Comment coupler observations et prédictions pour améliorer les prédictions d'épidémie de Septoriose sur le blé ?

Lygie Esquirol

AGROCAMPUS OUEST
CFR Rennes*

65 rue de Saint-Brieuc
CS 84215

35042 RENNES CEDEX

INRA_ Institut National de la Recherche en Agronomie

78850 THIVERVAL-*GRIGNON*

ARVALIS–Institut du Végétal
IBP – Université Paris Sud
Rue de Noetzlin – Bât. 630
91405 – ORSAY CEDEX

Mémoire de Fin d'Etudes*****

Diplôme d'Ingénieur de l'Institut Supérieur des Sciences Agronomiques, Agroalimentaires, Horticoles et du Paysage *

Année universitaire* : 2011.-2012.

Spécialisation : Statsitique

Comment coupler observations et prédictions pour améliorer les prédictions d'épidémie de Septoriose sur le blé ?

Par : Lygie ESQUIROL

Bon pour dépôt (version définitive) ❑

Date ; …./…/… Signature :
Autorisation de diffusion : Oui ❑ Non❑

***Devant le jury** :*

Soutenu à Rennes le* :11/ 09/ 2012

Sous la présidence de : Jérôme PAGES*

Maître de stage* : David MAKOWSKI, Davis GOUACHE.

Enseignant référent : David CAUSEUR

Autres membres du jury (Nom, Qualité) :

Diffusion du mémoire

A remplir par l'auteur avec le maître de stage

Limites de la confidentialité

Mémoire de fin d'études

Consultable sur place : * oui ❑ non

Reproduction autorisée : ❑ oui * non

Prêt autorisé : * oui ❑ non

Confidentialité absolue : ❑ oui * non
(ni consultation, ni prêt)

Durée de la confidentialité : . . . . . .

Fiche de résumé du mémoire de fin d'études :

Résumé diffusable : *oui ❑ non

Fait à : . . .Grignon. . . . . **Le :** . . . . . 22/08/2012 . . . . . .

Le Maître de stage :
Signature et cachet de l'entreprise

INRA AGROPARISTECH
UMR D'AGRONOMIE
Bat EGER
78850 THIVERVAL-GRIGNON

L'auteur :

REMERCIEMENTS.

Je tiens à remercier mes deux maîtres de stage messieurs David MAKOWSKI de l'INRA Versailles- Grignon et David GOUACHE d'Arvalis Institut du végétal pour l'aide qu'ils m'ont apporté tout au long du stage, pour leurs conseils et pour la relecture de mon rapport.
J'ai eu l'occasion de découvrir et d'essayer de nombreuses méthodes différentes tout au long du stage, qui aura été très enrichissant.

Je remercie également mon tuteur monsieur David CAUSEUR, ainsi que tous les enseignants de la spécialité « statistique » pour leurs cours de qualité qui m'ont donné goût aux statistiques !

Table des matières

Introduction :

Mon stage se déroule à l'INRA de Grignon et est réalisé en partenariat avec Arvalis, Institut du végétal. Dans le cadre de celui-ci **nous cherchons à améliorer les prédictions de la septoriose sur le blé, une maladie fongique.** Prédire les risques d'épidémie de manière plus précise permettrait d'utiliser les fongicides à meilleur escient : c'est-à-dire au moment où leur efficacité est maximale et seulement quand ils sont vraiment nécessaires.

La septoriose (Fig. 2,3) est une maladie courante qui nécessite de deux à trois traitements fongicides entre mars et mai. Pour établir à quel moment il doit traiter, l'agriculteur se base, entre autre, sur le Bulletin de Santé du Végétal (BSV) de sa région. Celui-ci est publié toutes les semaines par chaque chambre régionale d'agriculture. Il s'agit d'une fiche d'information gratuite sur l'état sanitaire des végétaux qui a pour but d'aider l'agriculteur à mieux protéger ses cultures et à raisonner l'utilisation de produits phytosanitaires. Il se base sur des observations réelles et sur des modélisations.

En 2010, 2000 Bulletins de Santé du Végétal ont été édités en France. Pour rédiger ces BSV, les experts de la chambre d'agriculture regardent en parallèle des observations obtenues par des mesures sur le terrain et des prédictions de contamination obtenues à partir de modèles épidémiologiques. Les observations terrain se font hebdomadairement. Sur une année, on suit environ 8000 sites répartis sur toute la France. Les réseaux de surveillance nationaux comptent environs 3000 personnes **(@1).** Toutes les observations recueillies depuis 2008 sont stockées dans une base de données nationale appelée Vigicultures® Ces observations sont faites, la plupart du temps, en « fréquence de plantes contaminées » : une note allant de 0, aucune plantes contaminées, à 10, toutes les plantes contaminées. On dispose également de quelques observations faites en intensités, c'est-à-dire le pourcentage de la surface de feuille atteint par la maladie. Le modèle épidémiologique principalement utilisé pour prédire les épidémies de septoriose s'appelle SeptoLIS®, il a été développé par Arvalis l'institut du végétal. Il prédit en intensité, c'est-à-dire en pourcentage de surface de feuille contaminé touché par la maladie.

Ces deux types de données, les prédictions en intensité et les observations de contaminations réelles, sont rarement confrontées. Dans le cadre de mon stage je teste différentes approches de couplage de ces deux types de données, dans le but d'améliorer les prédictions de Septoriose. Plus précisément, la question traitée dans ce rapport est :

- **Peut-on prédire les fréquences de contaminations à l'aide de SeptoLIS et d'autres variables disponibles avant le traitement**

Dans une première partie j'aborderai les enjeux de la prédiction de la septoriose : je détaillerai quelles sont les pratiques de préventions actuelles et dresserai un état des lieux sommaire des modèles existant de prédiction d'épidémies de Septoriose du blé et notamment de SeptoLIS®.
Dans une deuxième partie, je présenterais plus précisément une des stratégies de couplage des données SeptoLIS® et Vigicultures®, qui permet de prédire les fréquences de contamination de Septoriose.
Enfin dans une troisième partie je discuterais les résultats obtenus.

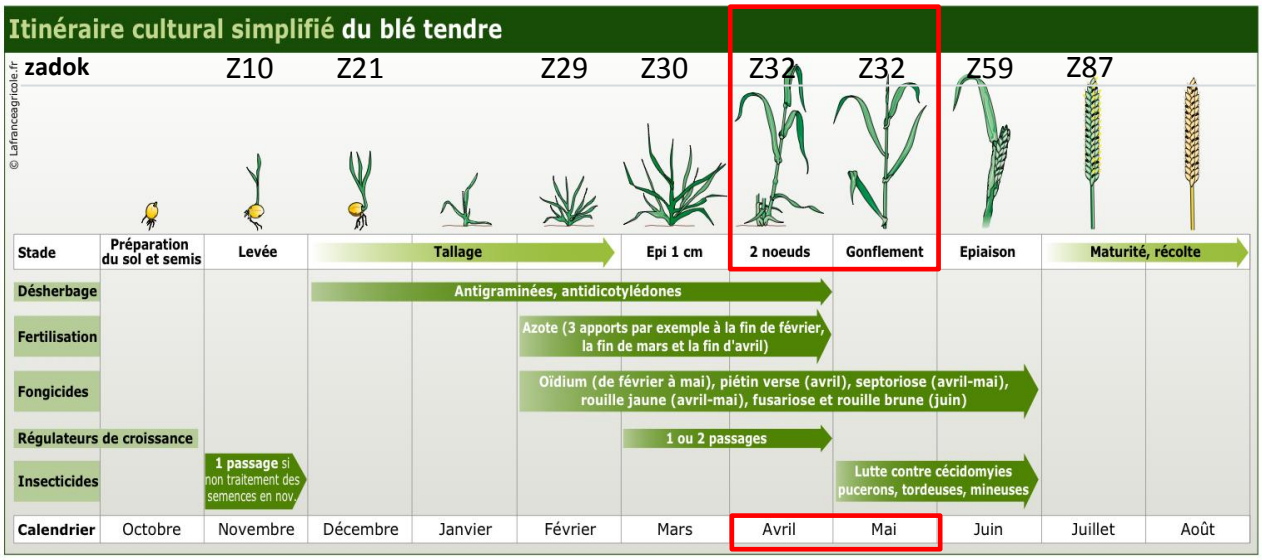


Figure1 : Itinéraire cultural simplifié du blé tendre, dans le carré rouge période où l'on traite contre la septoriose (source : La France agricole @2).

Figure 2 : Photo d'une lésion de septoriose sur une feuille. (source (9))

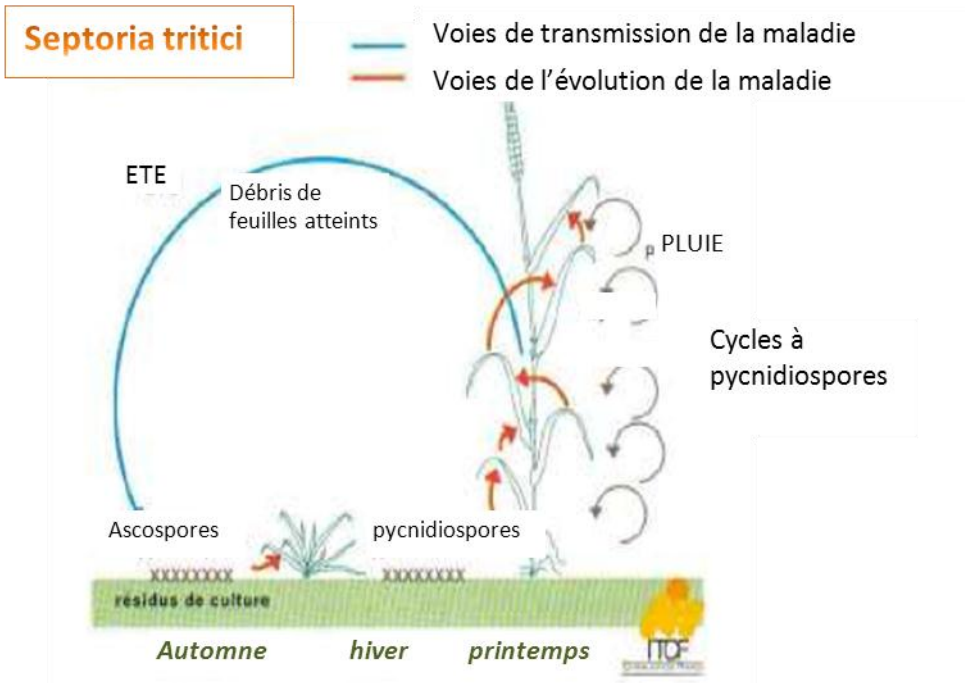


Figure 3 : Schéma du cycle de développement de la septoriose (source (1)).

1. Contexte du stage : Etat des lieux et enjeux de la prédiction d'épidémies de Septoriose (*Septoria tritici*) sur le blé en France.

1.1 La Septoriose du blé, une des principales maladies foliaires du blé en France pouvant causer d'importantes pertes de rendement.

Dans ce paragraphe nous donnerons une vision concrète de la maladie au sein du cycle du blé et des méthodes de luttes actuelles. En quoi la modélisation aide-t-elle dans la lutte contre cette maladie ?

Le blé est semé entre septembre et décembre selon la région. On peut voir en ***figure 1*** un exemple d'itinéraire technique et l'échelle de Zadoks* en ***ANNEXE I.1***, qui décrit à l'aide d'une note les différents stades de développement de l'épi. Les premières feuilles émergent en janvier et la tige commence à pousser en février/mars c'est le stade épis 1 cm (noté Z30 sur l'échelle de Zadoks*). Les dernières feuilles sortent vers avril-mai (stade noté Z41). Les dernières étapes importantes sont l'épiaison (la sortie de l'épi), et le remplissage du grain (notées Z59 et Z87). Cela a lieu vers mai ou juin, juste avant la récolte (juillet-août) (1).

La septoriose est une des maladies fongiques les plus fréquentes dans le monde (2). Elle va contaminer les feuilles et causer des lésions importantes sur celles-ci ***(figure2)***. Les lésions réduisent la surface verte, cela réduit la photosynthèse et impacte négativement la croissance de la plante et donc le rendement final (3,15). Les pertes de rendement peuvent aller jusqu'à 1,5 tonnes par hectares en l'absence de fongicide (4). La septoriose influe, non seulement sur la quantité de blé produit, mais aussi sur sa qualité (5).

Les deux espèces de champignons qui causent la septoriose existent sous forme soit sexuée, soit asexué: *Septoria tritici/Mycosphaerella* graminocola (forme sexuée/ asexuée respectivement) et *Stagonospora nodorum/Leptospheria nodorum*.(6)

La contamination se développe à partir des spores présentes sur les résidus de la récolte précédente, lorsque sont réunies les conditions température (entre 5 /7°C et 25°C) et l'alternance de périodes sèches et humides, qui produisent développement du champignon et dissémination de ses spores(2,6). Celles-ci stoppent leur développement si ces conditions cessent d'être favorables. Suffert et Sache (2) ont montré que selon le type de résidus laissés sur le sol, on obtient des contaminations d'ampleurs très différentes l'année suivante. Si le champignon est sous forme sexuée, il produira des ascospores* qui, transportées par le vent, se déposeront sur les feuilles ; s'il est sous forme asexuée, il produira des pycnidiospores* qui se déposeront sur des feuilles transportées par les éclaboussures de pluie ou « splashing ». Une fois sur la feuille, si les conditions le permettent, il y aura germination du champignon qui pénètrera dans la feuille. Après une période de latence, qui varie de 15 jours à 5 semaines en fonction des conditions, les symptômes apparaîtront sur la feuille sous forme de lésions (2,7). Celles-ci produiront alors de nouvelles spores et continueront de contaminer les autres étages foliaires, la septoriose est ainsi une maladie polycyclique (6), ***(figure 3)***.

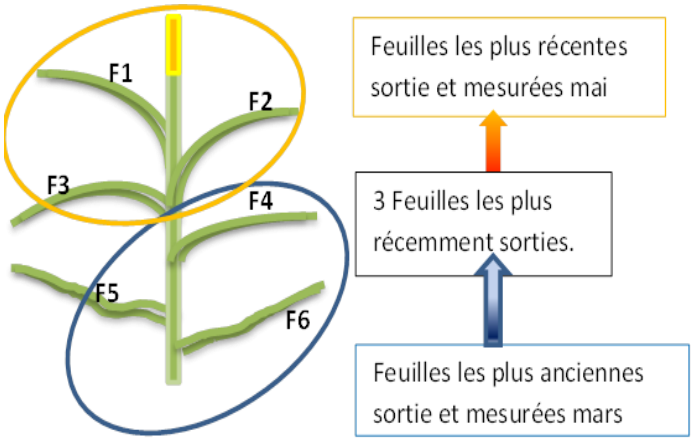
Le traitement de cette maladie se fait par applications de fongicides qui ont pour but de protéger les feuilles les plus impliquées dans le rendement c'est-à-dire les 3 dernières appelées F3, F2 et F1 (dernières feuilles sorties avant l'épi) (8,15). Les pulvérisations se font entre avril et mai ***(figure 1)***. Le traitement se fait en deux applications : une application quand les feuilles F4 et F3 sont sorties, pour contenir la maladie et limiter son développement. Une seconde application quand les feuilles F2 et F1 apparaissent. Le fongicide est le plus actif juste après la contamination. La maladie mettant parfois jusqu'à 5 semaines avant qu'apparaissent les symptômes, il est délicat de déterminer le moment optimum de traitement.

On conseille le traitement lorsque la contamination en fréquence mesuré sur les feuilles 3 atteint le seuil de 4/20, comme on peut le lire dans l'harmonisation des protocoles du réseau surveillance (@6).

OBSERVATION

1-Observation hebdomadaires des contaminations par le réseau de surveillance régional.

2-Un protocole d'observation en fréquence de contamination observée sur 3 niveaux de feuilles successifs.



3-Stockage de l'information observée dans la base de donnée nationale Vigicultures®.

MODELISATION

1-Variables d'entrée du modèle :

-date de semis
-météo

2- Prédiction de contamination en Intensité.

PUBLICATION HEBDOMADAIRE DU BSV

Figure4 : Rédaction du BSV avec des données observées (à gauche) et des prédictions modélisées (à droite). **(SOURCE PERSONNELLE ET @3)**

Dans le numéro de perspective agricole de 2010 et l'extrait de conférence (8,9), Arvalis décrit une expérimentation dans laquelle à cinq applications de fongicides sont réalisées à des moments différents du stade de développement de la plante pour analyser l'impact du positionnement du traitement sur le rendement. Selon le positionnement, la différence de rendement obtenu est de 5 quintaux par hectare. Le bon positionnement du traitement a donc un impact important sur le rendement final.

Dans la pratique la prédiction de la date optimale de traitement est délicate. D'une année sur l'autre, on constate des niveaux de contamination très irréguliers mais aussi des cibles différentes (7). Ainsi certaines années par exemple les variétés précoces seront plus touchées que les tardives et d'autres années ce sera l'inverse. Certaines années, quand la contamination est faible, une seule application au bon moment, suffirait à protéger la plante efficacement.
Le modèle septoLIS®, développé par Arvalis, permet d'estimer la date optimale du premier traitement (4,10). Dans le cas de la septoriose, la maladie étant irrégulière, et le temps de latence très long (durant lequel il est impossible d'observer à l'œil nu l'effet de la maladie), la modélisation devient un outil précieux. Le modèle vise à synthétiser les différentes informations (météo, développement de la plante, quantité d'inoculum initial) afin de prédire l'ampleur de la maladie pour aboutir à une décision pertinente de traitement.

1.2 Des outils de prédiction ont été développés pour prédire et limiter l'impact de la septoriose.

Description des acteurs impliqués dans la prévention du risque phytosanitaire : les réseaux de surveillance et les différents modèles de prévision de la septoriose.

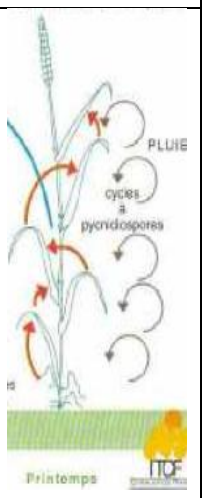
-La prévention du risque septoriose du blé:

Les chambres régionales d'agriculture organisent des réseaux d'observations des risques phytosanitaires et produisent des bulletins de santé du végétal hebdomadaires (B.S.V.) ***(figure 4)***. Ces documents résument les contaminations prédites par un modèle épidémiologique et les données d'observation recueillies. Ils concluent en donnant le risque épidémique sur la région : très élevé, moyennement élevé ou faiblement élevé.

