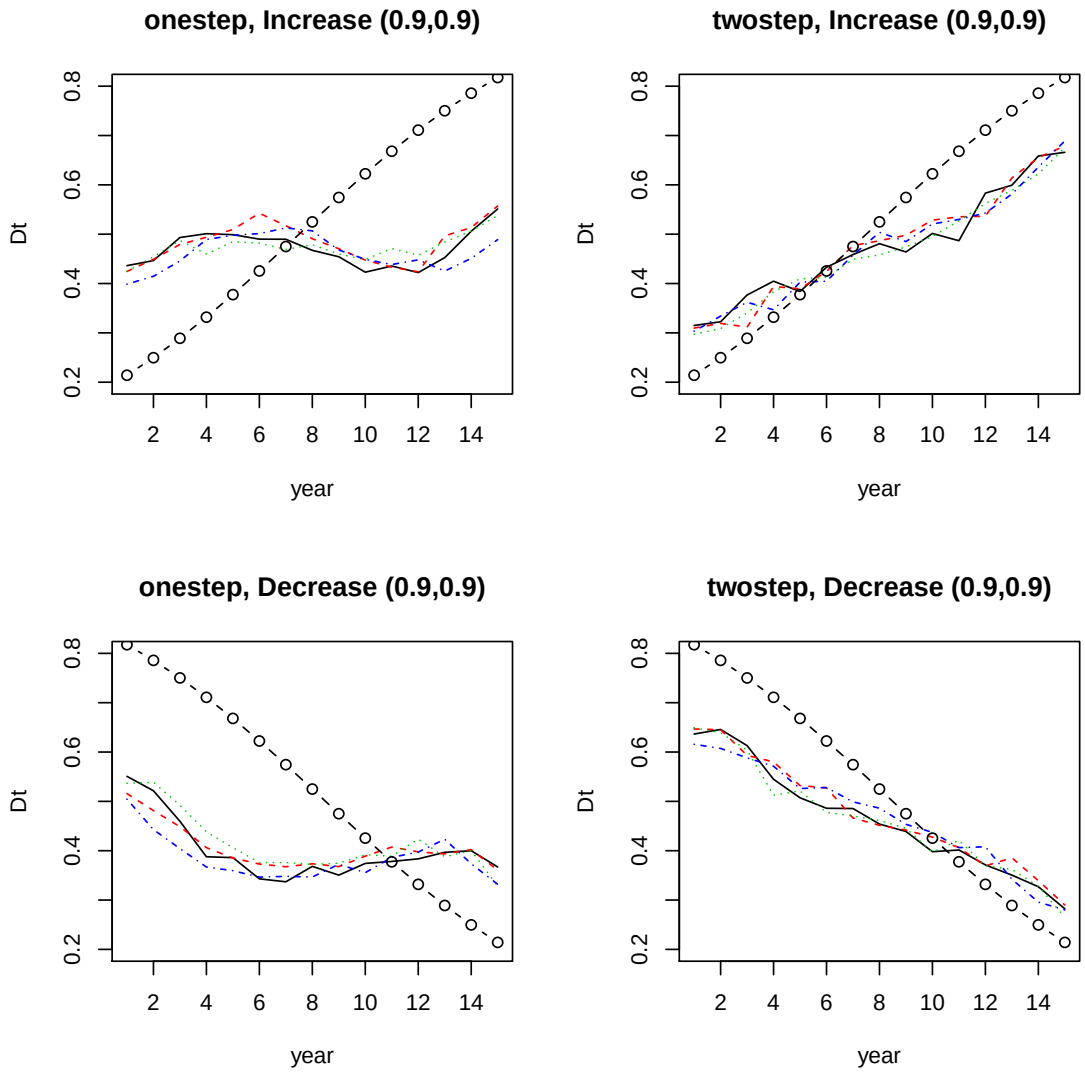


Figure 6.2: Comparison between large-scale model structure L_t (represented by black lines joined by circles) and mean of data structures D_t calculated by 4 PGS simulation chains (Each chain contain 11 simulations), they are presented by lines with different colors and different types. the small title of each figure denote the simulations of which model were drew in the figure, in the same way as before.

3. Maximization step: obtain ${\boldsymbol{\theta_1}}^l$ by maximizing the log pseudo(partial)-likelihood of $log(\frac{\mathbb{P}(Z_{i,t}=1|\mathbf{Z}_{t-1})}{\mathbb{P}(Z_{i,t}=0|\mathbf{Z}_{t-1})}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta}^{l-1} + \rho_2 Z_{i,t-1}^{*l-1}$ by Quasi-Newton.

- Step 2: To obtain the estimation of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \rho_1, \rho_2\}$, denoted by $\{\breve{\boldsymbol{\beta}}, \breve{\rho_1}, \breve{\rho_2}\}$, from two-step centered auto-logistic model.

  1. Initialization: Set initial values: $\boldsymbol{\theta}^0 = \{\tilde{\boldsymbol{\beta}}, 1, \tilde{\rho_2}\}$
  2. Expectation: Given $\boldsymbol{\theta}^{l-1}$, compute Z_{it}^* and Z_{it}^{**} by removing the corresponded trend.
  3. Maximization: Obtain $\boldsymbol{\theta}^l$ by maximizing the log pseudo-likelihood of $log(\frac{\mathbb{P}(Z_{i,t}=1|Z_{j,t;j\neq i},\mathbf{Z}_{t-1})}{\mathbb{P}(Z_{i,t}=0|Z_{j,t;j\neq i},\mathbf{Z}_{t-1})})$ by Quasi-Newton.

- Obtain estimates $\{\breve{\boldsymbol{\beta}}, \breve{\rho_1}, \breve{\rho_2}\}$

6.4.2 Simulation study

We used Perfect Sampling methods described in (Hughes et al., 2011) to generate 20 simulations of a $20 * 20$ grill for 8 years with one temporal covariate with large variation $(17, 10, 13, 33, 19, 14, 15, 21)$ (normalized.) The initial field is generated by Bernoulli distribution with parameter 0.1.

Moreover, since the large disagreement occurs when both ρ_1 and ρ_2 are big, here the initial value have been set $\beta_1 = -1.5, \beta_2 = 0.5, \rho_1 = 0.8, \rho_2 = 0.8$. Here we compare the simulation and estimation of one-step-centered model and our two-step-centered model depends on the past.

	β_1	β_2	ρ_1	ρ_2
mean	-1.7917(-1.5)	0.2322(0.5)	0.1941(0.8)	0.8400(0.8)
variance	0.0221	0.00129	0.0421	0.0159

Table 6.1: One-step centered model simulation estimated by PL with one iteration step, the true values have been presented in the brackets.

	β_1	β_2	ρ_1	ρ_2
mean	-1.4835(-1.5)	0.5167(0.5)	0.7858(0.8)	0.7994(0.8)
variance	0.0117	0.0057	0.0061	0.0084

Table 6.2: two-step centered model simulation estimated by PL with two iteration steps, the true values have been presented in the brackets.

6.5 Discussion

In this chapter, we developed a two-step centered spatio-temporal auto-logistic regression model to fit the binary spatio-temporal data over a lattice with exogenous covariates. The marginal data structure of the model under such centered

parametrization can better reflect the large-scale structure, and in this way, the correct interpretation of the parameters is ensured.

However, this model suffered two statistical drawbacks. The first one is that the existence of a unique joint distribution of such model is not proved. As presented in Chapter 2, a well-defined MCMF (Markov Chain Markov Field) is characterized by the potential functions of instantaneous interaction and of temporal interaction. These potential functions decide a unique joint distribution of dynamical auto-models and the likelihood function is base on them too. In our case, the centered term for the neighbors at the same year Z_{jt}^{**} include a past term $Z_{i,t-1}$, this arise an interaction spatio-temporelle between $exp(Z_{i,t-1})$ and $Z_{j,t}$ in the model and thus becomes difficult to identify the potential functions. One possible solution in perspective is to decompose the Z_{jt}^{**}s, and replace it by an approximate term which is separable to write in the classical form of the potential function.

The second drawback is that even the joint distribution of such model exists, due to the complex structure of the model and the normalizing constant in the likelihood function, such model is difficult to draw the inference. In the chapter, we benefited the easy calculation of pseudo-likelihood function to obtain a numerical estimation of the model which is statistically imperfect. In perspective, several methods based on approximation likelihood could be tried to estimate our models. Recently, Bee et al. (2015) proposed a AMLE (Approximate Maximize Likelihood Estimation) which is very interesting for our case. The advantage of AMLE method is since the only requirement of approximating maximum likelihood estimates is the ability to simulate the model to be estimated. It does not need to evaluate the likelihood function, only sufficient statistics but no joint distribution need to be clarified. When the inference of this model will be improved and adapted to large data set, we will apply the model to esca data in order to test the dependence on neighbors and the effects of environmental covariates.

Chapter 7

Discussion générale et conclusion

7.1 Rappel des objectifs

L'objectif général du travail présenté était d'améliorer la compréhension des processus qui gouvernent la dynamique spatio temporelle de l'esca à l'échelle de la parcelle, en particulier, questionner le rôle contagieux des ceps malades dans la propagation secondaire et détecter et quantifier l'effet de facteurs environnementaux sur l'expression de la maladie. Pour répondre à ces questions épidémiologiques, la stratégies mise en œuvre a été (1) d'analyser la structure spatiale et temporelle de la maladie dans plusieurs parcelles de la région de Bordeaux à l'aide de méthodes statistiques non paramétriques basées sur la dépendance spatiale (2) de développer des méthodes de géostatistique adaptées à la spécificité des sondages en vignoble, pour obtenir une information au cep pour plusieurs indicateurs liés à la plante ou au sol (3) de construire des modèles hiérarchiques bayésiens intégrant un processus latent de dépendance spatiale ou spatio-temporelle pour la modélisation de l'occurrence (ou première occurrence) des symptômes foliaires de l'esca.

Les résultats originaux de la thèse comprennent d'une part une meilleure compréhension de la dynamique de la maladie et d'autre part un apport d'outils de statistique non paramétrique, de géostatistique et de modélisation adaptés au modèle biologique étudié, au recueil spécifique des données et bien sûr à nos questions.

7.2 Apports de connaissances épidémiologiques

La synthèse des résultats des deux approches (1) & (3) nous apporte des arguments scientifiques en faveur d'un risque nul ou faible de contagion à partir de ceps malades d'esca, c'est-à-dire symptomatiques. La distribution des ceps malades est souvent aléatoire même pour des parcelles présentant un fort taux de maladie. Dans le cas d'une structure agrégée, il n'y a pas d'évidence d'augmentation de la taille des foyers comme on pourrait s'y attendre dans le cas d'un processus contagieux. Enfin, dans les modèles logistiques intégrant les facteurs environnementaux, les facteurs liés au voisinage de ceps ayant déjà exprimé des symptômes ne sont pas retenus. Ce résultat signifie que ces derniers ne modifient pas significativement la probabilité qu'un cep devienne symptomatique.

Les vignes étant plantées en rang, nous pouvions nous attendre dans le cas d'une structure en agrégat à des clusters de ceps malades le long des rangs. Si cette struc-

ture est relevée pour quelques parcelles, il n'y a pas non plus d'augmentation de la taille de ces clusters dans le temps. L'hypothèse d'un environnement local favorable à la maladie est plutôt retenue. Encore une fois, les résultats de la modélisation logistique confirment ceux de l'analyse exploratoire. Les modèles incluant le voisinage de ceps ayant déjà exprimé des symptômes et localisés sur le même rang ou hors rang ne sont pas non plus retenus. D'un point de vue de la gestion de la maladie, ces résultats nous donnent des indications pour les méthodes de prophylaxie : il ne semble pas que la propagation de la maladie le long des rangs par les outils de taille soit importante dans les parcelles étudiées.