Le modèle septoLIS® est mis à disposition sur une plateforme internet ***(@3)***. Cela permet de prédire de manière journalière, la date de traitement optimale. On peut également sortir des cartes de prédictions d'épidémie de septoriose.

Les observations sont effectuées à partir d'un protocole harmonisé depuis 2008, de manière à pouvoir récolter des données comparables au niveau national dans la base de données Vigicultures ®(11). Il y a deux types de notations, l'un pour les observateurs standards, qui notent en fréquence et l'autre pour les experts, qui notent en intensité. L'observateur doit avoir à sa disposition une zone non traitée et si ce n'est pas le cas il doit marquer le traitement subi par la parcelle.
L'observation « standard » consiste à ramasser 20 plantes. Sur chacune d'elles on note les trois dernières feuilles sorties (les plus récentes) : sur une feuille donnée, si une attaque de septoriose est visible, alors on note 1, sinon 0. Puis on additionne, pour un étage foliaire donné, les chiffres. On trouve une note sur 20 qui est ensuite divisée par deux pour arriver à **une note de fréquence** (aussi appelée **incidence**) sur 10. Si, sur les 20 plantes observées 6, ont de la septoriose sur la feuille numéro 4 « F4 », alors la note de fréquence finale pour la feuille F4 sera 3/10. On fera de même pour les deux autres étages de feuille à observer : par exemple F3 et F2.
La notation « expert » portera sur 60 plantes. Pour chacune, on va regarder à un niveau de feuille donné (F4 par exemple), quel pourcentage de la feuille est couvert par les lésions. On fait une moyenne des pourcentages mesurés sur les 60 plantes. On note alors l'attaque en **intensité**.

Tableau 1 : Résumé des principales équations du modèle SeptoLIS® et de leurs rôles (source : description issue de (4,12) et images (1), @2, PERSONNELLE).

Objectifs :	-Prédire les intensités de maladie sur les 6 dernières feuilles -Décider s'il est l'heure de traiter ou non.	
Inoculum :	Par rapport au modèle d'Audsley (13), SeptoLIS® utilise quelques équations supplémentaires pour **modéliser l'inoculum de départ, jour après jour, en fonction de la pluie et de la température**. *(cf : Equation ANNEXE I.3)* Dans les modèles d'Audsley et pour PROCULTURE, on doit observer la contamination en champ à l'émergence de la F3 et on en déduit rétrospectivement l'inoculum de départ.	
Gradient de disper-sion :	A une date donnée, correspond une quantité d'inoculum donnée. Une série d'équations permettent de **modéliser les contaminations**. L'inoculum va infecter les feuilles supérieures. Une fois la feuille touchée, celle-ci devient **source d'inoculum secondaire pour les feuilles supérieures et elle peut également s'auto contaminer**. La quantité d'inoculum transférée d'une feuille à l'autre va varier entre les différentes feuilles en fonction de **la taille des lésions**, en fonction de la **quantité de pluie tombée**, mais aussi en fonction de la **distance entre les feuilles** qui dépend du **nombre de degré-jours** écoulés. *(cf : Equation ANNEXE I.3).*	
Dynami-que de l'hôte :	Ce modèle tient compte du fait que seule une petite fraction d'infection se développera en fonction de l'hygrométrie et de la température. Il considère le **temps de latence** comme une fonction de la **température** et calcule l'importance de la lésion qui va émerger. *(cf : Equation ANNEXE I.3)* **L'augmentation de la taille de cette lésion est fonction de la croissance de la feuille et donc du nombre de degré jours** écoulés. *(cf : Equation ANNEXE I.3)* Les modifications contenues dans SeptoLIS® sont un **couplage du modèle qui prédit l'émergence des feuilles, l'épiaison et le remplissage du grain** (Gate 1995 (16)), avec celui décrivant le développement de la maladie Il ne tient pas compte de **la résistance du cultivar**. Les variétés cultivées en France ont des notes de résistances trop proches (scores entre 3 et 7 sur une échelle allant de 1 à 9) pour que cela influe sur les prédictions.	
Facteurs environne mentaux :	Le modèle intègre les variables température (en degré jour) et pluviométrie. **Si les conditions hygrométriques et de température ne sont pas suffisantes, les équations modélisent un arrêt du développement de la maladie**. *(cf : Equation ANNEXE I.3)*	

L'intensité de la contamination varie donc de 0 à 100% de la surface de la feuille. Ce type de mesure en intensité est délicat à réaliser demande de l'entrainement et du temps.

Les premiers modèles d'épidémie sur les plantes ont été développés dans les années 60 par Van Der Plank. Aujourd'hui la modélisation d'épidémie est devenue très commune. En faisant un modèle on peut viser différents objectifs : descriptif, prédictif ou conceptuel (12).
Pour la septoriose, les modèles existant visent surtout à prédire la contamination. Les principaux modèles existant sont presept, septoLIS® ou encore PROCULTURE *(tableau comparatif en ANNEXE I.2)*. En France on utilisait Presept, créé dans les années 80, jusqu'à ce qu'ARVALIS développe SeptoLIS®. En Belgique et au Luxembourg, on utilise plutôt PROCULTURE. Ces modèles sont tous des modèles mécanistes qui prennent comme paramètres d'entrée au minimum la météo, la date de semis.

-Le modèle septoLIS® un outil de prédiction:

SeptoLIS est principalement utilisé comme un outil d'aide à la décision (OAD). Il préconise les dates du premier traitement en fonction de la variété, de la date de semis et de la région de France concernée. Cet OAD se compose de deux éléments : du modèle qui donne l'intensité (ou sévérité) de contamination d'une feuille et de la règle de décision qui décrète s'il faut traiter ou non. La décision se base sur la valeur d'une variable interne du modèle, non visible par l'utilisateur. Si celle-ci dépasse un certain seuil, alors la décision suggérée sera traitement. On peut voir dans Gibert et al (10), un exemple d'utilisation de SeptoLIS® dans lequel il permet de prédire de manière précise l'intensité de contamination présente sur la feuille 3.

Dans sa construction, SeptoLIS® est aussi un modèle mécaniste qui décrit les processus impliqués dans le développement de la maladie et leur relation à la météo et au développement phénologique de la plante. Dans le cas de la septoriose, la maladie a été étudiée et est bien documentée. Certains modèles décrivent la progression de l'épidémie dans le couvert végétal (10). SeptoLIS® se base sur le modèle d'Audsley publié en 2005 (13), avec quelques modifications (calcul de l'inoculum de départ et intégration d'un modèle de développement de la plante publié dans Gate 1995 (16, 10)).

Le *tableau 1* ci-contre décrit brièvement le principe de septoLIS®, ses paramètres et leurs rôles. Une série d'équations permet de calculer la quantité d'inoculum de départ et sa dispersion, c'est-à-dire s'il contamine ou non la plante en fonction de la météorologiques. Enfin d'autres équations lient le développement de la maladie et la vitesse de développement de la plante, toujours en fonction des conditions météorologiques. Les équations du modèle issu de la publication de Gouache et al (4) sont présentées de manière plus détaillée en *ANNEXE I.3.* La description du modèle reprend les étapes importantes énoncées par Maanen et Zou (12) : objectif du modèle, modélisation de l'inoculum primaire, de la dispersion, prise en compte de la dynamique de l'hôte et des facteurs environnementaux.

1.3 Performances actuelles des modèles :

Les deux modèles principalement utilisés pour prédire les épidémies de Septoriose sur le blé sont SeptoLIS® et PROCULTURE. Ces deux modèles sont plus ou moins basés sur les mêmes principes *(tableau 1)*, mais PROCULTURE commence à prédire à partir de la feuille F3. Sa particularité est d'avoir pour variable d'entrée, la contamination observée dans le champ à la date d'émergence de la feuille F3. SeptoLIS®, lui, prédit les contaminations à partir des seules données météorologiques et de la date de semis.

Tableau 2 : Performance des modèles PROCULTURE et SeptoLIS® d'après la littérature (source 15,10) et d'après nos données.

Nb an : nombres d'année, Nb. Var : nombre de variétés différentes sur lesquels les observations ont été faites. MAE : mean absolute error en % de feuille contaminée. Specificité/sensibilité : dans (10) on teste 3 seuils. On définit une feuille F3 malade si sa surface contaminée>1%,3% ou 5%.

	Nb an	Nb sites	Nb var	MAE%	Spécificité Detect F3%	Sensibilité Detect F3%
PROCULTURE (15)	2	2	2	3-29	90-100	61-100
SeptoLIS®(10)	11	175	28	14	75-82	90
PROCULTURE(10)	11	175	28	8.75	NA	NA
SeptoLIS® d'après nos données.	2	191	47	2.14	62	82

Tableau 3 : (A) Extrait du tableau des Vigicultures® mesurées en intensité, (B) extrait du tableau SeptoLIS® correspondant.

(A) VIGICULTURES® en intensité	précocité variété	feuille	Altitude	…	Observation contaminat intensité
1	5	F5	172	…	2%
mesure ijk	…	…		…	…
1380	5	F2	300	…	20%

(B) SEPTOLIS® intensité	stade	feuille	âge	Prédiction contamination intensité
1	30	F5	19	2%
Prediction ijk	…	…	…	…
1380	54	F2	25	20%

Tableau 4 : (A) Extrait du tableau des Vigicultures® mesurées en fréquences, (B) extrait du tableau SeptoLIS® correspondant.

(A) VIGICULTURES® en fréquence	précocité variété	feuille	altitude	…	Observation contaminat fréquence	Observation contaminat « binarisée » >4/20=1 ou <4/20=0
1	6	F4	180	…	1/20	0
mesure ijk	…	…		…	…	…
26542	4	F3	50	…	20/20	1

(B) SEPTOLIS® intensité	stade	feuille	âge	Prédiction contamination intensité
1	32	F4	6	0%
Prediction ijk	…	…	…	…
26542	60	F3	23	70%

D'après Gilbert et al (10), ces deux modèles ont sensiblement les mêmes performances concernant les prédictions faites en intensité (pourcentage de feuille touchée), même si PROCULTURE est un peu plus performant avec une erreur moyenne de prédiction de l'intensité de 8.75% contre 14% pour SeptoLIS® *(tableau 2)*.

Ces deux modèles prévoient une intensité de maladie (c'est à dire un pourcentage de surface de feuille contaminée), mais on pourrait aussi imaginer les utiliser pour prédire une fréquence de contamination (si oui ou non la feuille est contaminée au-delà d'un seuil choisi, ici on a testé les détections à différents seuils). C'est ce qui est fait dans les deux publications de Gilbert et al (10) et de El Jarroudi (15), ils examinent les performance en terme de sensibilité et de spécificité. Est-ce que le modèle permet la détection de la contamination de la feuille 3 ?

Dans la publication de El Jarroudi (15), où ils évaluent les performances de PROCULTRE, ils observent un fort taux de faux positifs certaines années allant jusqu'à 39%. Il leur semble que PROCULTURE surestime les contaminations en intensité. L'auteur impute cela au fait qu'il y ait un décalage entre la prédiction du modèle et la vitesse de développement de la maladie au champ. Dans Gilbert et al (10), SeptoLIS® est dit surestimer sévèrement des prédictions dans 10% des cas et sous-estimer les contaminations dans 3% des cas. Au vu des premières observations de nos données *(figure 1)*, on retrouve bien le fait que SeptoLIS® semble surestimer les contaminations.

On peut voir, en *tableau 3 A*, un extrait des données issues de Vigicultures®. L'individu statistique est l'observation faite sur un site i, à la date j et au niveau de feuille k. Après tri, nous avons 1380 données mesurées en intensité (% feuilles contaminées). Leur équivalent est disponible en prédiction SeptoLIS® *tableau 3 B*, l'individu statistique est cette fois la prédiction de contamination sur un site i, à la date j et au niveau de feuille k.

Avec ces données, je peux calculer une erreur de prédiction de SeptoLIS® sur un site i, à la date j et pour un niveau de feuille k. On la définit comme la soustraction de la contamination observée et de la prédiction : $\boldsymbol{Erreur_{ijk}=mesure_{ijk}-prédiction_{ijk}}$. Le site i peut varier de 1 à N. Avec N=191, le nombre total de sites mesurés en intensité. La date *j* varie de 1 à Ji. Le nombre de dates auxquelles on a mesuré un site i donné J_i, est au maximum de 9. Au maximum pour un site donné il y a eu 9 dates ; en moyenne un site i donné est mesuré à 3 dates différentes. Le niveau de feuille k, pour un site i donné et à une date j donnée, varie de k à k+2. k prend des valeurs allant de 1 à 4. On aura le plus souvent 3 mesures faites sur 3 étages foliaires différents les plus récemment sortis : les 3 feuilles successives (k_{ij}), $(k_{ij})+1$, $(k_{ij})+2$.

Avec nos données, on trouve une erreur absolue moyenne de 2.15%, l'amplitude des erreurs que j'observe sur mes données est moins importante que celle calculée dans la publication de Gilbert et al (10). Pour ce qui est de la sensibilité on retrouve les résultats observés dans les publications *(tableau 2)*. On peut avoir une sensibilité de 82% pour une spécificité de 62%, on détecte donc bien les plantes malades, mais on range beaucoup de plantes saines dans la catégorie malades.

De nos jours, on essaie de limiter l'emploi des pesticides pour protéger l'environnement et pour limiter l'apparition de phénomènes de résistances d'un champignon aux fongicides (8). Dans le cas de la septoriose une efficacité maximale implique un traitement avant apparition des symptômes. Il est essentiel d'avoir un modèle performant pour prendre des décisions de traitement au bon moment et maximiser l'efficacité du traitement (10).

Tableau 5 : Résumé de 4 scénarios de prédictions : soit on prédit la contamination sur un site sur lequel on a de l'information, soit on prédit à partir de l'information prélevée sur un site voisin ; soit on veut prédire la contamination présente (en semaine J), soit on cherche à prédire la contamination de la semaine suivante (J+1).

Date pour laquelle on veut faire la prédiction.		**Origine de l'information disponible sur les variables, utilisée pour prédire la contamination su site i:**	
		Variables issues du **site i** que l'on cherche à prédire.	Variables issues d'un **site i'**, proche du site i, que l'on cherche à prédire.
	Date j	**SCENARIO 1 :** ***Variables :*** Date j, Site i → ***Contamination ?*** Date j, Site i -prédictions SeptoLIS® - région -feuille -résistance -précocité -âge de la feuille - stade de la plante -altitude *Variable indicatrice* → *Combinaison de variables Modèle glm* → **Probabilité que la contamination soit < 4/20 ou >4/20 ?**	**SCENARIO 2 :** ***Variables :*** Date j, Site i' → ***Contamination ?*** Date j, Site i -prédictions SeptoLIS® - région -feuille -résistance -précocité -âge de la feuille - stade de la plante -altitude -contamination en j *Variable indicatrice* → *Combinaison de variables Modèle glm* → **Probabilité que la contamination soit < 4/20 ou >4/20 ?**
	Date j+1	**SCENARIO 3:** ***Variables :*** Date j, Site i → ***Contamination ?*** Date j+1, Site i -prédictions SeptoLIS® - région -feuille -résistance -précocité -âge de la feuille - stade de la plante -altitude -contamination en J *Variable indicatrice* → *Combinaison de variables Modèle glm* → **Probabilité que la contamination soit < 4/20 ou >4/20 ?**	**SCENARIO 4 :** ***Variables :*** Date j, Site i' → ***Contamination ?*** Date j+1, Site i -prédictions SeptoLIS® - région -feuille -résistance -précocité -âge de la feuille - stade de la plante -altitude -contamination en j *Variable indicatrice* → *Combinaison de variables Modèle glm* → **Probabilité que la contamination soit < 4/20 ou >4/20 ?**

1.4 Stratégies envisagées pour répondre à l'objectif du stage :

On a vu dans cette première partie que dans les bulletins de santé du végétal, on utilisait à la fois des observations de fréquence de contamination (nombre de feuilles contaminées) et SeptoLIS®, qui donne des résultats en intensité de contamination (% de feuille contaminée). Ces deux types d'informations ne sont, à l'heure actuelle, jamais utilisées simultanément ensemble à l'aide d'une méthode formelle. Les experts de la chambre d'agriculture les examinent ensemble pour conseiller sur la nécessité de traiter ou non.
Outre les observations et les prédictions de contaminations, nous disposons dans la base Vigicultures® et dans les sorties de SeptoLIS® de nombreuses autres variables. Celles-ci concernent par exemple : la résistance de la variété, le stade de développement des plantes, l'âge de la feuille observée/prédite. **Au cours du stage, j'ai envisagé principalement deux couplages possibles des données Vigicultures® et SeptoLIS® : l'un pour corriger les erreurs de prédiction de SeptoLIS® (travail présenté en *ANNEXE IV.1* et l'autre pour prédire une fréquence de contamination à partir des simulations d'intensité du modèle *(tableau 4)*.**

Dans ce rapport je ne détaillerais que la seconde idée : A partir des simulations de septoLIS de l'intensité de maladie, est-il possible de prédire si la contamination en fréquence dépassera ou non le seuil de 4/20 ?
Une contamination en fréquence de 4/20 sur les feuilles F3, est le seuil de à partir duquel on préconise le traitement des champs. Quand on traite contre la septoriose, c'est pour protéger les feuilles F2 et F1 de la maladie, ce seuil-là garanti l'efficacité maximale du fongicide **(@6)**. C'est pourquoi notre objectif est de prédire si ce seuil de contamination est atteint sur un site donné appelé i. Une contamination sur un site i est toujours donnée pour une date j précise et pour un numéro de feuille donné k *(tableau 4)*. Nous avons envisagés 4 cas de figure *(tableau 5)* :

- SCENARIO 1 : On considère que l'on dispose d'informations sur le site i en semaine j : on connait la valeur de nos variables (prédiction de l'intensité de maladie obtenue avec SeptoLIS, mais aussi numéro de feuille, âge de la feuille, stade des plantes, résistance de la plante, précocité de la plante, région et altitude) (Tableau 6) et l'on souhaite prédire la probabilité que la contamination en fréquence dépasse 4/20 en cette semaine j sur la feuille considérée.
- SCENARIO 2 : Le plus souvent, on ne dispose pas de l'information pour le site qui nous intéresse (pour lequel on veut appliquer ou non un traitement fongicide). Nous avons donc testé si, à partir de donnée issues d'un autre site i' (raisonnablement proche de notre site i, c'est-à-dire dans la même région administrative) et recueillies en semaine j, on pouvait prédire la probabilité que contamination en fréquence dépasse 4/20 en cette semaine j sur notre site i.
- SCENARIO 3 : Le Bulletin de Santé du Végétal est publié le lundi et il donne des indications sur la semaine à venir. Il serait donc très utile de pouvoir prédire à partir des informations du BSV de la semaine j, non pas les contaminations de la semaine j déjà finie, mais plutôt celles de la semaine suivante j+1.
  Dans le scénario 3, on imagine que l'on connait les valeurs de nos variables sur le site i (dont la contamination en fréquence mesurée en semaine j) et l'on cherche à prédire la probabilité que contamination en fréquence dépasse 4/20 la semaine suivante j+1.
- SCENARIO 4 : Dans ce scénario, on dispose des variables mesurées à la date j sur un site i', situé dans la même région administrative que notre site i à prédire. On cherche à prédire, la probabilité que contamination en fréquence dépasse 4/20 la semaine suivante j+1 sur le site i.

A chaque fois nous avons testé si chacune des variables disponibles, prise comme indicatrice, pouvait nous permettre de bien prédire la contamination. Nous avons également essayé de prédire cette contamination en les combinant, à l'aide d'arbres de décisions et d'un modèle linéaire généralisé binomial.

Tableau 6 : Dictionnaire des variables.

Nom:	Valeurs minimum et maximum:	unité/ modalité:	nature
Variables observées issues de Vigicultures®:			
Date observation j	De mars à juin chaque année	jour/mois/an , transformée en numéro de semaine j.	date
Année	4 modalités	« 2008», «2009», « 2010», «2011 »	qualitative
latitude	Varie de 44,6 (Sud France) à 50 (Nord ~Lille)	degré d'angle projection WGS 84	coordonnée
longitude	Varie de -2,9 (Bretagne) à 6,8. (Est de la Lorraine)	degré d'angle projection WGS 84	coordonnée
altitude	de 1 à 1142	mètre	quantitative
Site i	numéro du site mesuré	1820 modalités	qualitative
résistance de la variété	Note de 4 peu résistante à 7 très résistante	note entre 0 et 9	quatitative
précocité de la variété	Notes de 4,5 peu précoces à 8 très précoce	note entre 0 et 9	quantitative
Feuille k	6 modalités	"F1","F2","F3","F4","F5","F6"	qualitative
Contamination observée en Fréquence	0/20 à 20/20	note entre 0 et 20	quantitative
Variables prédites issues de SeptoLIS®:			
stade	de z30 à z70	échelle de Zadok ***(ANNEXE I.1)***	quantitative
Prédiction de contamination	de 0 à 100%	% de surface de feuille touchée par les lésions	quantitative
date sortie d'un niveau de feuille k	entre février et mai chaque année	J/m/A	date
Variables crées:			
âge de la feuille : Date observation- Date de sortie	de 0 à 100 jours.	jours	quantitative
grandes régions	5 modalités.	"Nord", "Sud", "Ouest", "Est", "Ile de France" ***(ANNEXE II.3)***	qualitative

Tableau 7 : Extrait des notes de contaminations /20 mesurées sur l'étage foliaire 3, plusieurs semaines de suites, sur 11 sites en 2008 et en 2009.

		semaines																				
Année	site	2	3	4	5	6	7	8	9	10	11	12	16	17	18	19	20	21	23	24	25	26
2008	13	NA	4	0	0	8	16	20	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2008	15	2	4	8	2	18	20	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2008	19	NA	0	0	0	0	18	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2008	33	NA	0	0	0	8	14	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2008	61	NA	0	0	8	20	20	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2008	78	NA	2	8	20	20	10	20	20	20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2009	247	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	16	8	16	20	20	20
2009	308	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	2	2	4	6	8	14	NA
2009	309	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	2	2	8	8	14	16	NA
2009	378	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	2	4	20	20	16	20	NA	NA
2009	562	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	6	6	8	16	20	20

2. Matériel et Méthodes utilisées pour prédire des contaminations de septoriose en fréquence.

2.1 Tableau de données et dictionnaire des variables.

Un extrait des données disponibles est présenté dans les ***tableaux 3 et 4*** et le ***tableau 6*** présente un dictionnaire des variables. L'individu statistique est une « contamination $_{ijk}$ » soit mesurée, soit prédite, de septoriose sur une feuille de blé numérotée **(k), en un site donné (i) et à une date donnée (j)**. Cette contamination est notées soit en fréquence* pour les données observées, soit en intensité* pour les données prédites**.** Sur une plante, on considère 6 étages foliaires k. Ils varient de la feuille F6, la plus ancienne sortie, à la feuille F1, la plus récente feuille sortie, juste avant l'épi.

A chaque mesure, en plus de la contamination, l'observateur note **des informations supplémentaires** :

- Le **stade de développement** de la plante, variable quantitative, il est décrit à l'aide de l'échelle de Zadok ***(ANNEXE I.1)***. Ce stade varie entre Z31 (stade 1er nœud*) et Z85 (grain pâteux).
- Les caractéristiques de la variété de blé utilisé (note de **résistance** à la septoriose, note de **précocité**). Elles sont toutes les deux considérées comme des variables qualitatives.
- Le **précédent cultural***, variable qualitative à 7 modalités (pouvant être soit blé, colza, autre et NA, pois, maïs, pomme de terre ou orge)**,** connue pour influencer le degré de contamination **(2)**.
- D**es informations de type spatiales** : le site et les coordonnées de la commune. Afin de prendre en compte les variations de contamination dues à la zone géographique nous avons créé une variable région qui découpe la France en 5 zones ***(ANNEXE II.3)***.
- Des **informations de type temporelles** : l'année, la date de l'observation et la date de semis.

Pour chaque prédiction, SeptoLIS® calcule en tenant compte des conditions météorologiques, la date estimée de sortie d'une feuille. Il est également capable d'estimer le stade probable de la plante. A partir de la date de sortie de la feuille et de la date où l'observation a eu lieu, j'ai créé une variable quantitative **âge de la feuille à la date d'observation,** en jour. Celle-ci se calcule en soustrayant à la date observation jk, la date de sortie de la feuille j'k estimée par SeptoLIS *: $(âge)_k$=(date observation) $_{j'k}$-(date de sortie) $_{jk}$*.

Au départ dans la base de données Vigicultures® 75882 mesures d'observation de contaminations étaient disponibles. Elles ont été effectuées hebdomadairement, entre mars et juin, depuis 2008 et jusqu'en 2011. Les données recueillies en intensité ne commencent qu'en 2010 et représentent seulement 5514 mesures. Nous avons enlevé les mesures faites sur des plantes traitées aux fongicides, car ils ne sont pas pris en compte dans les calculs de prédictions faits par SeptoLIS®. Nous avons enlevé les mesures dont la variété était inconnues et les observations qui ont été faites au-delà du stade « grain formé (Z 70) », car au-delà de ce stade, les prédictions septoLIS® ne sont plus utilisées. **Il nous reste au final 1380 observations faites en intensité et 26542 en fréquences.**

Selon le protocole Vigicultures®**(11)**, **seulement les 3 feuilles les plus récentes sont notées à chaque observation**. Nous n'avons donc pas d'information sur toutes les feuilles à chaque date, mais seulement sur 3 étages de feuilles consécutifs. Les nombres d'observations de contamination sont donc différents d'une feuille à l'autre et varient aussi d'une période à une autre ***(ANNEXE II.1)***. On note que les observations sur F5 et F6 (feuilles les plus anciennes) sont moins nombreuses que les autres et sont réalisées plutôt entre mars et avril.

D'autre part, un site a été mesuré à différentes périodes pendant une année. En moyenne, pour les mesures **en fréquence, un même site est mesuré 5 fois** en une année. On peut voir ***tableau 7*** un exemple des 11 sites ayant le plus de mesures de contaminations consécutives. Le même site n'est jamais mesuré deux années de suite.

Tableau 8 : Description du critère utilisé pour créer le tableau utilisé pour chaque scénario et sa taille.

Scénario	Critère	Taille
1 : **site i et Date j** **➔site i et Date j ?**	On garde toutes les observations faites en fréquences	**26542** lignes et **27** colonnes
2: **site i' et Date j** **➔site i et Date j ?**	On garde l'information de 5 régions. On garde les sites sur lesquels au moins 3 niveaux consécutifs sont mesurés. Sur chaque période de deux semaines, je note tous les sites d'une région. On se sert de chaque site comme i' successivement pour deviner les autres sites i. *(tableau 9)*	**192 752** lignes et **27*2** colonnes (les variables du site i'sont mises en face de celles autres sites i)
3 : **site i et Date j** **➔site i et Date j+1 ?**	On garde tous les sites sur lesquels on a au moins deux mesures consécutives de trois niveaux de feuilles. A chaque ligne on ajoute une colonne : contamination mesurée en J+1.	**14 339** lignes et **30** colonnes
4 : **site i' et Date j** **➔site i et Date j+1 ?**	On associe, comme dans le scénario 2, un site i'à des sites i, mais on ne garde que les sites ayant au moins deux mesures consécutives.	**67 782** lignes et **30*2** colonnes.

Tableau 9 : Exemple de « duplication du jeu de donné » dans le scénarios 2 .

	Données : variables du site i'	**Données : variables des sites i**
Période : de 01/03/2008 au 15/03/2008	S1	S2
Région : Lorraine	S1	S3
Site ayant 3 mesures: S1, S2, S3 et S4	S1	S4
Taille du tableau initial: 4	…	…
Taille du tableau final:16		
	S4	S1
	S4	S2
	S4	S3

Les prédictions ont été calculées avec le modèle SeptoLIS®. Les variables d'entrée nécessaires au calcul d'une « $prédiction_{ijk}$ » sont : la météo mesurée 7 jours avant la date j et prédites sur 7 jours après cette date et la date de semis de la variété cultivée sur le site i. On peut réaliser les prédictions météorologiques à l'aide de deux méthodes : spatialisée ou non spatialisée. Les écarts de prédictions entre l'une ou l'autre des méthodes sont très faibles ***(ANNEXE II.2)***, mais les prédictions faites avec météo spatialisée font légèrement moins d'erreur. Nous avons donc retenu les prédictions septoLIS® faites avec météo spatialisée.

2.2 Description des Tableaux de données utilisés dans les 4 scénarios.

Chaque scénario a sa spécificité et nos données mesurées n'étant pas forcément très régulière ***(tableau 7)***, nous avons dû trier et adapter le jeu de donnée pour chaque situation. Ici nous allons préciser le contenu de chaque jeu de donnée ***(tableau 8)***

Dans le scénario 1, nous avons gardé dans la base de donnée que les données en observées en fréquence. Cela nous laisse donc 26542 mesures.

Pour le scénario 2, nous voulons pouvoir prédire la semaine J+1 avec de l'information recueillie en J. Nous n'avons donc gardé que les 14339 mesures concernant des sites ayant des mesures consécutives.

Dans les scénarios 3 et 4, il nous a fallu dupliquer le jeu de donnée. En effet on imagine avoir l'information d'un site i' disponible pour prédire plusieurs sites alentours i. Nous avons décidé de nous placer à l'échelle régionale. Nous n'avions pas assez d'information sur toutes les régions pour faire cela, nous avons donc gardé 5 régions ayant beaucoup de mesures : Champagne-Ardenne, Poitou-Charentes, Lorraine, Normandie et Picardie. Pour réaliser la validation croisée j'ai dupliqué mon jeu de donnée.

On peut voir un exemple en ***tableau 9 (ANNEXE II.4 B)***. Si l'on se place en Lorraine, sur une période de deux semaines, j'ai listé tous les sites comptant au moins trois niveaux de feuille d'observation. Imaginons que l'on observe 4 sites. On va prendre tour à tour l'information de chaque site et la considérer comme la seule disponible : ce site joue alors le rôle du « site i' ». Les autres seront, eux, considérés comme les sites à deviner i.

Dans le cas du scenario 2, lors de la « duplication », on se trouve dans le cas où l'on utilise le même site pour observer et prédire (site 1 en j= site 1 en j) ! Cela n'étant pas réaliste j'ai enlevé les cas où le site i'=site i.

Dans le scénario 4 en revanche, Le cas site i= site i' n'est pas irréaliste. Il revient juste à dire que l'on utiliser la contamination mesurée dans le champ i la semaine précédente pour prédire le même site la semaine suivante (site 1 en j= site 1 en j+1). J'ai donc laissé les cas où site i'=site i.

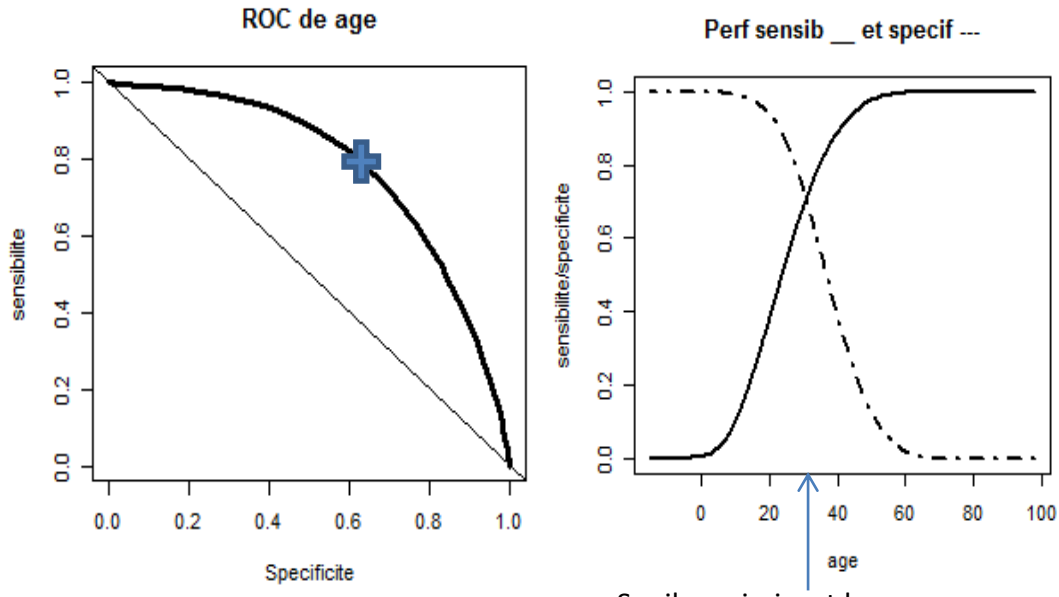


Figure 5 : Courbe ROC de la variable indicatrice «âge de la feuille » (à gauche), courbe de sensibilité (___) et spécificité (- - -) (à droite). La flèche indique le seuil maximisant la somme de la sensibilité et de la spécificité pour la variable âge de la feuille.

Tableau 10 : Matrice des confusions. (Source (20))

		Décision optimale (réalité)	
		Appartenance réelle R=0	*Appartenance réelle R=1*
Décision prise avec notre règle de décision	*Décision : Y=0 si (valeur de la variable) < seuil s.*	*Vrais négatifs*	*Faux positifs*
	Décision : Y=1 si (valeur de la variable) >= seuil s.	*Faux positifs*	*Vrais positifs*

- Sensibilité =P (Y=1|R=1)=1-P(D=0|R=1)
  = P(Indicatrice>Seuil s|R=1).
- Spécificité : =P (Y=0|R=0)=1-P(D=1|R=0)
  = P(Indicatrice<Seuil s|R=0).

Figure 6 : Définition de la sensibilité et de la specificité. (Source (20))

2.3 La courbe ROC, un outil pour évaluer la capacité de discrimination de chaque variable.

2.3.1 Le principe :

L'analyse à l'aide de courbe ROC (Receiver Operating Characteristics) permet d'optimiser une règle de décision binaire. Ces courbes sont souvent utilisées dans le cadre de la protection des cultures (24). On choisit un indicateur : une variable, qui, si elle dépasse un seuil fixé, permet de prendre la décision de traitement. Afin d'optimiser le taux de bon classement on teste ensuite différents seuils s et dans notre exemple de la ***figure 5***, le seuil choisi maximise la somme de la sensibilité et de la spécificité. On aurait tout aussi bien pu choisir un seuil plus bas conduisant à un taux de faux négatif plus faible. (20). L'aire sous la courbe (AUC) maximum ***(figure 5)*** est un critère souvent utilisé pour évaluer et comparer les performances de différents indicateurs (Réf). Il est indépendant du seuil s choisi et est compris entre zéro et 1. Un classement aléatoire est caractérisé par une AUC de 0.5.

Nous allons utiliser cette méthode pour voir quelles performances de classement peut être atteinte en utilisant chacune de nos variables, prise comme indicatrice. Nous allons tester les variables quantitatives suivantes : âge de la feuille, stade de la plante, résistance de la plante, sensibilité de la plante à la septoriose, altitude (influant sur la contamination à cause des différences de température) et la contamination en intensité prédite par SeptoLIS®. On cherche la variable qui permettra d'atteindre une AUC maximum.

On cherche à prédire si, sur un étage foliaire donné k, d'un site donné i à une date donnée j on dépasse ou non le seuil de contamination de 4/20. La variable contamination observée (fréquence de feuilles infectées), Yijk, a été transformée en variable binaire. Si la contamination est supérieure à 4/20, alors sa modalité sera « 1 », sinon elle sera «0 ». Une variable indicatrice, par exemple «âge de la feuille », associée à un seuil s, ici 35 jours, permet de classer une feuille k parmi deux catégories. Si l'âge de la feuille est supérieur à 35 jours, alors on la classera dans la catégorie contaminée « 1 ». L'efficacité de notre classement peut être réalisée en répétant cette opération sur l'ensemble des données en établissant la matrice des confusions ***(tableau 10),*** et en calculant la sensibilité et la spécificité (définies en ***figure 6***) pour l'ensemble des seuils s possibles.

2.3.2 Application de cette méthode à nos quatre scénarios.

Nous nous plaçons toujours dans les 4 cas de figures détaillés en partie 1.4. Je vais maintenant prendre un exemple de classification avec la variable âge de la feuille choisie comme indicatrice dans nos 4 scénarios.

- SCENARIO 1 : J'utilise l'âge de la feuille mesuré sur un site i à une date j et un étage foliaire k. Quelles sont les performances de mon classement pour classer la contamination Yijk ?
- SCENARIO 2 : Je dispose de l'âge de la feuille mesuré sur un site i' à une date j et un étage foliaire k. Quelles sont mes performances de classement pour classer la contamination Yijk d'un site voisin?
- SCENARIO 3 : Je dispose de l'âge de la feuille mesuré sur un site i à une date j et un étage foliaire k. Quelles sont mes performances de classement pour classer la contamination Yi(j+1)k de la semaine suivante sur le même site?
- SCENARIO 4 : Je dispose de l'âge de la feuille mesuré sur un site i' à une date j et un étage foliaire k. Quelles sont mes performances de classement pour classer la contamination Yi'(j+1)k de la semaine suivante sur un site voisin?

Tableau 11 : Définition des trois fonctions de lien utilisées. Avec Φ la fonction de répartition de la loi normale (Source (20)).