Tous ces résultats nous conduisent à ne pas négliger le rôle de l'environnement intra parcellaire dans la dynamique spatio-temporelle. Ceci est montré pour une des deux parcelles (Vignoble 12). Dans ce cas, nous confirmons l'hypothèse proposée sur l'effet significatif de la vigueur sur le risque de maladie. Pour cette parcelle, la variable "total des précipitations" est aussi retenue dans le modèle logistique le plus performant, par contre cette variable n'est pas retenue pour l'autre parcelle. Nous pensons que l'effet du climat existe bel et bien sur la maladie, qu'il agit de manière complexe à différents niveaux du système mais qu'il est difficile à mettre en évidence. Nous admettons que le choix de la variable "température" n'était pas très pertinent, au vu de la faible variabilité de cet indicateur. Néanmoins, la sélection de cet indicateur dans les modèles nous incite à poursuivre la recherche d'indicateur intégrant la température. Pour une telle maladie pérenne, l'effet à moyen terme du climat n'est pas négligeable. L'effet du climat pourra aussi être intégré via un modèle biophysique afin de déterminer soit à l'échelle de la plante ou à l'échelle de la parcelle un indicateur de contrainte hydrique. Une première indication est cependant donnée avec la variable estimée au cep delta C13 qui est sélectionnée par le modèle logistique multiple REG-all (Vignoble 8). Dans ce cas, le risque d'expression est augmenté avec une diminution de la contrainte hydrique. En perspective, la variable réserve utile du sol spatialisée à l'échelle du cep pour la parcelle St Valerin intégrée avec les données climatiques au modèle biophysique de (Lebon et al., 2003) devrait nous conduire à explorer le lien entre l'état hydrique de la plante et la maladie. Conscients des difficultés méthodologiques, cette approche pluri-disciplinaire devra être réalisée en prenant en compte les dernières avancées en écophysiologie de la vigne. Une première approche des relations entre "réserve utile" et maladie montre un lien complexe non monotone ni de surcroît linéaire (Figure 7.1). La fonction parabolique observée devra être expliquée.

7.3 Apports méthodologiques

Partant de très peu de connaissance sur la maladie, nous avons adapté et décliné le test du Join count pour répondre à nos différentes questions épidémiologiques. Le test est décliné en quatre types de tests selon des objectifs différents : les Distance tests sont utilisés pour tester la taille et la direction de l'agrégation à une échelle plus grande, les Neighbors tests explorent le lien entre ceps nouvellement et précédemment malades, afin d'identifier l'ordre de la propagation et la direction.

Pour répondre à des questions globales de propagation sur toute la période et/ou tout le rang, et/ou toutes les distances ou ordre de voisinage, nous avons développé un test global agrégeant les statistiques des tests à l'année-parcelle-rang et permet-

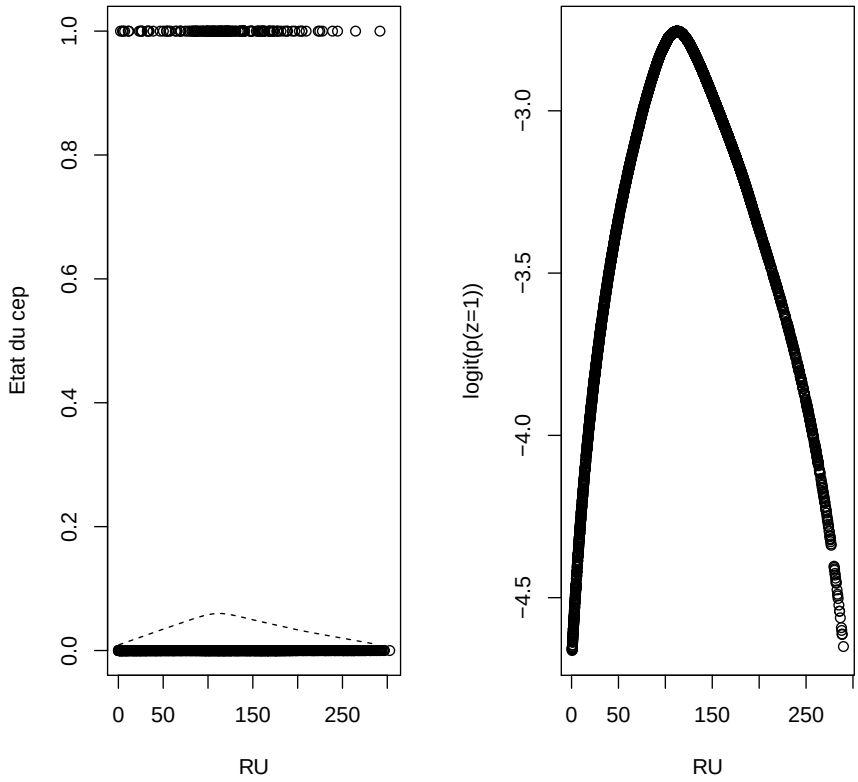


Figure 7.1: y-axis, a non-parametric estimation of presence-absence data for esca disease; x-axis, estimated available water capacity for each vine. The figure showed a non-linear relationship between the disease and the available water capacity.

tant de hiérarchiser des objectifs de tests dans une structure ascendante, afin d'éviter des tests croisés et parallèles.

Les résultats de cette étude exploratoire nous ont aidés à construire des modèles hiérarchiques spatio-temporels adaptés à nos données et aux questions épidémiologiques posées. L'approche par modélisation dynamique nous a permis d'identifier des covariables spatiales ou temporelles d'intérêt. Nous avons pu étudier les deux hypothèses biologiques en intégrant dans un modèle incluant un processus latent les deux types de facteurs. Ce modèle est le premier de ce type développé pour étudier la dynamique spatio-temporelle de l'esca de la vigne en intégrant les facteurs environnementaux. Il servira à tester d'autres variables d'intérêt liées à l'environnement. Ce modèle ne considère que la transition d'état entre asymptomatique et symptomatique pour la première fois (nouveau cas) ou non. En perspective, il pourrait intégrer plus de complexité en modélisant simultanément la transition entre plusieurs états (sain, symptomatique, asymptomatique après symptôme, mort,..) par exemple lorsque nous aurons plus de recul temporel dans les données.