Fonction de lien	***Fonction :***	***Probabilité : π =P(Y=1 | $X_{1=x}$,…,$X_{p=x'}$)***
logit	$\log\left(\frac{\pi}{1-\pi}\right)=\mu+X_1+\ldots+X_p$	$\pi = \frac{\exp(\mu + X1 + \cdots + Xp)}{1 + \exp(\mu + X1 + \cdots + Xp)}$
probit	$\Phi^{-1}(\pi)= \mu+X_1+\ldots+X_p$	$\pi = \Phi(\mu + X1 + \cdots + Xp)$
cauchy	$Tan\,(\pi\,(\pi - \frac{1}{2}))= \mu+X_1+\ldots+X_p$	$\pi = \frac{1}{\pi}\arctan(\mu + X1 + \cdots + Xp) + \frac{1}{2}$

2.4 Les arbres de décision: une méthode combinant nos variables indicatrices pour prendre une décision.

L'arbre de décision est une méthode qui permet de visualiser graphiquement la classification d'individus en fonction de différentes variables quantitatives ou qualitatives (23). Dans notre cas on va chercher à classer une feuille dans la catégorie contamination >4/20 ou non, à l'aide de nos variables.

Cette méthode est basée sur un algorithme séquentiel, basé sur CART (22), qui construit des classes d'individus. Les classes sont construites grâce à des règles binaires elles-mêmes construites à partir des variables explicatives, de manière à ce que nos classes soient les plus homogènes possible du point de vu de la variable d'intérêt (23) Il semble intéressant d'appliquer cette méthode après avoir observé les performances de nos variables individuellement avec les courbes ROC.

Nous l'appliqueront à nos 4 scénarios.

2.5. Le modèle glm un outil permettant de combiner plusieurs variables dans le but de prédire le niveau de contamination.

2.5.1 Principe :

Comme écrit dans Agresti (21) et Makowski et Monod (20), le modèle linéaire généralisé permet de décrire le lien existant entre une variable catégorielle Y et des variables d'entrées X1,… Xp. Il n'existe pas de relation linéaire directe, mais il existe une transformation de l'espérance de la variable d'intérêt Y, qui permet d'exprimer celle-ci comme une combinaison linéaire des variables d'entrées. Cette transformation se fait par l'intermédiaire d'une fonction de lien.

J'ai utilisé le modèle de binomial linéaire car on cherche à expliquer une variable binaire Y_{ijk}=1 ou 0 : la contamination d'une feuille k d'un site i à une date j donnée est-elle supérieure à 4/20 (modalité 1) ou inférieure à 4/20 (modalité 0). Nous l'expliquons grâce à la combinaison linéaire de nos différentes variables : âge de la feuille, stade de la plante, résistance de la plante, sensibilité de la plante à la septoriose, altitude, contamination en intensité prédite par SeptoLIS®, numéro de la feuille, région, l'année et parfois la contamination mesurée. D'un point de vue formel, cela revient à supposer que Y suit une loi de Bernouilli de paramètre π_{ijk}, où $\pi_{ijk} = E(Y|X_1,\ldots,X_p)$ est égal à la probabilité que Y=1 (20).

Nous avons testé trois différentes fonctions de lien : logit, probit ou cauchy. Chaque fonction permet de modéliser l'effet des variables X1,… Xp, sur la probabilité que Yijk=1, probabilité qu'il y a ait une contamination supérieure ou égale à 4/20. Pour chacun de nos 4 scénarios, on évalue les performances de prédiction, en regardant l'aire sous la courbe ROC (AUC) obtenue avec les 3 fonctions de lien présentées en *tableau 11*.

Tableau 12: Enoncé des différents modèles glm utilisés. Pour chaque scénario nous avons testé les 3 fonctions de lien, nous désignons donc cette fonction par la lettre G.

Scénario :	Enoncé des modèles :
1 : on prédit la contamination du site i, avec les infos du site i.	$G(\pi_{ijk})=\beta_0^{(1)}+ Feuille_k^{(1)}+Région_i^{(1)}+ Annee_j^{(1)}+ \beta_1^{(1)}\ \underline{stade}X_{1ij} + \beta_2^{(1)}\underline{age}X_{2k} + \beta_3^{(1)}\underline{Resist}X_{3i} + \beta_4^{(1)}\underline{Precocit}X_{4i} + \beta_5^{(1)}\underline{Altitude}X_{5i}$
2 : on prédit la contamination du site i, avec les infos du site i'.	$G(\pi_{ijk})=\beta_0^{(2)}+ Feuille_k^{(2)}+Région_{i'}^{(2)}+ Annee_j^{(2)} + \beta_1^{(2)}\underline{stade}X_{1i'j} + \beta_2^{(2)}\underline{age}X_{2k} + \beta_3^{(2)}\underline{Resist}X_{3i'} + \beta_4^{(2)}\underline{Precocit}X_{4i'} + \beta_5^{(2)}\underline{Altitude}X_{5i'} +\beta_6^{(2)}\underline{contamination}X_{6i'jk}$
3 : on prédit la contamination de la sem j+1, avec les infos de la sem j.	$G(\pi_{i(j+1)k})=\beta_0^{(3)}+Feuille_k^{(3)}+Région_i^{(3)}+ Annee_j^{(3)}+\beta_1^{(3)}\ \underline{stade}X_{1ij} + \beta_2^{(3)}\underline{age}X_{2k} + \beta_3^{(3)}\underline{Resist}X_{3i} + \beta_4^{(3)}\underline{Precocit}X_{4i} + \beta_5^{(3)}\underline{Altitude}X_{5i} +\beta_6^{(3)}\underline{contamination}X_{6ijk}$
4 : on prédit la contamnation de la sem j+1 et site i, avec les infos de la sem j et du site i'.	$G(\pi_{i(j+1)k})=\beta_0^{(4)}+Feuille_k^{(4)}+Région_{i'}^{(4)}+ Annee_j^{(4)}+\beta_1^{(4)}\underline{stade}X_{1i'j} + \beta_2^{(4)}\underline{age}X_{2k} + \beta_3^{(4)}\underline{Resist}X_{3i'}+ \beta_4^{(4)}\ \underline{Precocit}X_{4i'}+ \beta_5^{(4)}\underline{Altitude}X_{5i'} +\beta_6^{(4)}\underline{contamination}X_{6i'jk}$

2.5.2 Enoncé des différents modèles glm utilisés:

Nous avons sélectionné le meilleur modèle sur le critère AIC à l'aide de la fonction stepAIC du package MASS à partir du modèle complet. Dans chaque cas, le modèle avec toutes les variables a conduit à l'AIC le plus faible. Il a donc été conservé ***(ANNEXE II.4)***.

Pour chaque scénario, on cherche à estimer $\pi = E(Y|X_1,\ldots,X_p)$ égal à la probabilité que y=1, mais les informations avec lesquelles on cherche à prédire sont différentes, il y a donc des variations dans l'énoncé des modèles ***(tableau 12)***. La signification de chaque terme du modèle se trouve ci-dessous :

-π_{ijk} est la probabilité que Y=1 à un site i, une date j et une feuille k.

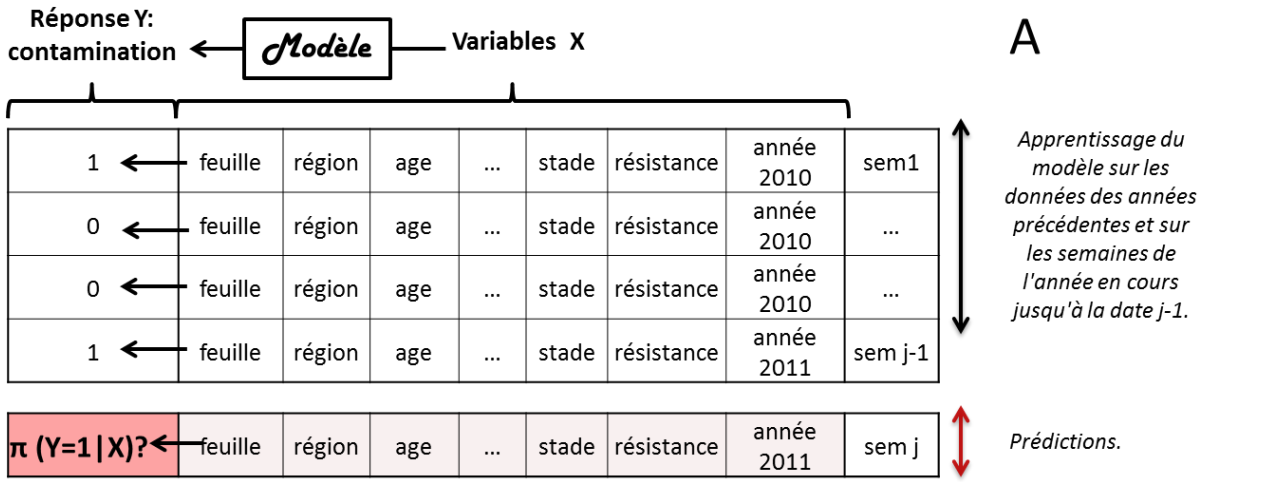
-β_0 est l'intercept : soit la moyenne globale de la probabilité de contamination.

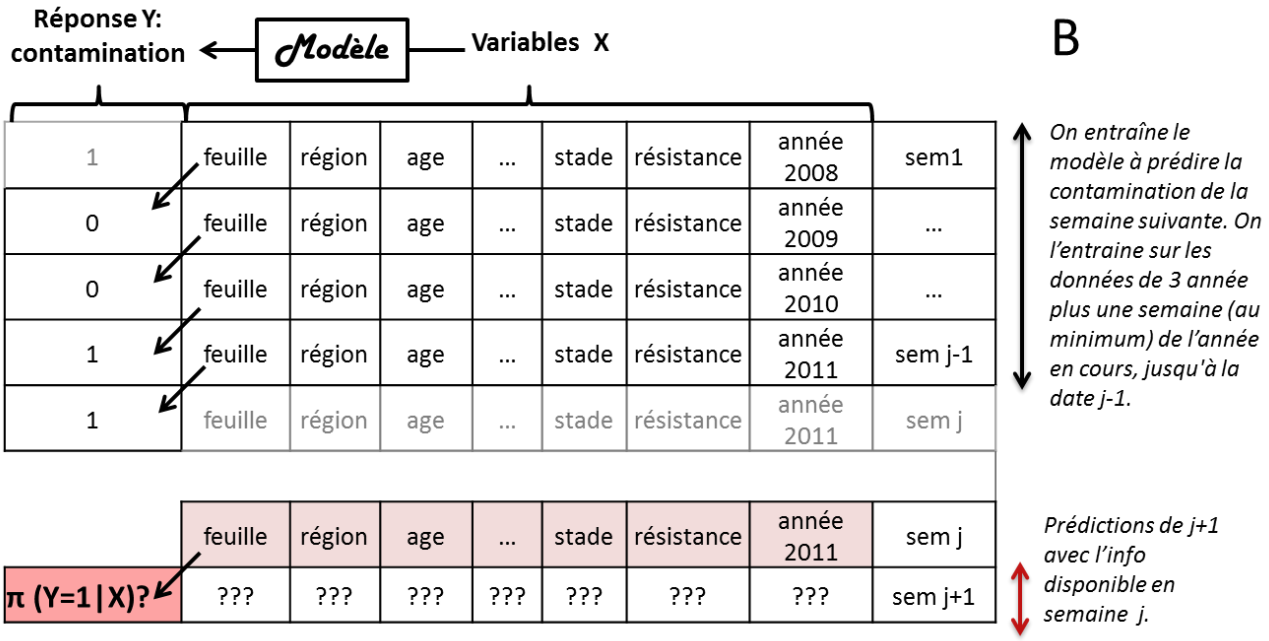
-$Feuille_k$ est **l'effet étage foliaire** k sur la probabilité de contamination. $Région_i$ est **l'effet région** du site i sur la probabilité de contamination et $Année_j$ **l'effet année** sur la probabilité de contamination. D'une année à l'autre l'amplitude des contaminations varie, nous estimons donc toujours l'effet année, en utilisant au minimum les données mesurées sur la première semaine de l'année.

-β_1 est le coefficient de **l'effet « stade de la plante »** X_{1ij}, à la date j, sur le site i. β_2 est le coefficient de **l'effet « âge de la feuille »** X_{2k}, sur la probabilité de contamination ; Il est spécifique d'un étage foliaire k. β_3 est le coefficient de **l'effet « résistance » de la plante à la septoriose** X_{3i}. β_4 est le coefficient de **l'effet « précocité » de la plante** X_{4i}. β_5 est le coefficient de **l'effet « altitude » du site** X_{5i}. β_6 est le coefficient de **l'effet « prédiction de la contamination par SeptoLIS® »** X_{6ijk} sur la probabilité que le taux de contamination soit supérieur ou égal à 4/20. Ces prédictions SeptoLIS® sont faites en intensité pour le site i à la date j sur un étage foliaire k. Enfin β_7 est le coefficient de **l'effet « contamination de septoriose »,** elle est mesurée en fréquence sur le site i à la date j sur un étage foliaire k, X_{7ijk}.

Chacun des 4 modèles ***(tableau 12)*** correspond à nos 4 scénarios.

Tableaux 13 : A) Illustration de la validation croisée utilisée dans le scénario 1, où l'on prédit une semaine j avec les informations de la semaine j B) Illustration de la validation croisée utilisée dans le scénario 3, où l'on prédit une semaine j+1 avec les informations de la semaine j.





2.6 Validation croisée utilisée pour tester nos capacités prédictives en conditions réelles.

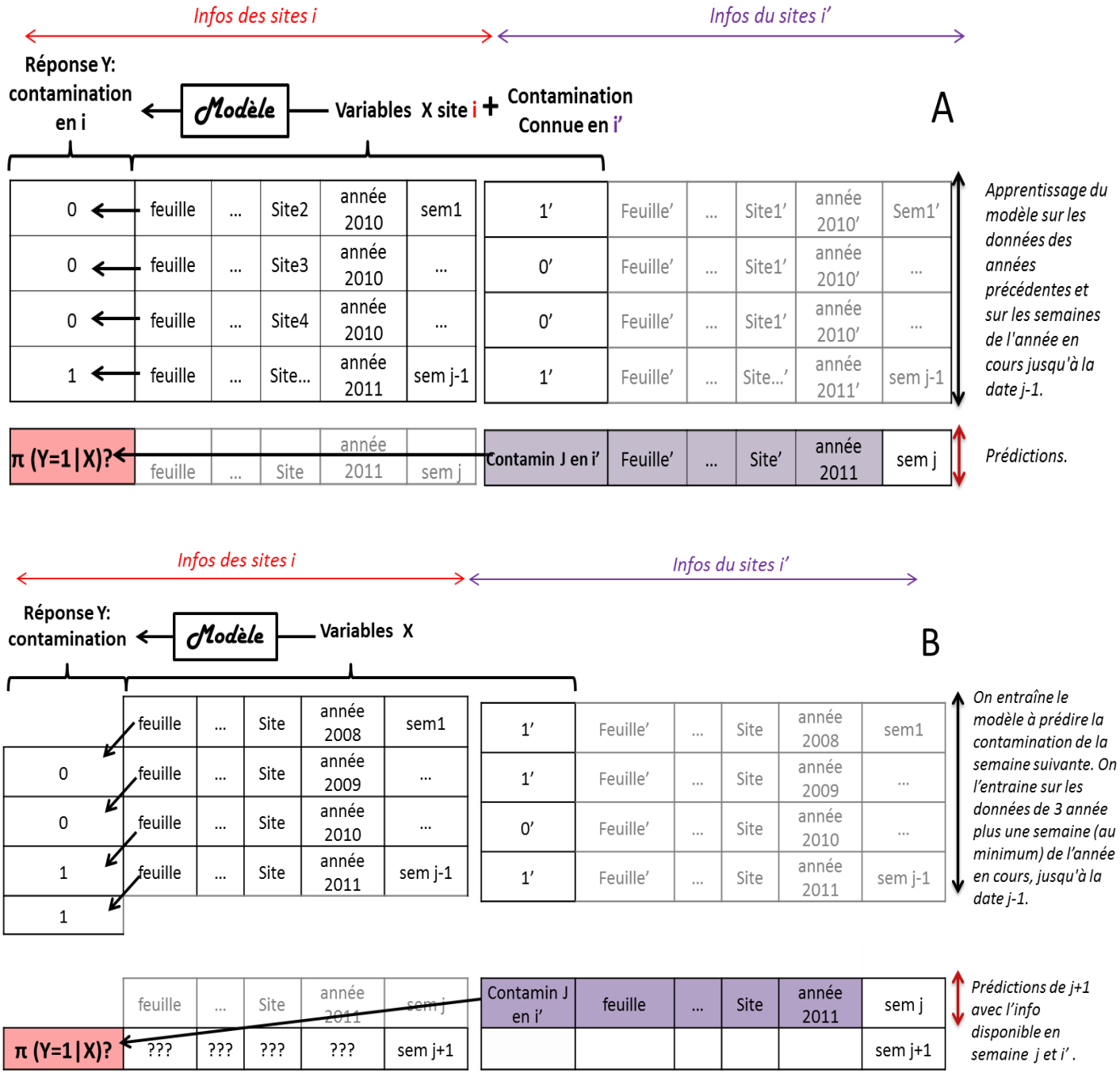
L'utilisation des mêmes données pour estimer les valeurs des paramètres et pour calculer la sensibilité et la spécificité conduit à une surestimation des performances d'un modèle **(20)**. Pour éviter cela on va réaliser une validation croisée plus réaliste, c'est à dire que l'on va estimer les paramètres du modèle sur une partie des données et prédire sur l'autre partie.
Dans le cas de nos données, nous disposons de 4 années d'informations. Il y a un ordre chronologique à respecter.

2.6.1 Validation croisée des scénarios 1 et 3 : observations et prédictions sur un seul et même site i :

- SCENARIO 1: On souhaite prédire π_{ijk} de la semaine j, à l'aide des valeurs prises par les variables en cette même semaine. On entraine donc le modèle sur les informations de 3 années complètes (par exemple 2008, 2009, 2010) plus une semaine de l'année en cours (dans l'exemple ***tableau 13 A*** : la première semaine de l'année 2011, notée semaine j-1). On fait cela dans le but de pouvoir estimer toutes les modalités de l'effet année, on veut aussi pouvoir estimer celui de 2011. Une fois l'apprentissage fait on peut prédire la probabilité de contamination pour la semaine suivante, exclue de l'apprentissage du modèle, dans notre exemple il s'agit des valeurs des variables de la semaine j. Au tour suivant on recommence l'apprentissage du modèle sur 2008, 2009, 2010 et les deux premières semaines de 2011, puis on prédit la semaine suivante exclue du modèle : la troisième semaine de 2011. Ainsi de suite jusqu'à la fin de 2011.
Une fois 2011 fini, on prédira une autre année. Par exemple on va utiliser 2009, 2010, 2011 et une semaine de 2008 pour prédire la semaine numéro 2 de 2008…etc.

- SCENARIO 3 : La particularité des scénarios 3 tient dans le fait qu'on prédit la contamination de la semaine suivante j+1, avec les valeurs des variables de la semaine j. Pour cela j'entraine mon modèle à prédire la contamination en semaine j+1, avec les variables de la semaine j ***(tableau 13 B)***. Pendant l'apprentissage par exemple, on associe la variable réponse « contamination » de la semaine 2 avec les valeurs des variables mesurées en semaine 1.
Au niveau de la séparation groupe d'apprentissage et groupe à prédire, je m'arrête à j-1 pour prédire j. On utilise les valeurs mesurées en semaine j et le modèle va prédire la probabilité que la contamination dépasse 4/20 en semaine j+1.

Tableau 14 : A) Illustration de la validation croisée utilisée dans le scénario 2, où l'on prédit une semaine j en i (en rose) avec les informations de la semaine j en i' (en vert) B) Illustration de la validation croisée utilisée dans le scénario 4, où l'on prédit une semaine j+1 en site i avec les informations de la semaine j du site i' voisin.



2.6.2 Validation croisée pour les scénarios 2 et 4 : observations sur un site i'et prédictions sur des sites i voisins:

Dans ces deux scénarios, on ne dispose pas des vraies valeurs des variables du site i, on dispose seulement des informations d'un site voisin i', choisi dans la région de i. Pour réaliser la validation croisée, on se place à l'échelle d'une de nos 5 régions : Lorraine, Poitou-Charentes, Champagne-Ardenne, Picardie et Normandie.

-SCENARIO 2 : On peut voir en ***tableau 14 A***, l'illustration de la validation croisée utilisée. L'entraînement du modèle se fait comme précédemment sur les données disponibles de 3 années complètes (par exemple 2008, 2009, 2010) plus une semaine de l'année en cours. Une des différences de ce scénario est que l'on doit utiliser la contamination du site i' pour estimer l'effet contamination du site i (**β6** dans le modèle ***tableau 12***).
Une fois nos paramètres estimés, j'utilise les valeurs des variables d'un seul site : i' pour estimer π_{ijk} les probabilités de contaminations de tous les autres sites de la région autour. Je décide qu'à tout site i, voisin de i', sera affecté la même $\pi_{i'jk}$ que celle calculée pour le site i'.
Cette situation est très semblable à ce qui se fait à l'heure actuelle. Chaque région publie dans le BSV des informations contaminations concernant quelques sites et l'on en déduit la contamination probable des sites alentours.

- SCENARIO 4 : Le scénario 4 ressemble exactement à la 3 dans l'entraînement du modèle : on prédit la contamination de la semaine suivante j+1, avec les valeurs des variables de la semaine j. ***(tableau 14 B)***. Pour prédire par contre on va utiliser les données de la semaine j et issue d'un site voisin pour prédire la probabilité que la contamination dépasse 4/20 en semaine j+1 au site i.

2.7 Calcul des performances obtenues pour les cas de référence

On dispose, après la validation croisée expliquée en 2.4, d'une évaluation des performances de nos modèles. Mais pour évaluer leur réelle plus-value, il nous faut maintenant les comparer à un scénario de référence. Le scénario de référence est celle d'un agriculteur ou d'un conseiller n'ayant à sa disposition ni courbe ROC, ni modèle linéaire généralisé, mais ayant lu la contamination publiée dans le BSV:
-Référence du SCENARIO 1 : On considère qu'il dispose de la contamination exacte de son champ i en semaine j. La référence de la situation 1 implique d'aller mesurer sur le terrain la contamination exacte.
-Référence du SCENARIO 2 : Dans ce cas on suppose que l'agriculteur ou le conseiller dispose de l'information mesurée sur son site i en semaine j, par exemple 2/20, et avec celle-ci, il en déduit que la contamination gardera cette même valeur la semaine suivante j+1.
-Référence du SCENARIO 3 : En semaine j, on suppose qu'il dispose d'une mesure de contamination de 2/20 faite sur un site de la région i' et il en déduit que cette valeur est valable également dans ces propres parcelles, cette même semaine.
-Référence du SCENARIO 4 : En semaine j, on suppose que l'agriculteur ou le conseiller dispose d'une mesure de contamination de 2/20 faite sur un site de la région (site i') et il en déduit que la contamination gardera cette valeur la semaine suivante j+1 et que cette valeur sera également valable dans ces propres parcelles (sites i).

2.8 Description des outils informatiques utilisés et des scripts rendus.

J'ai principalement utilisés les packages suivants : ROCR et les fonctions prediction et performance, Epi pour les courbes ROC, rpart et randomForest pour les arbres de prédiction, MASS pour les fonctions stepAIC et le package stats pour les fonctions glm et predict.

J'ai rendu 4 scripts qui concernent les 4 scénarios décrits précédemment. Dans chacun on a une partie tri du tableau réarrangement des données, une partie prédiction avec les indicatrices et les courbes ROC, une partie Sélection de modèle et validation croisée.

J'ai aussi rendu un script « prêt à l'emplois » et son guide d'utilisation. Il permet l'utilisation directe des indicatrices, d'un arbre de décision et d'un modèle pour prédire une contamination en semaine J ou une contamination en semaine J+1.

On trouve en ***ANNEXE II.4*** les schémas des algorithmes principaux utilisés dans ces scripts.

Tableau 15: Récapitulatif des Aires sous la courbe obtenues dans les 4 scénarios avec l'étage foliaire 3.

SCENARIO 1 : J=J i=i	estimé sur 7 754 feuilles		SCENARIO 2 : J=J i'=i	estimé sur 53 204 feuilles	
VARIABLE INDICATRICE			**VARIABLE INDICATRICE**		
		AUC			**AUC**
	SeptoLIS	**0.816**		**SeptoLIS i'**	**0.904**
	Stade	**0.807**		**Stade i'**	**0.915**
	Age	**0.794**		**Age i'**	**0.897**
	resistance	0.468		**resistance i'**	0.487
	precocite	0.532		**precocite i'**	0.548
	altitude	0.472		**Altitude i'**	0.457
REFERENCE=		1	**REFERENCE=**	**contmin J i'**	0.877

SCENARIO 3 : J=J+1 i=i	estimé sur 2 264 feuilles		SCENARIO 4 : J=J+1 i'=i	estimé sur 19 958 feuilles	
VARIABLE INDICATRICE			**VARIABLE INDICATRICE**		
		AUC			**AUC**
	SeptoLIS	0.820		**septoLIS i'**	**0.810**
	Stade	0.835		**stade i'**	**0.814**
	Age	0.846		**age i'**	**0.800**
	resistance	0.482		**résistance i'**	0.494
	precocite	0.492		**précocite i'**	0.525
	altitude	0.421		**altitude i'**	0.497
REFERENCE=	**contamin J**	0.857	**REFERENCE=**	**contmin J i'**	0.765

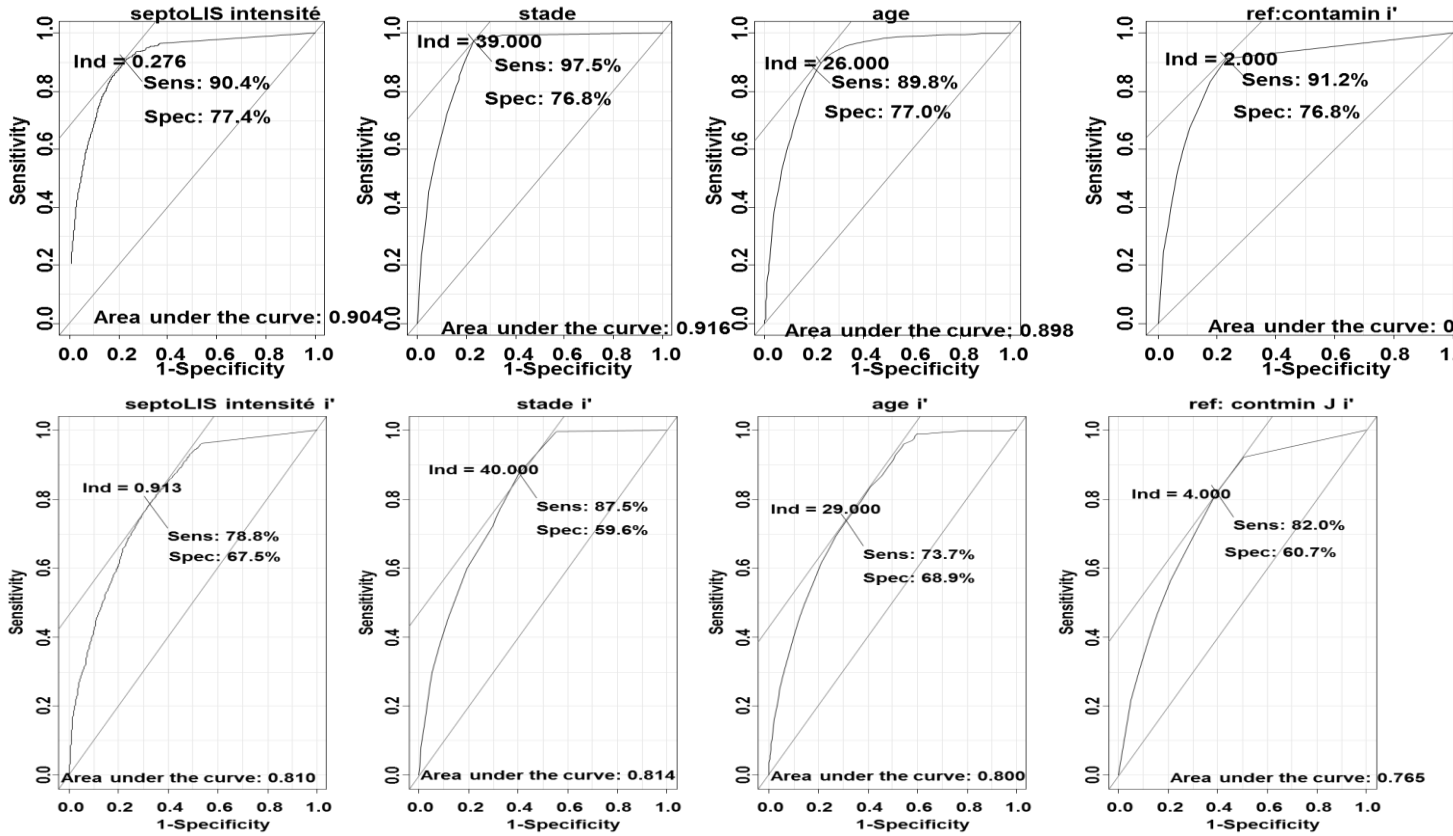


Figure 7: Courbes ROC et courbes de sensibilité spécificité pour les situations 2 (en haut) et 4 (en bas). Les graphiques de gauches à droite sont : septoLIS®, stade, âge et la référence.

3. Résultats:

3.1 Le classement par variables indicatrices :

Dans le ***tableau 14***, on peut voir les résultats des prédictions faites avec les variables indicatrices. Pour chaque variable, on calcule un seuil qui permet de classer une feuille, soit dans la catégorie contamination supérieure à 4/20 soit dans la catégorie contamination inférieure à 4/20. L'observation de l'aire sous la courbe (AUC) permet d'avoir une idée de la performance de chacune de ces variables pour classer notre variable d'intérêt.
On tient surtout à bien prédire les contaminations sur l'étage foliaire 3. En réalisant les analyses mais seulement avec ces feuilles 3, on trouve de meilleures performances que si l'on analyse toutes les feuilles ensembles. ***(ANNEXE3.1 B)***.

Quand on observe les AUC des indicatrices dans le scénario 3 ou dans le 1, on peut voir qu'elles sont toutes moins bonnes que les AUC obtenues avec les scénarios de référence : 85.7% cas 3 et 100% pour le cas 1. Il est cependant rare et coûteux en temps d'aller mesurer les contaminations dans tous les champs. On a donc rarement la contamination exacte présente dans notre champ une semaine avant, comme nous le supposons dans le cas référence des scénarios 3 ou 1, si c'est le cas on peut se passer des indicatrices !
Si en revanche on se place dans le cas plus fréquent où notre information provient d'un site voisin (scénarios 2 et 4), on peut voir que les variables « prédiction SeptoLIS® »,« âge de la feuille » et « stade » sont les plus performantes pour classer les contaminations sur les feuilles 3 que les références. Dans le cas 2 par exemple, on atteint une AUC de 91.5% avec l'indicatrice « stade » contre une référence avec un AUC de 88%. Nos performances restent cependant proches des performances obtenues par le scénario de référence.

On peut voir en ***figure 7***, les courbes ROC et les sensibilités et spécificité. L'indicatrice stade est la plus performante pour classer dans les cas 3 et 4. On observe que par rapport à leurs références respectives, on obtient de meilleures AUC. Quand on compare les taux d'erreurs trouvés grâce à « stade », on atteint les mêmes performances que les référence en termes de spécificité mais que notre sensibilité est meilleure.

On trouve les résumés des sensibilités et de spécificités dans le ***tableau 16.***
En général on observe que le score de sensibilité, la proportion de feuilles malades parmi les feuilles prédites comme malade, est plus important que le score de spécificité, la proportion de plantes réellement saines parmi les feuilles prédites en dessous de 4/20. On peut cependant choisir un seuil qui permettrait de mieux détecter les plantes malades si nécessaire. Pour le cas 1, avec la variable stade, on atteint une sensibilité de 96%. Les plantes contaminées sont très bien diagnostiquées, mais beaucoup d'entre elles sont considérées infectées à tort puisque la spécificité n'est que de 56%.

Tableau 16: Récapitulatif des seuils, des pourcentages de sensibilité et de spécificité des meilleures indicatrices pour nos 4 scénarios, d'après la fonction ROC du package Epi.

	cas 1:			cas 2:		
variable	seuil	sensib	specif	seuil	sensib	specif
SeptoLIS	0.730	79.6	69.4	0.276	90.4	77.4
stade	39.00	96.0	56.2	39	**97.5**	**76.8**
age	28.00	85.7	61.0	26	89.8	77.0
contamin ref		**1**	**1**	2.00	91.2	76.8

	cas 3:			cas 4:		
SeptoLIS	0.008	78.5	75.8	0.913	78.8	67.5
stade	39	72.5	79.9	40	**87.5**	**59.6**
age	20	81.9	72.1	29	73.7	68.9
contamin ref	4	**69.2**	**95.6**	4	82.0	60.7

Figure 8: Importance des variables pour les arbres de décisions. A gauche cas 1, à droite cas 3.

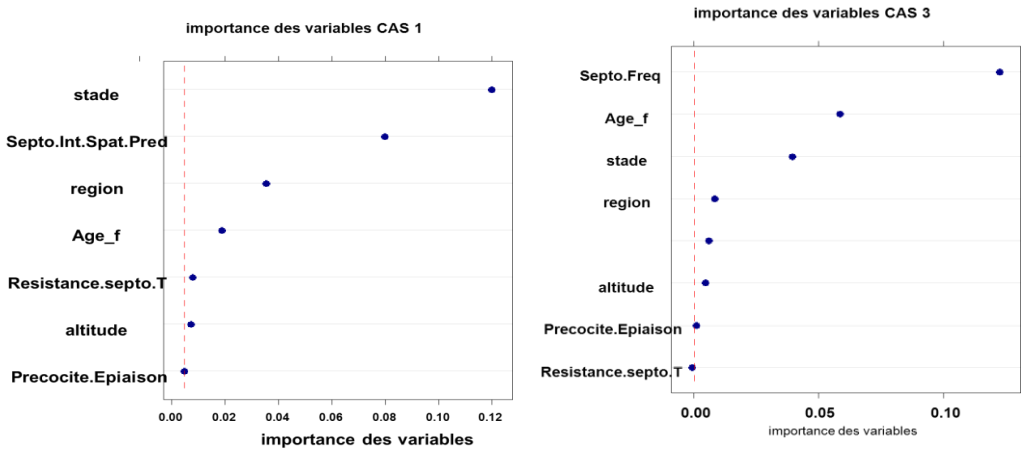
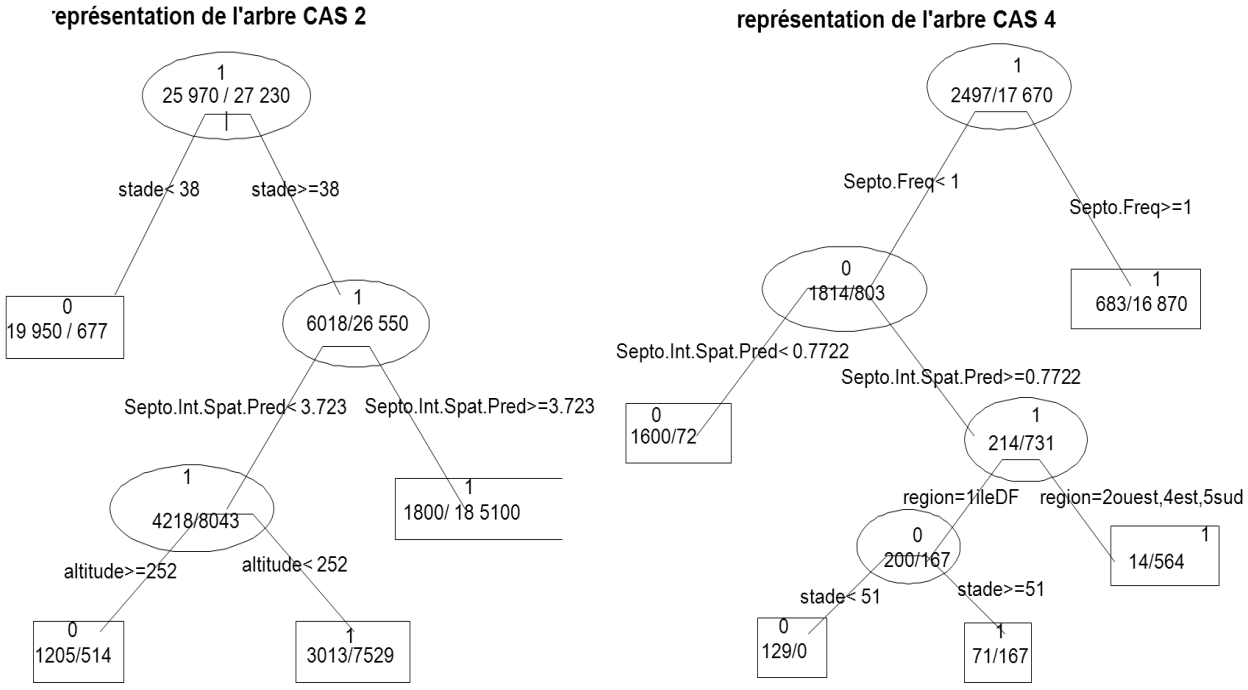


Figure 9: Arbre de décision des cas 2 et 4.