L'interpolation des facteurs liés au sol ou à la plante était un objectif ambitieux de la thèse, compte-tenu de l'échantillonnage très spécifique des variables étudiées. Nous avons proposé dans ce travail de thèse une méthode géostatistique originale pour estimer la réserve utile en chaque point de la parcelle à partir de données de résistivité et d'un échantillonnage ponctuel orienté de données de réserve utile du sol. Le challenge méthodologique était de prendre en compte la structure spatiale des données de résistivité organisée en ligne, les lignes étant espacées de plusieurs mètres. Nous avons combiné plusieurs méthodes géostatistiques, dont une méthode de krigeage « median polish » dans un sens qui s'adapte bien pour les données mesurées selon le rang. Cette méthode pourra être utilisée pour toute disposition similaire des données mesurées en ligne dans un environnement de culture pérenne (vergers par exemple).

Dans ce travail, les covariables que nous avons utilisées sont estimées par un modèle de krigeage. Dans nos modèles explicatifs de la maladie, elles sont intégrées comme si elles avaient été mesurées exactement en tout point. Les études futures doivent aussi prendre en compte l'incertitude des résultats de krigeage intégrés à un modèle caractérisé par une structure auto-corrélée. Ceci est d'autant plus important si elles doivent être utilisées pour faire de la prévision et donner une marge de confiance à cette prévision. Bien sûr cela n'empêche pas d'intégrer d'autres indicateurs pouvant être mesurés de façon systématique sur l'ensemble des ceps "sains" de la parcelle et produire une estimation plus précise. Par exemple, pour la mesure de la vigueur, il est possible d'utiliser des capteurs optiques type N-testeurs ou Greenseeker.

Enfin, nous avons proposé une autre approche de modélisation de la maladie grâce à la mise en œuvre d'un modèle autologistique adapté pour modéliser la dynamique de données sur réseau. Il modélise une dépendance spatiale par régression sur les individus eux-mêmes. Cependant, ce type de modèle n'est pas facilement estimable car sa vraisemblance contient une constante normalisée. Toutefois, il est plus adapté pour modéliser une structure spatiale à petite distance explicite. Dans ce travail, nous n'avons pas cherché à résoudre les problèmes de l'estimation de la constante normalisée. Par contre, nous nous sommes concentrés sur le problème d'interprétation des paramètres pour ce type de modèles avec régression :

la structure des données générées par les modèles ne reflète pas la structure du modèle lui-même. Et l'interprétation des paramètres ρ_1 et ρ_2 du modèle peut être trompeuse. C'est pourquoi, nous avons proposé une approche en deux étapes centrées pour améliorer ce problème d'interprétation. Même si le paramétrage en deux étapes centrées n'est pas décomposable pour la formalisation théorique de " chaine de Markov et champ de Markov", nous avons pu tester le modèle par des simulations. La pertinence de ce modèle pour répondre à nos questions devra être validée. Il faudra le tester avec nos données d'observation.

7.4 Perspectives

Ce travail de thèse démontre que l'analyse des données d'esca mérite encore beaucoup d'attention et de développement. D'une part, nos résultats montrent de la variabilité entre les différentes années et entre les différentes parcelles. Cette variabilité doit être expliquée : quelle est la part incompressible due au caractère erratique de l'esca ? Quelle est la part qui peut être réduite par la recherche d'autres facteurs, quelles covariables (environnementales mais aussi de conduite de la vigne) peuvent expliquer les variations spatio-temporelles de la maladie ? Pour ce faire, d'autres études, en utilisant des données supplémentaires provenant de différents vignobles, doivent être menées pour confirmer et étendre les résultats obtenus dans la thèse et répondre à d'autres questions. C'était l'un des objectifs à long terme de ce travail, construire des modèles qui pourront permettre de faire des prédictions de prévalence ou d'incidence des symptômes pour une année donnée en fonction des années précédentes et d'un scénario climatique par exemple.

Bibliography

Abou-Mansour, E., Couche, E., and Tabacchi, R. (2004). Do fungal naphthalenones have a role in the development of esca symptoms? *Phytopathologia Mediterranea*, 43(1):75–82.

Acevedo-Opazo, C., Ortega-Farias, S., and Fuentes, S. (2010). Effects of grapevine (vitis vinifera l.) water status on water consumption, vegetative growth and grape quality: An irrigation scheduling application to achieve regulated deficit irrigation. *Agricultural Water Management*, 97(7):956–964.

Agustí-Brisach, C., León, M., Garcia-Jimenez, J., and Armengol, J. (2014). Detection of grapevine fungal trunk pathogens on pruning shears and evaluation of their potential for spread of infection. *Plant Disease*, pages PDIS–12.

Andolfi, A., Cimmino, A., Evidente, A., Iannaccone, M., Capparelli, R., Mugnai, L., and Surico, G. (2009). A new flow cytometry technique to identify phaeomoniella chlamydospora exopolysaccharides and study mechanisms of esca grapevine foliar symptoms. *Plant Disease*, 93(7):680–684.

Andolfi, A., Mugnai, L., Luque, J., Surico, G., Cimmino, A., and Evidente, A. (2011). Phytotoxins produced by fungi associated with grapevine trunk diseases. *Toxins*, 3(12):1569–1605.

Ayachi, S. (2010). Caractérisation de la réserve utile des sols viticoles bourguignons dans le réseau suivi pour les maladies du bois. Master's thesis, Université de Bourgogne.

Baize, D. and Girard, M. C. (2008). Référentiel pédologique 2008. page 405.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Bassanezi, R. B., Bergamin Filho, A., Amorim, L., Gimenes-Fernandes, N., Gottwald, T. R., and Bové, J. M. (2003). Spatial and temporal analyses of citrus sudden death as a tool to generate hypotheses concerning its etiology. *Phytopathology*, 93(4):502–512.

Bee, M., Espa, G., and Giuliani, D. (2015). Approximate maximum likelihood estimation of the autologistic model. *Computational Statistics & Data Analysis*, 84:14–26.