Dans le cas du scénario 3, on peut voir que connaître la contamination présente dans son propre champ, et supposer qu'elle restera la même la semaine suivante, permet d'assez bien prédire. 88% des plantes déclarées malades le sont réellement et 76% de plantes saines sont bien diagnostiquées.
Pour les scénarios 2 et 4, où l'on utilise les informations d'un champ voisin, mieux vaut utiliser la variable stade qui donne les meilleures sensibilités pour la même spécificité. Pour le scénario 2, au-delà du stade Z39, on considèrera la feuille contaminée, en deçà comme non contaminée. On atteint 98% de sensibilité et 77% de spécificité. Pour le scénario 4, on peut considérer que la contamination dépassera 4/20 la semaine suivante si le stade atteint dépasse Z40. On atteint des performances de détection un peu moins bonne que dans le cas 2 : 88% de sensibilité et seulement 60% de spécificité. Ceci est dû au fait que l'on prévoit la contamination de la semaine suivante. ***(tableau 16)***

Les indicatrices utilisées sont calculées à partir des sorties de SeptoLIS®. L'observation du stade de développement d'une plante n'est pas toujours chose aisée, le programme septoLIS® la calcule donc à partir de la date de semis et des conditions météorologiques. La variable Âge est elle aussi calculée grâce à septoLIS®. SeptoLIS® est donc être assez efficaces pour prédire une contamination la semaine même ou même une semaine à l'avance.

3.2 Le classement à partir d'arbres de décisions :

Après avoir noté les performances de nos indicatrices, il semble intéressant d'essayer de combiner nos différentes variables à l'aide des arbres de décision, afin de voir si cela améliore nos capacités de prédictions. Comme dans le cas des prédictions par indicatrice, nous avons réalisé des arbres sur toutes les feuilles et seulement à partir des feuilles 3. Les scores de classifications sont meilleurs si l'on ne considère que l'étage foliaire 3.

Comme les arbres de décisions sont assez instables, c'est-à-dire que des changements légers dans les données peuvent parfois avoir de grosses conséquences, il est intéressant de regarder les variables concurrentes qui auraient pu être choisi à chaque nœud pour classer les feuilles. Nous avons également réalisé les forêts aléatoires de Breiman (23)) dans les cas 1 et 3 pour avoir une idée des variables les plus importantes dans chacun des cas. ***(figure 8) (ANNEXE3.2)***. Dans le cas 1, on retrouve les variables stade et septoLIS® comme étant les plus utile pour classer, puis la variable « région », suivie de l'âge de la feuille. Dans le cas 3, où l'on prédit la semaine suivante, l'information de contamination de la semaine précédente joue un rôle primordial dans le classement, suivie de l'âge et du stade.

Nous allons maintenant examiner l'arbre de décision du cas 2, où l'on prédit les contaminations à l'aide de données mesurées sur un champ proche. La première variable à regarder est le stade. Si celui-ci est inférieur à 38, alors notre feuille est en dessous du seuil avec une probabilité de 96%, si cela dépasse Z38, alors on va regarder SeptoLIS®, si celui-ci prédit au-dessus de 3.72%, alors la feuille est classée parmi les contaminées (91% de cette nœud sont effectivement bien contaminés). Si SeptoLIS® est inférieur à 3.72% on peut alors observer l'altitude si l'on est en deça de 252 mètres, alors on classe la feuille dans contaminée (71 % des feuilles de ce nœud le sont effectivement) et sinon notre feuille est considérée comme saine (70% des feuilles le sont effectivement) ***(figure 9 gauche).***
Pour l'arbre correspondant au scénario 4, on utilise les données d'un champ voisin à la date J pour prédire la contamination en J+1 dans un autre champ régionalement proche. Si, en semaine J, on a une prédiction septoLIS® au-dessus de 3.12%, alors la feuille est classée comme dépassant 4/20, 97% des feuilles de ce nœud le sont. Sinon on peut regarder le stade, si celui-ci est en dessous de Z32 rangera notre feuille dans la catégorie saine (80% des feuilles le sont), sinon une altitude au-dessous de 233 mètres et un âge au-delà de 18.5 jours classe notre feuille en 1 (seulement 55% des feuilles de ce nœuds sont contaminées).

Tableau 17: Récapitulatif des seuils, des pourcentages de sensibilité et de spécificité des arbres de décisions pour nos 4 scénarios.

	Scénario 1:		Scénario 2:		Scénario 3:		Scénario 4:	
		REF		REF		REF		REF
sensibilité	70	100	84	42	88	54	93	66
spécificité	81	100	94	86	73	84	65	69
taux d'erreur	24	0	12	21	17	19	10	31

Tableau 18: Récapitulatif des AUC des glm pour nos 4 scénarios. On peut aussi y voir les performances si l'on ne considère les que les prédictions faites sur l'étage foliaire 3.

SCENARIO 1			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		32 341
1 logit	0.830	0.833	23 954
1 probit	0.830	0.833	**23 953**
1 cauchit	0.829	0.829	24 331
MODELE 1 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	0.838	0.832	
1 probit	**0.838**	**0.832**	
1 cauchit	0.836	0.827	
REFERENCE			
J connu…	1	1	

SCENARIO 2			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		253 282
1 logit	0.877	0.871	162 887
1 probit	0.877	0.872	**162 785**
1 cauchit	0.876	0.866	166 361
MODELE 2 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	**0.934**	0.922	
1 probit	0.933	**0.923**	
1 cauchit	0.934	0.915	
REFERENCE			
	0.884	0.876	

SCENARIO 3			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.500		18 775
1 logit	0.889	0.905	**11 263**
1 probit	0.887	0.904	11 376
1 cauchit	0.890	0.906	11 251
MODELE 3 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	**0.897**	**0.879**	
1 probit	0.896	0.877	
1 cauchit	0.898	0.881	
REFERENCE			
J connu…	0.857	0.859	

SCENARIO 4			
	AUC	**AUC ac CV**	**AIC**
modele nul			92 896
1 logit	0.943	0.852	**40 639**
1 probit	0.942	0.854	41 208
1 cauchit	0.943	0.847	41 204
MODELE 4 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	**0.963**	**0.804**	
1 probit	0.962	0.807	
1 cauchit	0.963	0.799	
REFERENCE			
J connu…	0.771	0.769	

En revanche, si l'âge de la feuille est inférieur à 18.5 jours et que la mesure n'est pas réalisée dans l'ouest de la France, cette feuille est classée dans les saines (61% des feuilles de ce nœud le sont). Sinon on la considèrera comme malade (61% des feuilles de ce nœud le sont). ***(figure 9 à droite)***.

Si l'on regarde maintenant les scores de sensibilité et spécificités atteints grâce aux arbres de décisions dans les 4 scénarios, on peut voir que l'on a amélioré la sensibilité par rapport aux performances obtenues avec les indicatrices seules. Cependant cela se fait au détriment de la sensibilité par exemple pour le scénario 3. ***(tableau17)***.

Pour le cas 1, on n'a pas beaucoup amélioré la sensibilité puisqu'on atteint 70%, mais on reste à 81% de sensibilité. Pour ce qui est du scénario 2, on obtient encore une fois une bonne spécificité 84% et 94% en sensibilité, la performance est tout juste meilleure que celle atteinte par l' indicatrice « stade », qui arrive à une sensibilité de 98%., tout en gardant une spécificité de 77% . Pour le scénario 3 en revanche, nos performances sont meilleures que celles de la plupart des indicatrices. Mais utiliser la référence, c'est-à-dire décider que notre champ aura la même contamination en semaine J+1, que celle présente en J, reste le meilleur moyen de bien classer la feuille, avec 70%en sensibilité et 96%en spécificité ***(tableau16)***. Pour le scénario 4, l'arbre apparait avoir de meilleure performances que nos indicatrices. On augmente et en sensibilité et en spécificité, même si cette dernière reste assez faible : seulement 65% des plantes détectées comme le sont, beaucoup d'entre elles sont considérées infectées à tort.

3.3 Le classement par modèles linéaires généralisés.

On retrouve en ***ANNEXE III.3*** les coefficients des modèles les plus performants et leurs odd ratios**.** On peut voir que tous les modèles semblent être relativement bien ajustés au vu des pvalues < 0.05 des anova ***(ANNEXE III.4)***. On peut également voir les résidus de nos modèles en ***ANNEXE III.5***.

Pour ce qui est des AUC obtenues par les modèles linéaires généralisés, on obtient de meilleurs résultats qu'avec les seules indicatrices ***(tableau 18 et 19)***. A chaque fois on regarde les performances des modèles avec 3 fonctions de liens différentes sans validation croisée, puis avec. Nous avons aussi regardé l'AUC obtenu si l'on n'estime les paramètres à partir des seules feuilles 3, car on base généralement la décision de traitement sur l'observation de leurs contaminations. Contrairement à ce que l'on avait observé pour les indicatrices et les arbres, on obtient de meilleurs résultats en conservant toutes les feuilles. ***(ANNEXE III.1B)***.

Pour choisir le meilleur modèle prédisant la probabilité de contamination d'une feuille au-delà du seuil de 4/20, nous regardons le critère d'Akaïké (AIC) et l'AUC en situation de validation croisée.

Pour les scénarios où l'on prédit la même semaine que l'on fait l'observation (cas 1/ 2), le modèle avec la fonction de lien probit semble le plus adapté avec une AUC de 87% (contre 79% en référence) pour le scénario 2. Pour ce qui est des scénarios 3 et 4, où l'on souhaite prédire l'avenir, le modèle avec cauchit est le meilleur pour le scénario 3. Si l'on considère le cas de figure où l'on prédit avec les données d'un site voisin régionalement éloigné (scénario 4), le modèle avec probit est le meilleur. Ces deux scénarios sont deux cas extrêmes, l'un où on utilise son propre champ pour prédire et l'autre où l'on utilise un lieu voisin parfois très éloigné. Dans les deux cas, la fonction de lien logit a une performance très proche du maximum et son AIC est le plus faible. Nous allons donc garder la fonction de lien logit.

Figure 10: Probabilités de contamination des Feuille 3, calculées par validation croisées, en fonction des contaminations réelles.

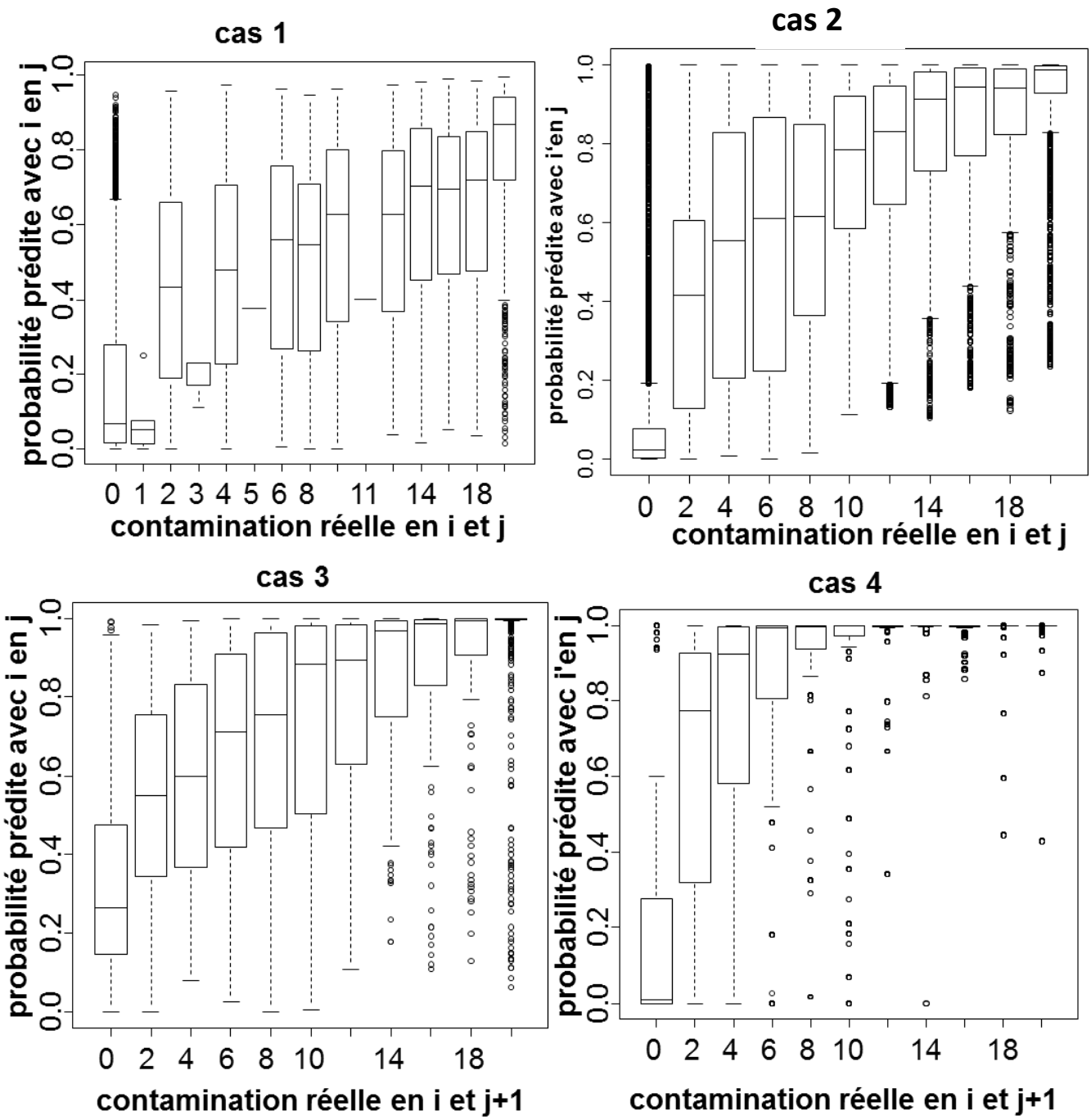


Tableau 19: Récapitulatif des sensibilités et spécificités trouvées avec les glm (en pourcentage), par la fonction ROC du package Epi.

	Scénario 1					
	sans CV	avec CV	REF	sans CV F3	avec CV F3	REF
sensibilité	81.3	76.1	1	81.3	73.7	1
spécificité	69.3	74.3	1	67.0	77.5	1

	Scénario 2					
	sans CV	avec CV	REF	sans CV F3	avec CV F3	REF
sensibilité	85.4	76.4	78.7	92.2	**84.3**	91.2
spécificité	75.9	80.9	73.2	79.6	**84.5**	76.8

	Scénario 3					
sensibilité	76.4	75.2	68	72.8	72.3	69.2
spécificité	85.3	89.5	90.8	92.3	91.7	**95.6**

	Scénario 4					
sensibilité	84.5	76.6	73.4	83.5	**84.5**	82.0
spécificité	90.9	78.7	78.3	97.6	**62.4**	60.7

En ***figure 10***, on trouve les graphiques des probabilités de contamination des feuilles 3 obtenues par validation croisée, en fonction des contaminations réelles observées. Dans les cas 1 et 2, on trouve les probabilités de contamination de la semaine J sont en fonction des fréquences de contamination réellement observées cette même semaine. Dans les cas 3 et 4, on observe les probabilités estimées de contamination de la semaine suivante (J+1) en fonction des contaminations ayant eue lieue en J+1. La fréquence de contamination observée étant une variable discrète, on l'a représentée à l'aide de boxplot.
On note que dans l'ensemble les probabilités augmentent bien régulièrement avec la fréquence de contamination effectivement observée. Pour le cas 4, nos probabilités semblent augmenter beaucoup plus vite que les contaminations, on doit surestimer les contaminations. On remarque également qu'il y a une forte variabilité au niveau de nos probabilités estimées, comme l'indiquent les quantiles 0% et 100% des boxplots. Ils coïncident avec 0 et 1 exactement, sauf pour les fréquences extrêmes comme 0/20-1/20 et 19/20-20/20.

On peut observer les performances du seuil choisi par la fonction ROC et examiner les scores du ***tableau 19***. Ils sont calculés pour toutes les feuilles, mais aussi plus spécifiquement pour les F3. Pour le scénario 1, on observe que l'on a une détection des feuilles contaminées ou non contaminées assez moyenne. Les scores restent dans le même ordre de grandeur qu'avec les autres méthodes.
Dans le cas 2 en revanche, les scores de sensibilité et spécificité atteints sont de 84%. Notre AUC est supérieure à celui de la référence et on atteint des performances en spécificité et sensibilité supérieures à celles obtenues par l'indicatrice stade. Cependant il semble qu'avec l'arbre on puisse atteindre une spécificité de 94% tout en gardant une sensibilité à 84%.
Pour ce qui est des performances du scénario 3, il semble qu'utiliser la référence ou notre modèle ait les mêmes performances.
Enfin pour le scénario 4, on observe un AUC mais aussi des scores de sensibilités et de spécificités bien meilleurs qu'avec l'indicatrice ou qu'avec l'arbre. On remarque que la spécificité est assez faible. Cela confirme bien l'observation faite dans la ***figure 10***, on a tendance à surestimer, il y a probablement beaucoup de faux positif.

Dans les scénarios 2 et 4, il semble que l'utilisation des outils, l'arbres (pour le scénario 2) et modèle glm (pour le 4), ait une vraie plus-value. Combiner de l'information sur la contamination existante aux autres variables, même si l'information vient d'un champ voisin, améliore donc de beaucoup nos performances de prédictions.

Dans le scénario 1 et dans le 3, cas où l'on a des informations concernant sa propre parcelle, on observe en revanche une faible plus-value de nos outils. Dans le scénario 3, le modèle glm arrive à dépasser de justesse la performance de référence, obtenue avec la contamination présente sur le champ la semaine précédente.

4. Discussion :

Les meilleures variables indicatrices sont issues des sorties SeptoLIS® et sont le stade, l'âge de la feuille et l'intensité de contamination. D'après les AUC et dans l'exemple pour lequel nous avons calculé les spécificités et sensibilité, on peut voir qu'elles sont plus performantes que les références dans les scénarios 2 et 4, c'est-à-dire quand on se place dans une situation où l'on prédit une contamination à l'aide d'un champ voisin. Les arbres sont une méthode de classification graphique assez intuitive, dans le scénario 2 il atteint une meilleure performance de classement que le modèle glm. Les modèles linéaires généralisés offrent la meilleure performance si l'on tient compte de l'aire sous la courbe ROC. Au grès du choix du seuil on pourra donc optimiser la spécificité pour être sûr de détecter le maximum de feuilles 3 contaminées au-dessus de 4/20, tout en gardant une sensibilité raisonnablement forte.
Dans le cas des indicatrices et des arbres de prédiction, il est nécessaire de n'utiliser que l'information disponible sur l'étage foliaire 3 pour avoir de meilleures capacités prédictives. En revanche pour les modèles, on utilise l'information disponible sur toutes les feuilles pour optimiser les performances des prédictions de contamination de l'étage foliaire 3.

Grâce à nos 4 scénarios, nous avons essayé de tester les performances de nos trois outils de manière la plus réaliste possible. Les scénarios 2 et 4 représentent des situations dans lesquelles nous n'avons pas beaucoup d'information sur les parcelles à prédire. Pour réaliser la duplication du jeu de données des scénarios 2 et 4, nous avons choisi une échelle régionale, mais celle-ci est discutable. Le choix de l'échelle influe beaucoup sur les performances du modèle. Dans la réalité il est plus probable qu'une échelle plus petite, départementale, voire même communale soit plus réaliste.

Nous avons dû réarranger le jeu de donnée de manière à pouvoir vérifier si nos prédictions à j+1 (basées sur des observations faites en J) étaient exactes ou non. Cela a conduit à un grand tri de données, pour ne laisser que les sites mesurés consécutivement pour au minimum trois étages foliaires. Il est donc possible que cela ait conduit à un biais.

Enfin il semble que sur les régions du nord potentiellement très fortement contaminées telles que la Picardie, au-delà d'un certain seuil de contamination en fréquence, le protocole change et les mesures sont réalisées en intensité. Cela a également dû créer un biais dans nos résultats, même si nous avons créé une variable « région » pour y remédier.

Nous comparons toujours nos performances à des situations de référence. Dans celle-ci on ne tient compte que de la connaissance de la contamination mesurée en fréquence pour conclure sur la contamination future. Dans la réalité, des experts prédisent sûrement avec plus de performance, mais sans méthode formelle. Le mélange formel d'informations issues de relevés physiques et de résultats de modélisation de SeptoLIS® semble être assez prometteur et pourrait constituer une aide supplémentaire synthétisant les différents types d'information. Grâce aux modèle linéaire généralisé, il est possible de prédire le dépassement d'un seuil de contamination, ici 4/20, une semaine à l'avance.

On pourrait imaginer aller encore plus loin dans la prédiction des probabilités de contamination. Ainsi, au lieu de se fixer un seuil et de vouloir savoir si oui ou non celui-ci est dépassé, on pourrait connaître la probabilité le risque d'apparition de chacune des fréquences de contamination possibles. Les fréquences sont discrètes et vont de 0/20 à 20/20, si on considère chaque note comme une modalité de la variable réponse, on peut faire de la régression ordinale logistique à risque cumulé. On aurait ainsi le risque d'occurrence de chaque note, par rapport au risque d'être inférieure. Cela donnerait une vision plus précise des risques, une semaine à l'avance.

D'autre couplages sont encore possible entre les données Vigicultures® et les modélisations septoLIS®. SeptoLIS® prédit les contaminations en intensité, mais fait parfois des erreurs, surtout quand les contaminations prédites sont importantes. Nous avons essayé de prédire mieux les pourcentages de feuilles contaminées, en rajoutant des informations issues d'autres variables, notamment de Vigicultures®. Nous avons ainsi un peu amélioré les prédictions de SeptoLIS® ***(ANNEXEX IV.1)***. Nous ne disposons jamais simultanément de contaminations observées à la fois en fréquence et à la fois en intensité. Mais si nous l'avions, il est probable que l'on pourrait corriger les prédictions septoLIS® de manière encore plus précises.

Conclusion

Nos objectifs étaient de prévoir les contaminations en fréquences à partir de variables issues de la modélisation septoLIS® et d'observations faites en champ. Nous voulions prédire les contaminations, soit pour la semaine en cours, soit pour la semaine à venir. Pour ce faire nous avons testé 3 outils différents: les variables indicatrices, des combinaisons de variables dans des arbres de décision ou des combinaisons de variables par le modèle linéaire généralisé. Nous avons cherché à nous placer dans des conditions les plus proches possibles de la réalité, en créant 4 scénarios différents, afin d'attester de la réelle plus-value de nos outils de prédictions par rapport à des situations de référence.

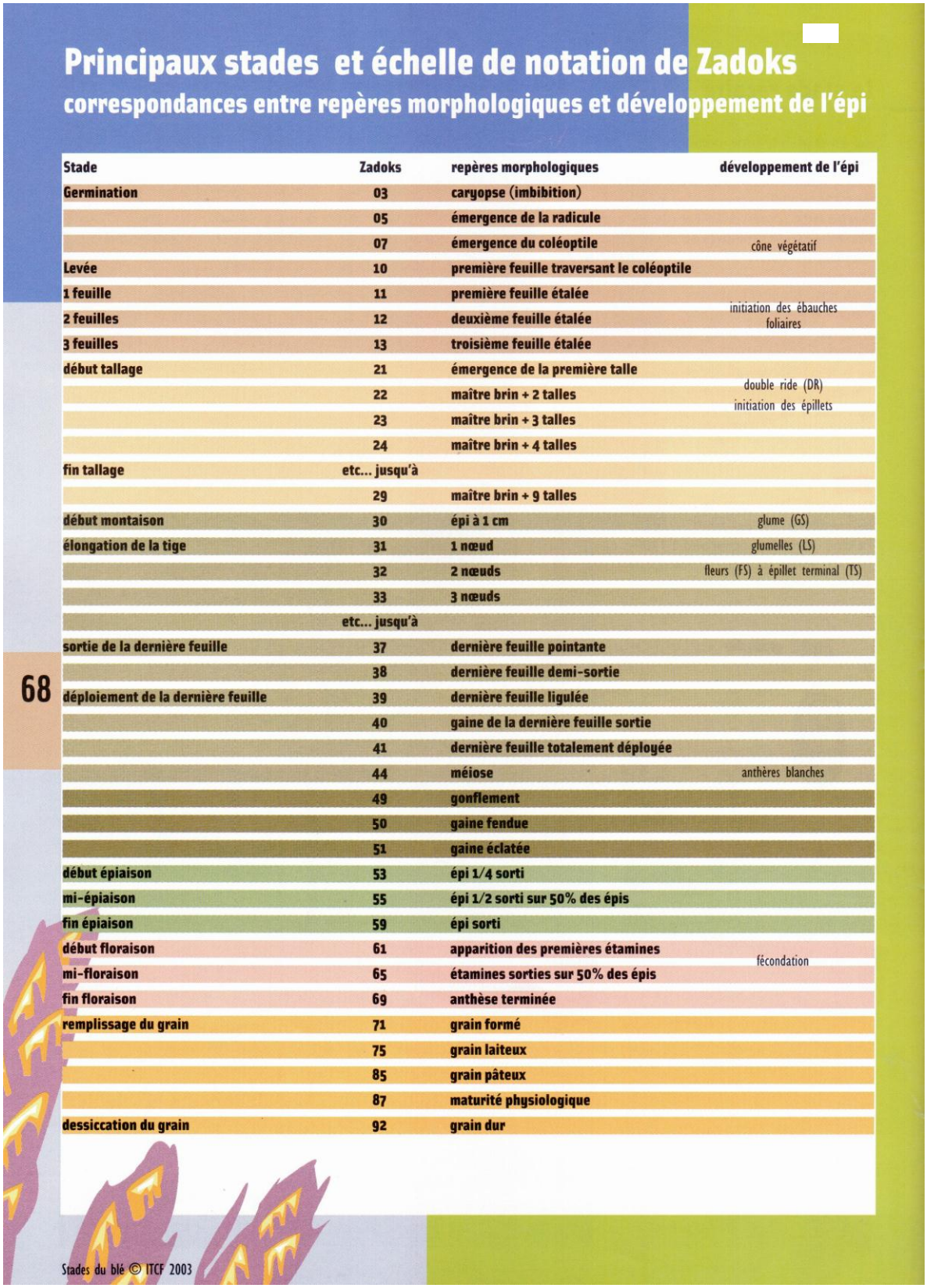
Nous avons comparé les performances de ces outils à des situations de référence, à l'aide de l'aire sous la courbe ROC. Les meilleures performances de prédictions, pour des cas où l'on a peu d'information sur les sites, sont obtenues avec les arbres de décision (scénario 2) ou avec le modèle linéaire généralisé (scénario 4 et 3). Néanmoins les variables issues de SeptoLIS® telles que le stade, considérées comme variables indicatrices, peuvent servir de à prédire des contaminations la semaine suivante de manière assez performante.

Nous avons montré que l'on pouvait prédire la probabilité que la contamination dépasse un seuil donné de 4/20, grâce à un modèle synthétisant les différents types d'informations disponibles. Le couplage des différents types de données disponibles : soit des mesures de terrain stockées années après années dans la base de données nationale Vigicultures®, soit les prédictions issues de modèles, pourrait certainement améliorer les prédictions d'épidémies.

ANNEXES.

ANNEXE I :

ANNEXE I.1 : Echelle de Zadoks (Source : Stade du blé ICTF Arvalis (source(1))



Principaux stades et échelle de notation de Zadoks

correspondances entre repères morphologiques et développement de l'épi

Stade	Zadoks	repères morphologiques	développement de l'épi
Germination	03	caryopse (imbibition)	
	05	émergence de la radicule	
	07	émergence du coléoptile	cône végétatif
Levée	10	première feuille traversant le coléoptile	
1 feuille	11	première feuille étalée	
2 feuilles	12	deuxième feuille étalée	initiation des ébauches foliaires
3 feuilles	13	troisième feuille étalée	
début tallage	21	émergence de la première talle	
	22	maître brin + 2 talles	double ride (DR)
	23	maître brin + 3 talles	initiation des épillets
	24	maître brin + 4 talles	
fin tallage	etc... jusqu'à		
	29	maître brin + 9 talles	
début montaison	30	épi à 1 cm	glume (GS)
élongation de la tige	31	1 nœud	glumelles (LS)
	32	2 nœuds	fleurs (FS) à épillet terminal (TS)
	33	3 nœuds	
	etc... jusqu'à		
sortie de la dernière feuille	37	dernière feuille pointante	
	38	dernière feuille demi-sortie	
déploiement de la dernière feuille	39	dernière feuille ligulée	
	40	gaine de la dernière feuille sortie	
	41	dernière feuille totalement déployée	
	44	méiose	anthères blanches
	49	gonflement	
	50	gaine fendue	
	51	gaine éclatée	
début épiaison	53	épi 1/4 sorti	
mi-épiaison	55	épi 1/2 sorti sur 50% des épis	
fin épiaison	59	épi sorti	
début floraison	61	apparition des premières étamines	fécondation
mi-floraison	65	étamines sorties sur 50% des épis	
fin floraison	69	anthèse terminée	
remplissage du grain	71	grain formé	
	75	grain laiteux	
	85	grain pâteux	
	87	maturité physiologique	
dessiccation du grain	92	grain dur	

68

Stades du blé © ITCF 2003

ANNEXE I.2 : résumé des caractéristiques de 3 modèles prédisant la septoriose. (source (14))

	Presept (SPV, France)	*Proculture (UCL, Belgique)*	*DESSAC (ADAS, Royaume Uni)*
Maladies concernées	Septoriose	Septoriose	Septoriose, Oïdium, Rouille brune, Rouille jaune
Objectifs du modèle	- Identification des contaminations du blé - Simulation de l'épidémie en fonction du climat - Ajustement de la stratégie fongicide	- Compréhension du développement du champignon - Ajustement de la stratégie fongicide	- Prédiction des risques d'infection - Ajustement de la stratégie fongicide
Echelle	Régionale	« à la parcelle »	« à la parcelle »
Type de modèle	Mécaniste	Mécaniste	Mécaniste
Données d'entrée	- Données météorologiques journalières : Température Pluviométrie - Observations phénologiques	- Données météorologiques horaires : Température Pluviométrie Humidité relative - Observations sur la parcelle (stade, % de feuilles atteintes…)	- Données météorologiques journalières : Température Pluviométrie Humidité relative Vitesse du vent - Positionnement des traitements réalisés - Observations des symptômes sortie hiver
Sorties du modèle	- Diagramme de contaminations - Indice de risque	- Courbe de la maladie visible et en incubation - Prévision de la maladie visible et en incubation - Prévision de la perte de rendement	- Risques d'infections - Doses et positionnement des applications - Prévision du rendement
Prévision de la nuisibilité	NON	Modèle statistique	Modèle mécaniste
Sensibilité variétale	NON	NON	OUI

ANNEXE I.3 : Description des équations du modèle de SeptoLIS®, issue de la publication de Gouache et al en cours d'acceptation (source(4)).

The calculation of winter accumulation of ground inoculum begins at tillering (GS21) and continues until the appearance of leaf L6. The algorithm is:

$$A(d) = (-0.0493 \times T(d)^2 + 1.8759 \times T(d) - 8.4949) \times (0.0008 \times P(d)^2 + 0.0104 \times P(d) + 0.828) \quad T(d) > \theta_{Tmin}$$
$$A(d) = 0 \quad T(d) \leq \theta_{Tmin}$$

$$V_g(d+1) = V_g(d) + A(d) + \theta_{mult} \times V_g(d - d_{latency}) \quad A(d) > 0 \text{ and } V_g(d - d_{latency}) > \theta_{winter.th}$$
$$V_g(d+1) = V_g(d) + A(d) \quad \text{otherwise}$$

$$U_g(d+1) = V_g(d+1) / \theta_{scale}$$

Time d here is expressed in calendar days. $T(d)$ and $P(d)$ are respectively mean daily temperature (°C) and daily precipitation (mm) on day d. In the above equations, $P(d)$ values greater than 25 are set equal to 25. $A(d)$ is an intermediate variable that expresses the effect of daily temperature and rainfall on ground inoculum. $V_g(d)$ is proportional to accumulated ground inoculum. It is initialized to 0 at the start of the calculation. $d_{latency}$ is the number of days of latency. $U_g(t)$ is actual ground inoculum.

After the appearance of L6, the model describes the disease dynamics for each leaf from L6 to L1. The number of potential infections produced by leaf l on day d is

$$U_l(d) = 1 - \exp(-\theta_{prod.inoc} * S_l(d))$$

where $S_l(d)$ is disease severity (percentage of leaf area showing symptoms) for leaf l on day d.

The number of new infections reaching leaf l on day d is given by

$$F_l(d) = U_g(d) \sum_{l=6}^{1} \exp(-h_l(d) * \theta_{height} * e^{(\theta_{rain} * P(d))}) + \sum_{l'=l+1}^{6} U_{l'}(d) \exp(-(h_l(d) - h_{l'}(d)) * \theta_{height} * e^{(\theta_{rain} * P(d))}) + \theta_{mult} * U(d)$$

where $h_l(d)$ is the height above ground level of leaf l on day d. The three terms on the right hand side represent respectively transfer from ground inoculum, from other leaves and from infections on the same leaf. Here as throughout, the sum is only over visible leaves. $F_l(d)$ is set to zero unless ($Tmin(d) \geq \theta_{min.today}$ or $Tmin(d-1) \geq \theta_{min.yesterday}$) and $P(d) + P(d-1) \geq \theta_{min.rain}$, where $Tmin(d)$ is minimum temperature day d. Infection on L1 is set to zero ($F_1(d) = 0$) unless $P(d) \geq \theta_{limitF1}$.

$U_g(t)$ is depleted by inoculum transfer to upper leaves, but cannot fall below a certain limit:

$$U_g(d+1) = \max\left[U_g(d) - \theta_{deplete} * U_g(d) * \sum_{l=6}^{1} \exp(-h_l(d) * \theta_{height} * e^{(\theta_{rain} * P(d))}), \theta_{min.inoc} \right]$$

The number of successful infections on leaf l, day d is

$$I_l(d) = L_l^v(d) L_l^h(d) \min(P(d) / \theta_{rain.infect}, 1)$$

$L_l^h(d)$ is the fraction of the leaf surface that is not diseased; $L_l^h(d) = 1.0 - S_l(d)/100$. $L_l^v(d)$ is the fraction of leaf l visible, calculated as

$$L_l^v(d) = \left[TT(d) - TT(d_{appear}) \right] / \left(\theta_{leaf.growth} * phyllochron \right)$$

where *phyllochron* is the number of degree-days between emergence of successive leaf layers and d_{appear} is the day of emergence of the leaf l, calculated in the plant model.

Lesion size, expressed as a fraction of final size, is a logistic function of thermal time ΔTT (base 0°C):

$$s(\Delta TT) = \theta_{initial} / \left[\theta_{initial} + (1 - \theta_{initial}) \exp(-\theta_{expansion} * \Delta TT) \right]$$

Finally, the severity on leaf l on day $d+1$ is the sum of infections from previous days

$$S_l(d+1) = \sum_{d'=1}^{d} I_l(d) * s(TT(d) - TT(d'))$$

where $TT(d)$ is accumulated thermal time up to day d. In the model, the original function for latency period is replaced by that proposed by Lovell et al. (2004) : the fraction of the latent period passed by a lesion on day d is $L_l(d) = 0.0039 \times T_{mean}(d) + 0.0108$, with T_{mean}(d) daily mean temperaure. For a given lesion, the latent period is finished once the cumulative sum of L_l(d) reaches a value of 1.

ANNEXE II:

ANNEXE II.1 : Tableau résumant le nombre de mesures intensité disponibles selon l'année pour chaque étage foliaire.

	F1	F2	F3	F4	F5	F6	total
2010	129	173	215	134	103	49	803
2011	103	126	164	84	70	30	577
total	232	299	379	218	173	79	1380

ANNEXE II.2 : Tableau résumé des erreurs de prédiction SpetoLIS® pour chaque feuille en prenant en compte la moyenne de la météo (en haut) ou en prenant les données météo spatialisées (en bas).