Bertsch, C., Ramírez-Suero, M., Magnin-Robert, M., Larignon, P., Chong, J., Abou-Mansour, E., Spagnolo, A., Clément, C., and Fontaine, F. (2013). Grapevine trunk diseases: complex and still poorly understood. *Plant pathology*, 62(2):243–265.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Bierkens, M., Finke, P., and De Willigen, P. (2000). Upscaling and downscaling methods for environmental research.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170.

Borie, B., Jacquiot, L., Jamaux-Despréaux, I., Larignon, P., and Péros, J.-P. (2002). Genetic diversity in populations of the fungi phaeomoniella chlamydospora and phaeoacremonium aleophilum on grapevine in france. *Plant Pathology*, 51(1):85–96.

Bourennane, H. and King, D. (2003). Using multiple external drifts to estimate a soil variable. *Geoderma*, 114(1):1–18.

Bourennane, H., Nicoullaud, B., Couturier, A., Pasquier, C., Mary, B., and King, D. (2012). Geostatistical filtering for improved soil water content estimation from electrical resistivity data. *Geoderma*, 183:32–40.

Brillante, L., Bois, B., Mathieu, O., Bichet, V., Michot, D., and Lévêque, J. (2014). Monitoring soil volume wetness in heterogeneous soils by electrical resistivity. a field-based pedotransfer function. *Journal of Hydrology*, 516:56–66.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, pages 481–483.

Bruand, A., Duval, O., and Cousin, I. (2004). Estimation des propriétés de rétention en eau des sols à partir de la base de données solhydro: Une première proposition combianant le type d'horizon, sa texture et sa densité apparente. *Etude et gestion des Sols*, 11:3–323.

Bruand, A., Fernandez, P. P., Duval, O., Quétin, P., Nicoullaud, B., Gaillard, H., Raison, L., Pessaud, J.-F., and Prud'Homme, L. (2002). Estimation des propriétés de rétention en eau des sols: utilisation de classes de pédotransfert après stratifications texturale et texturo-structurale. *Etude et gestion des sols*, 9:2–105.

Bruez, E., Lecomte, P., Grosman, J., Doublet, B., Bertsch, C., Fontaine, F., Ugaglia, A., Teissedre, P.-L., Da Costa, J.-P., Guerin-Dubrana, L., et al. (2013). Overview of grapevine trunk diseases in france in the 2000s. *Phytopathologia Mediterranea*, 52(2):262–275.

Bruez, E., Vallance, J., Gerbore, J., Lecomte, P., Da Costa, J.-P., Guerin-Dubrana, L., and Rey, P. (2014). Analyses of the temporal dynamics of fungal communities colonizing the healthy wood tissues of esca leaf-symptomatic and asymptomatic vines. *PloS one*, 9(5).

Bruno, G. and Sparapano, L. (2006a). Effects of three esca-associated fungi on vitis vinifera l.: I. characterization of secondary metabolites in culture media and host responses to the pathogens in calli. *Physiological and Molecular Plant Pathology*, 69(4):209–223.

Bruno, G. and Sparapano, L. (2006b). Effects of three esca-associated fungi on vitis vinifera l.: Ii. characterization of biomolecules in xylem sap and leaves of healthy and diseased vines. *Physiological and Molecular Plant Pathology*, 69(4):195–208.

Bruno, G., Sparapano, L., and Graniti, A. (2007). Effects of three esca-associated fungi on vitis vinifera l.: Iv. diffusion through the xylem of metabolites produced by two tracheiphilous fungi in the woody tissue of grapevine leads to esca-like symptoms on leaves and berries. *Physiological and Molecular Plant Pathology*, 71(1):106–124.

Buvat, S., Thiesson, J., Michelin, J., Nicoullaud, B., Bourennane, H., Coquet, Y., and Tabbagh, A. (2014). Multi-depth electrical resistivity survey for mapping soil units within two 3ha plots. *Geoderma*, 232:317–327.

Calzarano, F. and Di Marco, S. (2007). Wood discoloration and decay in grapevines with esca proper and their relationship with foliar symptoms. *Phytopathologia Mediterranea*, 46(1):96.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Caragea, P. C. and Kaiser, M. S. (2009). Autologistic models with interpretable parameters. *Journal of agricultural, biological, and environmental statistics*, 14(3):281–300.

Celano, G., Palese, A., Ciucci, A., Martorella, E., Vignozzi, N., and Xiloyannis, C. (2011). Evaluation of soil water content in tilled and cover-cropped olive orchards by the geoelectrical technique. *Geoderma*, 163(3):163–170.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*, volume 44. Pion London.

Cordeau, J. (1998). Création d'un vignoble. *Greffage de la vigne et portegreffes. Élimination des maladies à virus, Editions Féret.*

Cortesi, P., Fischer, M., and Milgroom, M. G. (2000). Identification and spread of fomitiporia punctata associated with wood decay of grapevine showing symptoms of esca. *Phytopathology*, 90(9):967–972.

Cressie, N. (1993). *Statistics for Spatial Data: Wiley Series in Probability and Statistics*. Wiley-Interscience New York.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Destrac-Irvine, A., Laveau, C., Goutouly, J., and Guérin-Dubrana, L. (2007). L'écophysiologie de la vigne pour mieux comprendre les maladies de dépérissement. *Union Girondine des Vins de Bordeaux*, 1035:28–32.

Dillon, W. W., Haas, S. E., Rizzo, D. M., and Meentemeyer, R. K. (2014). Perspectives of spatial scale in a wildland forest epidemic. *European journal of plant pathology*, 138(3):449–465.

Dubos, B. (1999). *Les maladies cryptogamiques de la vigne: les champignons parasites des organes herbacés et du bois de la vigne*. Éditions Féret.

Edwards, J., Marchi, G., and Pascoe, I. G. (2001). Young esca in australia. *Phytopathologia Mediterranea*, 40(3):303–310.

Eskalen, A. and Gubler, W. D. (2001). Association of spores of «phaeomoniella chlamydospora»,«phaeoacremonium inflatipes», and «pm. aleophilum» with grapevine cordons in california. *Phytopathologia Mediterranea*, 40(3):429–431.

Evidente, A., Sparapano, L., Andolfi, A., Bruno, G., et al. (2000). Two naphthalenone pentaketides from liquid cultures of phaeoacremonium aleophilum, a fungus associated with esca of grapevine. *Phytopathologia mediterranea*, 39(1):162–168.

Feliciano, A. J., Eskalen, A., and Gubler, W. D. (2004). Differential susceptibility of three grapevine cultivars to phaeoacremonium aleophilum and phaeomoniella chlamydospora in california. *Phytopathologia Mediterranea*, 43(1):66–69.

Fischer, M. (2002). A new wood-decaying basidiomycete species associated with esca of grapevine: Fomitiporia mediterranea (hymenochaetales). *Mycological Progress*, 1(3):315–324.

Fischer, M. and Kassemeyer, H.-H. (2015). Water regime and its possible impact on expression of esca symptoms in vitis vinifera: growth characters and symptoms in the greenhouse after artificial infection with phaeomoniella chlamydospora. *VITIS-Journal of Grapevine Research*, 51(3):129.

Fussler, L., Kobes, N., Bertrand, F., Maumy, M., Grosman, J., and Savary, S. (2008). A characterization of grapevine trunk diseases in france from data generated by the national grapevine wood diseases survey. *Phytopathology*, 98(5):571–579.

Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiales*, volume 63. Springer.

Garrett, K., Dendy, S., Frank, E., Rouse, M., and Travers, S. (2006). Climate change effects on plant disease: genomes to ecosystems. *Annu. Rev. Phytopathol.*, 44:489–509.

Gaudillère, J.-P., Van Leeuwen, C., and Ollat, N. (2002). Carbon isotope composition of sugars in grapevine, an integrated indicator of vineyard water status. *Journal of Experimental Botany*, 53(369):757–763.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Geoffrion, R. (1971). L'esca de la vigne dans les vignobles de l'ouest. *Phytoma*, 23:21–31.

Geoffrion, R. and Renaudin, I. (2002). Anti-esca pruning. a useful measure against outbreaks of this old grapevine disease. *Phytoma. La Défense des Végétaux (France)*.

Gibson, G. (1997). Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. *Phytopathology*, 87(2):139–146.

Gibson, G. and Austin, E. (1996). Fitting and testing spatio-temporal stochastic models with application in plant epidemiology. *Plant Pathology*, 45(2):172–184.

Gilligan, C. A. and van den Bosch, F. (2008). Epidemiological models for invasion and persistence of pathogens. *Annu. Rev. Phytopathol.*, 46:385–418.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford university press.

Gosme, M. (2007). *Modélisation du développement spatio-temporel des maladies d'origine tellurique*. PhD thesis, Agrocampus-Ecole nationale supérieure d'agronomie de rennes.

Gramaje, D. and Armengol, J. (2011). Fungal trunk pathogens in the grapevine propagation process: potential inoculum sources, detection, identification, and management strategies. *Plant Disease*, 95(9):1040–1055.

Gray, S., Moyer, J., and Bloomfield, P. (1986). Two-dimensional distance class model for quantitative description of virus-infected plant distribution lattices. *Phytopathology*, 76(2):243–248.

Guérin-Dubrana, L., Labenne, A., Labrousse, J. C., Bastien, S., Patrice, R., and Gégout-Petit, A. (2013). Statistical analysis of grapevine mortality associated with esca or eutypa dieback foliar expression. *Phytopathologia Mediterranea*, 52(2):276–288.

Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 131–156.

Guyon, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media.

Guyon, X. and Hardouin, C. (2002). Markov chain markov field dynamics: models and statistics. *Statistics: A Journal of Theoretical and Applied Statistics*, 36(4):339–363.

Haran, M. (2011). Gaussian random field models for spatial data. *Handbook of Markov Chain Monte Carlo*, pages 449–478.

Hardouin, C. (2011). *Quelques contributions à la modélisation et l'analyse statistique de processus spatiaux*. PhD thesis, Université Paris Ouest Nanterre La Défense.

Hengl, T., Heuvelink, G., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1):75–93.

Hengl, T., Heuvelink, G. B., and Stein, A. (2003). Comparison of kriging with external drift and regression-kriging. *Technical note, ITC*, 51.

Hesse, A., Jolivet, A., and Tabbagh, A. (1986). New prospects in shallow depth electrical surveying for archaeological and pedological applications. *Geophysics*, 51(3):585–594.

Huffer, F. W. and Wu, H. (1998). Markov chain monte carlo for autologistic regression models with application to the distribution of plant species. *Biometrics*, pages 509–524.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871.

Jousimo, J., Tack, A. J., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., and Laine, A.-L. (2014). Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science*, 344(6189):1289–1293.

Kaiser, M. S., Pazdernik, K. T., Lock, A. B., and Nutter, F. W. (2014). Modeling the spread of plant disease using a sequence of binary random fields with absorbing states. *Spatial Statistics*, 9:38–50.

Kang, Y., Khan, S., and Ma, X. (2009). Climate change impacts on crop yield, crop water productivity and food security–a review. *Progress in Natural Science*, 19(12):1665–1674.

Keeling, M. J. and Ross, J. V. (2008). On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, 5(19):171–181.

Larignon, P. and Dubos, B. (1997). Fungi associated with esca disease in grapevine. *European Journal of Plant Pathology*, 103(2):147–157.

Larignon, P. and Dubos, B. (2000). Preliminary studies on the biology of phaeoacremonium. *Phytopathologia Mediterranea*, 39(1):184–189.

Larignon, P., Fontaine, F., Farine, S., Clément, C., and Bertsch, C. (2009). Esca et black dead arm: deux acteurs majeurs des maladies du bois chez la vigne. *Comptes Rendus Biologies*, 332(9):765–783.

Laveau, C., Letouze, A., Louvet, G., Bastien, S., and Guerin-Dubrana, L. (2009). Differential aggressiveness of fungi implicated in esca and associated diseases of grapevine in france. *Phytopathologia Mediterranea*, 48(1):32–46.

Lavigne, A. (2013). *Modélisation statistique régionale de l'activité avalancheuse*. PhD thesis, AgroParisTech.

Lebon, E., Dumas, V., Pieri, P., and Schultz, H. R. (2003). Modelling the seasonal dynamics of the soil water balance of vineyards. *Functional Plant Biology*, 30(6):699–710.

Lecomte, P., Darrieutort, G., Laveau, C., Blancard, D., Louvet, G., Goutouly, J., Rey, P., and Guérin-Dubrana, L. (2011). Impact of biotic and abiotic factors on the development of esca decline disease."integrated protection and production in viticulture". *IOBC/wprs Bulletin*, 67:171–180.

Lecomte, P., Darrieutort, G., Liminana, J.-M., Comont, G., Muruamendiaraz, A., Legorburu, F.-J., Choueiri, E., Jreijiri, F., El Amil, R., and Fermaud, M. (2012). New insights into esca of grapevine: the development of foliar symptoms and their association with xylem discoloration. *Plant Disease*, 96(7):924–934.

Lejeune, M. (2006). *Analyse statistique des données spatiales*. Editions TECHNIP.

Lele, S. R. and Dennis, B. (2009). Bayesian methods for hierarchical models: are ecologists making a faustian bargain. *Ecological Applications*, 19(3):581–584.

Li, S., Bonneu, F., Chadoeuf, J. J., Picart, D., Gégout-Petit, A., and Guerin-Dubrana, L. (2015). Spatial and Temporal Pattern Analyses 1 of Esca Grapevine Disease in Vineyards in France.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lorrain, B., Ky, I., Pasquier, G., Jourdes, M., Dubrana, L. G., Gény, L., Rey, P., Donèche, B., and TEISSEDRE, P.-L. (2012). Effect of esca disease on the phenolic and sensory attributes of cabernet sauvignon grapes, musts and wines. *Australian Journal of Grape and Wine Research*, 18(1):64–72.

Madden, L. V., Hughes, G., and Van den Bosch, F. (2007). *The study of plant disease epidemics*. American Phytopathological Society St. Paul.

Maher, N., Piot, J., Bastien, S., Vallance, J., Rey, P., and Guérin-Dubrana, L. (2012). Wood necrosis in esca-affected vines: types, relationships and possible links with foliar symptom expression.

Marchi, G., Peduto, F., Mugnai, L., Di Marco, S., Calzarano, F., and Surico, G. (2006). Some observations on the relationship of manifest and hidden esca to rainfall. *Phytopathologia Mediterranea*, 45(4):117–126.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

McGrory, C. A., Titterington, D. M., Reeves, R., and Pettitt, A. N. (2009). Variational bayes for estimating the parameters of a hidden potts model. *Statistics and Computing*, 19(3):329–340.

McKinley, J. M., Ofterdinger, U., Young, M., Barsby, A., and Gavin, A. (2013). Investigating local relationships between trace elements in soils and cancer data. *Spatial Statistics*, 5:25–41.

Michot, D., Benderitter, Y., Dorigny, A., Nicoullaud, B., King, D., and Tabbagh, A. (2003). Spatial and temporal monitoring of soil water content with an irrigated corn crop cover using surface electrical resistivity tomography. *Water Resources Research*, 39(5).

Moller, J. (1999). Perfect simulation of conditionally specified models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 251–264.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Moyo, P., Allsopp, E., Roets, F., Mostert, L., and Halleen, F. (2014). Arthropods vector grapevine trunk disease pathogens. *Phytopathology*, 104(10):1063–1069.

Mugnai, L., Graniti, A., Surico, G., et al. (1999). Esca (black measles) and brown wood-streaking: two old and elusive diseases of grapevines. *Plant disease*, 83(5):404–418.

Murolo, S. and Romanazzi, G. (2014). Effects of grapevine cultivar, rootstock and clone on esca disease. *Australasian Plant Pathology*, 43(2):215–221.

Nelder, J. A. and Baker, R. (1972). *Generalized linear models*. Wiley Online Library.

Péros, J.-P., Jamaux-Despréaux, I., Berger, G., et al. (2000). Population genetics of fungi associated with esca disease in french vineyards. *Phytopathologia Mediterranea*, 39(1):150–155.

Peyrard, N., Pellegrin, F., Chadoeuf, J., and Nandris, D. (2006). Statistical analysis of the spatio-temporal dynamics of rubber tree (hevea brasiliensis) trunk phloem necrosis: no evidence of pathogen transmission. *Forest pathology*, 36(5):360–371.

Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1).

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*.

Pollastro, S., Dongiovanni, C., Abbatecola, A., Faretra, F., et al. (2000). Observations on the fungi associated with esca and on spatial distribution of esca-symptomatic plants in apulian (italy) vineyards. *Phytopathologia Mediterranea*, 39(1):206–210.

Pollastro, S., Dongiovanni, C., Habib, W., and Faretra, F. (2009). Long-term observations on the spatial distribution of esca disease in vineyards. *Journal of Plant Pathology*, 91(4).

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2):223–252.

Purse, B. V. and Golding, N. (2015). Tracking the distribution and impacts of diseases with biological records and distribution modelling. *Biological Journal of the Linnean Society*, 115(3):664–677.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S. (2008). Analysis of near-surface atmospheric variables: Validation of the safran analysis over france. *Journal of applied meteorology and climatology*, 47(1):92–107.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redondo, C., Tello, M., Avila, A., and Mateo-Sagasta, E. (2001). Spatial distribution of symptomatic grapevines with esca disease in madrid region (spain)[vitis vinifera l.]. *Phytopathologia Mediterranea (Italy)*.

Reisenzein, H., Berger, N., Nieder, G., et al. (2000). Esca in austria. *Phytopathologia Mediterranea*, 39(1):26–34.

Riebler, A. and Held, L. (2009). The analysis of heterogeneous time trends in multivariate age–period–cohort models. *Biostatistics*, page kxp037.

Rivoirard, J. (2002). On the structural link between variables in kriging with external drift. *Mathematical geology*, 34(7):797–808.

Rossi, R., Amato, M., Pollice, A., Bitella, G., Gomes, J., Bochicchio, R., and Baronti, S. (2013). Electrical resistivity tomography to detect the effects of tillage in a soil with a variable rock fragment content. *European Journal of Soil Science*, 64(2):239–248.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Samouëlian, A., Cousin, I., Tabbagh, A., Bruand, A., and Richard, G. (2005). Electrical resistivity survey in soil science: a review. *Soil and Tillage research*, 83(2):173–193.

Shaddick, G. and Zidek, J. V. (2014). A case study in preferential sampling: Long term monitoring of air pollution in the uk. *Spatial Statistics*, 9:51–65.

Shaw, M. (1994). Modeling stochastic processes in plant pathology. *Annual review of phytopathology*, 32(1):523–544.

Sofia, J., Gonçalves, M. T., and Oliveira, H. (2006). Spatial distribution of esca symptomatic plants in dão vineyards (centre portugal) and isolation of associated fungi. *Phytopathologia Mediterranea*, 45(4):87–92.

Sparapano, L., De Leonardis, S., Campanella, A., and Bruno, G. (2001). Interaction between esca-associated fungi, grapevine calli and micropropagated shoot cultures of grapevine. *Phytopathologia Mediterranea*, 40(3):423–428.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spolti, P., Valdebenito-Sanhueza, R., Laranjeira, F., and Del Ponte, E. (2012). Comparative spatial analysis of the sooty blotch/flyspeck disease complex, bull's eye and bitter rots of apples. *Plant Pathology*, 61(2):271–280.

Stefanini, F. M., Surico, G., Marchi, G., et al. (2000). Longitudinal analysis of symptom expression in grapevines affected by esca. *Phytopathologia Mediterranea*, 39(1):225–231.

Surico, G., Marchi, G., Braccini, P., Mugnai, L., et al. (2000a). Epidemiology of esca in some vineyards in tuscany (italy). *Phytopathologia Mediterranea*, 39(1):190–205.

Surico, G., Marchi, G., Ferrandino, F. J., Braccini, P., Mugnai, L., et al. (2000b). Analysis of the spatial spread of esca in some tuscan vineyards (italy). *Phytopathologia Mediterranea*, 39(1):211–224.

Surico, G., Mugnai, L., and Marchi, G. (2006). Older and more recent observations on esca: a critical overview. *Phytopathologia Mediterranea*, 45(4):68–86.

Surico, G., Mugnai, L., and Marchi, G. (2008). The esca disease complex. In *Integrated management of diseases caused by fungi, phytoplasma and bacteria*, pages 119–136. Springer.

Tabacchi, R., Fkyerat, A., Poliart, C., Dubin, G.-M., et al. (2000). Phytotoxins from fungi of esca of grapevine. *Phytopathologia mediterranea*, 39(1):156–161.

Tabbagh, A., Dabas, M., Hesse, A., and Panissod, C. (2000). Soil resistivity: a non-invasive tool to map soil structure horizonation. *Geoderma*, 97(3):393–404.

Thébaud, G., Peyrard, N., Dallot, S., Calonnec, A., and Labonne, G. (2005). Investigating disease spread between two assessment dates with permutation tests on a lattice. *Phytopathology*, 95(12):1453–1461.

Turechek, W. W. and McRoberts, N. (2013). Considerations of scale in the analysis of spatial pattern of plant disease epidemics. *Annual review of phytopathology*, 51:453–472.

Van Leeuwen, C., Friant, P., Chone, X., Tregoat, O., Koundouras, S., and Dubourdieu, D. (2004). Influence of climate, soil, and cultivar on terroir. *American Journal of Enology and Viticulture*, 55(3):207–217.

Van Leeuwen, C., GOUTOULY, J.-P., PERNET, D., de RESSEGUIER, L., and FRIANT, P. (2011). Spatialisation of vine water and nitrogen status at the estate level or at the block level. *Proceedings 17th international Symposium GIESCO*, pages 255–258.

Van Leeuwen, C., Tregoat, O., Choné, X., Bois, B., Pernet, D., Gaudillère, J.-P., et al. (2009). Vine water status is a key factor in grape ripening and vintage quality for red bordeaux wine. how can it be assessed for vineyard management purposes. *J. Int. Sci. Vigne Vin*, 43(3):121–134.

Viala, P. (1926). *Recherches sur les maladies de la vigne*. Revue de Viticulture.

Wackernagel, H. (2003). *Multivariate geostatistics*. Springer.

Walter, C., McBratney, A. B., Douaoui, A., and Minasny, B. (2001). Spatial prediction of topsoil salinity in the chelif valley, algeria, using local ordinary kriging with local variograms versus whole-area variogram. *Soil Research*, 39(2):259–272.

Wang, Z. and Zheng, Y. (2013). Analysis of binary data via a centered spatial-temporal autologistic regression model. *Environmental and Ecological Statistics*, 20(1):37–57.

Weber, E. A., Trouillas, F. P., and Gubler, W. D. (2007). Double pruning of grapevines: a cultural practice to reduce infections by eutypa lata. *American journal of enology and viticulture*, 58(1):61–66.

White, C.-L., Halleen, F., and Mostert, L. (2011). Symptoms and fungi associated with esca in south african vineyards. *Phytopathologia Mediterranea*, 50(4):236–246.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.

Xu, X., Harwood, T. D., Pautasso, M., and Jeger, M. J. (2009). Spatio-temporal analysis of an invasive plant pathogen (phytophthora ramorum) in england and wales. *Ecography*, 32(3):504–516.

Zanzotto, A., Gardiman, M., Serra, S., Bellotto, D., Bruno, F., Greco, F., and Trivisano, C. (2013). The spatiotemporal spread of esca disease in a cabernet sauvignon vineyard: a statistical analysis of field data. *Plant Pathology*, 62(6):1205–1213.

Zheng, Y. and Zhu, J. (2008). Markov chain monte carlo for a spatial-temporal autologistic regression model. *Journal of Computational and Graphical Statistics*, 17(1).

Zhu, J., Huang, H.-C., and Wu, J. (2005). Modeling spatial-temporal binary data using markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2):212–225.

Zhu, J., Zheng, Y., Carroll, A. L., and Aukema, B. H. (2008). Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood. *Journal of agricultural, biological, and environmental statistics*, 13(1):84–98.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470.