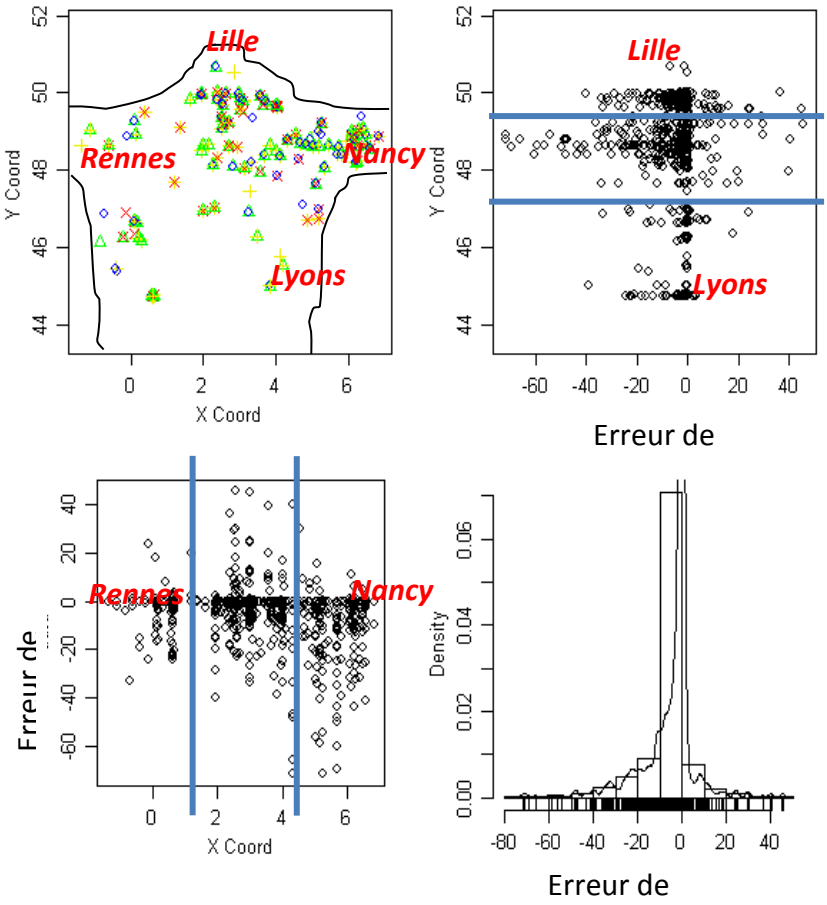
	mean	sd	0%	25%	50%	75%	100%	n	NA
F1err	-1.04	5.37	-49.26	0.00	0.00	0.00	25.00	233	466
F2err	-3.30	8.90	-52.23	-3.32	-0.30	0.00	38.75	299	400
F3err	-5.29	14.75	-73.77	-6.10	-0.25	0.00	89.09	379	320
F4err	-0.51	5.63	-38.91	-0.40	0.00	0.00	23.38	218	481
F5err	-2.71	8.60	-41.36	-5.32	-0.36	0.00	47.99	173	526
F6err	-2.58	11.74	-43.65	-4.42	-0.68	0.54	38.65	79	620

	mean	sd	0%	25%	50%	75%	100%	n	NA
F1err_spat	-1.16	5.56	-48.38	0.00	0.00	0.00	25.00	233	466
F2err_spat	-3.23	9.26	-54.21	-3.02	-0.32	0.00	39.84	299	400
F3err_spat	-4.92	14.48	-71.78	-5.77	-0.14	0.00	88.45	379	320
F4err_spat	-0.35	5.29	-32.11	-0.44	0.00	0.00	25.00	218	481
F5err_spat	-2.31	8.19	-33.23	-4.11	-0.35	0.00	45.71	173	526
F6err_spat	-2.53	11.73	-35.97	-6.71	-0.62	0.73	40.44	79	620

Il n'y a une petite différence entre les erreurs/résidus obtenus par les données météo et météo spatialisée (en bas), par la suite nous allons donc **garder météo spatialisée car c'est là que les erreurs sont les moins fortes.**

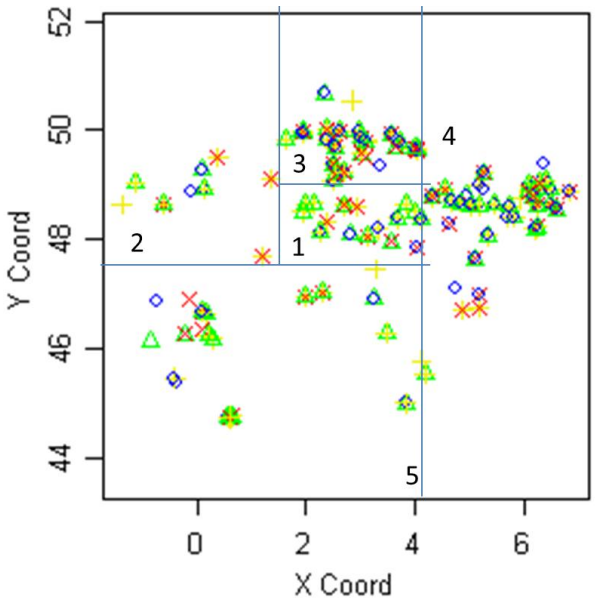
ANNEXE II.3 : création de la variable qualitative « Région ».

J'ai créé une variable région, sensée rendre compte des disparités de contaminations/prédiction/d'erreur de prédiction observées d'une région à l'autre. On observe dans la ***figure A*** ci-dessous, créé à l'aide du package geoR, que certaines zones font, par exemple, des erreurs de prédictions plus importantes que d'autre (***FigA***). Il parait donc important de prendre en compte une variable spatiale. Les régions administratives ne reflétant pas forcément les différentes zones, j'ai découpé la France en 5 zones listée en figure B. ■



FigA : Observation des mesure (haut à gauche), latitude des mesures en fonction de l'erreur (haut droite), erreur en fonction de la longitude (bas à gauche), histogramme des erreurs.

Fig B : Création des 5 zones.

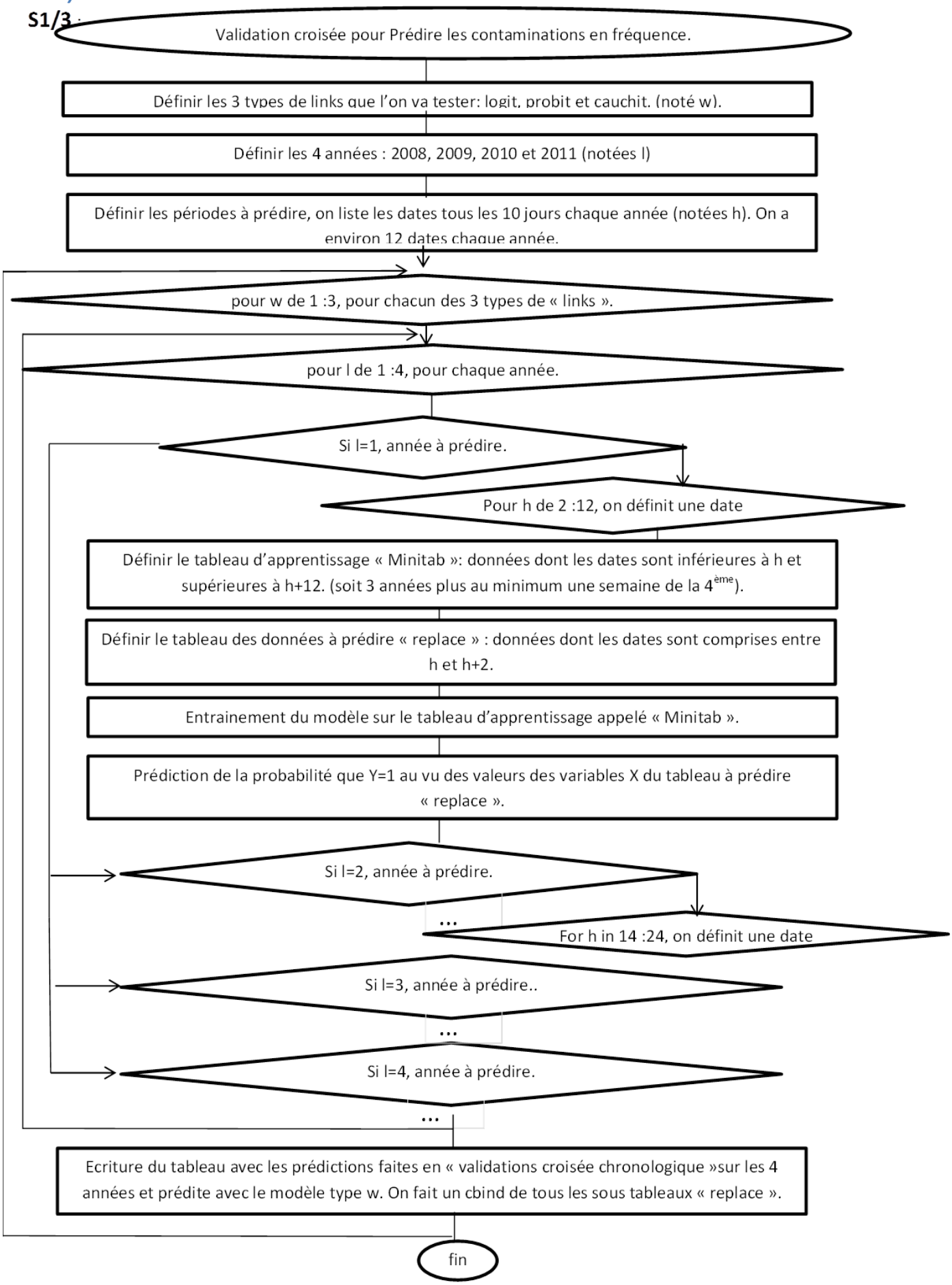


1: ile de France
2: ouest
3: Nord
4: Est
5:Sud

➔J'ai découpé en 5 zones qui semblaient avoir des erreurs différentes les unes des autres.

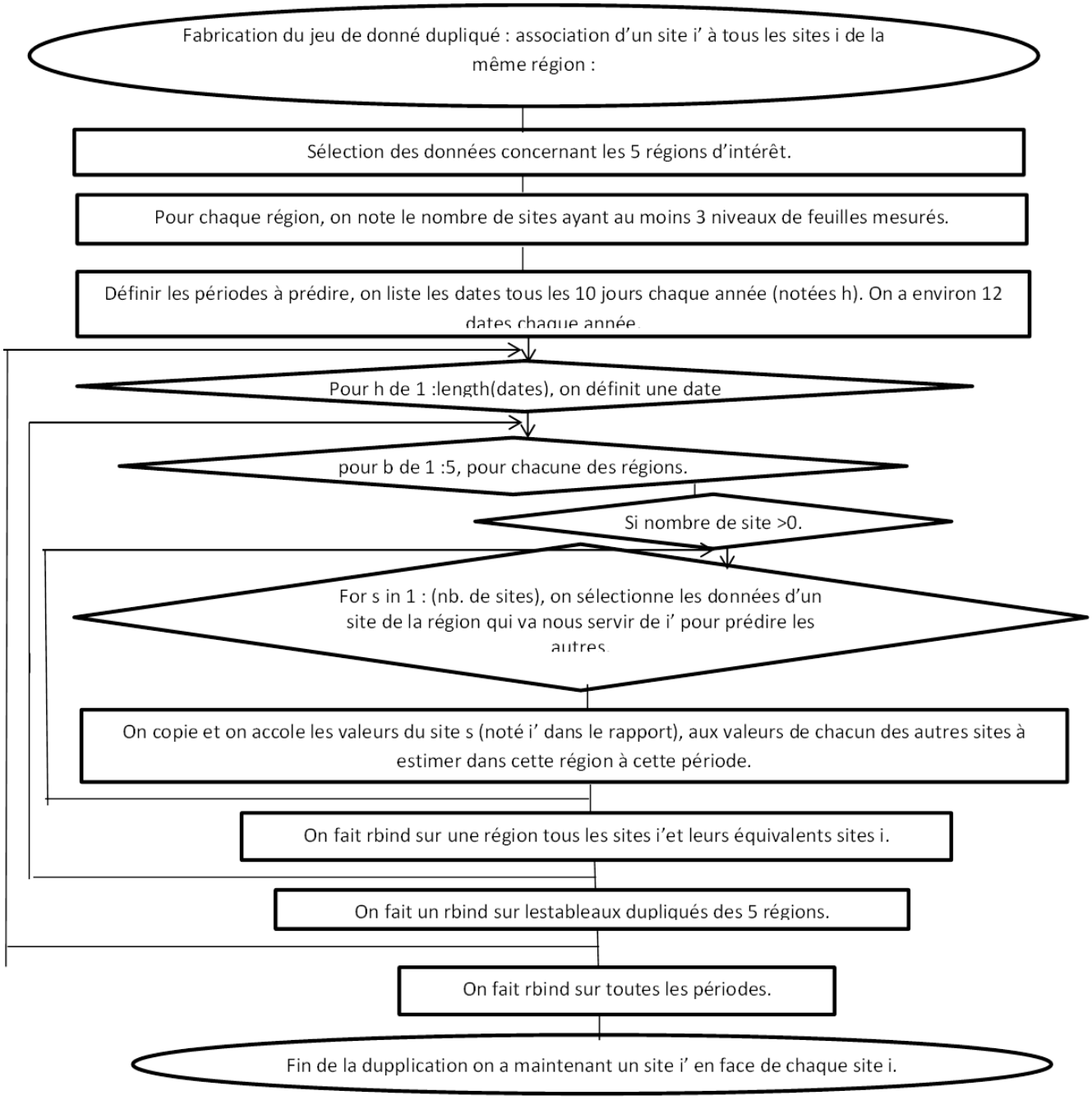
ANNEXE II.4 : Schéma des deux principaux algorithmes utilisés pour la validation croisée (A) et l'autre pour fabriquer les tableaux dupliqués utilisés dans les scénarios 2 et 4 (B).

A)



B)

S2/4 : La seule différence va se faire au niveau du tableau initial. On ne va garder que les données concernant 5 régions administratives : Lorraine, Champagne Ardenne, Normandie, Picardie et Poitou Charente. On va associer un site i' à tous les autres sites i de la région. On prédira avec les données de i'la contamination existante en i.



ANNEXE III :

ANNEXE III.1 : A) résultats des Aires sous la courbe obtenus quand seules les feuilles de l'etage foliaire 3 sont considérées. B) Résultats des AUC pour toutes les feuilles.

SITUATION 1 : J=J i=i	**estimé sur toutes les 26542 feuilles**
VARIABLE INDICATRICE	
	AUC
SeptoLIS	**0.816**
Stade	0.807
Age	0.794
resistance	0.468
precocite	0.532
altitude	0.472

MODELE 1			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		10118
1 logit	0.848	0.835	7067
1 probit	0.848	0.834	7057
1 cauchit	**0.848**	**0.832**	**7178**
REFERENCE			
J connu…	1	1	

SITUATION 2 : J=J i'=i	**estimé sur toutes les 53204 feuilles**
VARIABLE INDICATRICE	
	AUC
SeptoLIS i'	0.904
Stade i'	**0.915**
Age i'	0.897
resistance i'	0.487
precocite i'	0.548
Altitude i'	0.457
contmin J i'	0.884

MODELE 2			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		73728
1 logit	**0.943**	**0.927**	**32105**
1 probit	0.943	0.927	32278
1 cauchit	0.943	0.920	32891
REFERENCE			
J connu en i'	0.884	0.876	

SITUATION 3 : J=J+1 i=i	**estimé sur toutes les F3: 2264 feuilles**
VARIABLE INDICATRICE	
	AUC
SeptoLIS	0.820
Stade	0.835
Age	**0.846**
resistance	0.482
precocite	0.492
altitude	0.421
contamin J	0.857

MODELE 3			
	AUC	**AUC ac CV**	**AIC**
modele nul			2850
1 logit	0.863	0.876	1958
1 probit	0.863	**0.875**	1959
1 cauchit	0.862	0.876	1962
REFERENCE			
J connu en i	0.857	0.859	

SITUATION 4 : J=J+1 i'=i	**estimé sur toutes les 19958 feuilles**
VARIABLE INDICATRICE	
	AUC
septoLIS i'	0.810
stade i'	**0.814**
age i'	0.800
résistance i'	0.494
précocite i'	0.525
altitude i'	0.497
contmin J i'	0.771

MODELE 4			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		15107
1 logit	0.968	**0.759**	5647
1 probit	0.968	0.731	5609
1 cauchit	0.967	0.699	5864
REFERENCE			
J connu en i'	0.771	**0.769**	

SITUATION 1 : J=J i=i	estimé sur toutes les 26542 feuilles		
VARIABLE INDICATRICE			
	AUC		
SeptoLIS	**0.784**		
Stade	0.603		
Age	0.778		
resistance	0.473		
precocite	0.506		
altitude	0.446		
MODELE 1			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		32 341
1 logit	0.830	0.833	23 954
1 probit	0.830	**0.833**	**23 953**
1 cauchit	0.829	0.829	24 331
MODELE 1 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	0.838	0.832	
1 probit	0.838	**0.832**	
1 cauchit	0.836	0.827	
REFERENCE			
J connu…	1	1	

SITUATION 2 : J=J i'=i	estimé sur toutes les 192752 feuilles		
VARIABLE INDICATRICE			
	AUC		
SeptoLIS i'	**0.814**		
Stade i'	0.686		
Age i'	0.801		
resistance i'	0.489		
precocite i'	0.526		
Altitude i'	0.449		
contmin J i'	0.799		
MODELE 2			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.5		253 282
1 logit	0.877	0.871	162 887
1 probit	0.877	**0.872**	**162 785**
1 cauchit	0.876	0.866	166 361
MODELE 2 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	0.934	0.922	
1 probit	0.933	**0.923**	
1 cauchit	0.934	0.915	
REFERENCE			
	0.799	0.799	

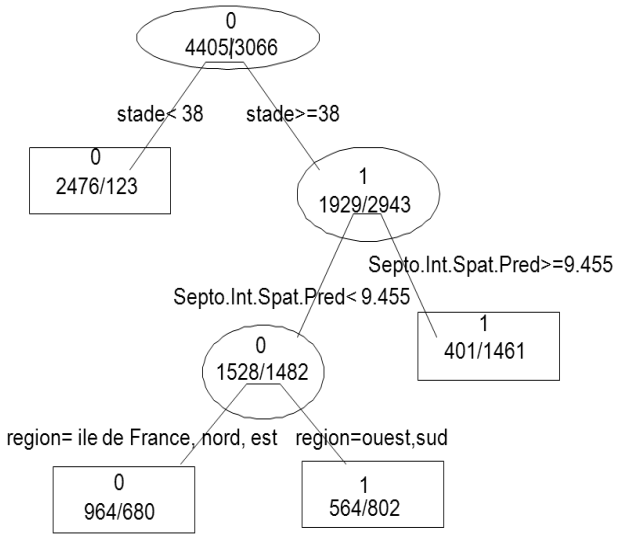
SITUATION 3 : J=J+1 i=i	estimé sur toutes les 14339 feuilles		
VARIABLE INDICATRICE			
	AUC		
SeptoLIS	0.722		
Stade	0.545		
Age	0.722		
resistance	0.471		
precocite	0.510		
altitude	0.433		
contamin J	**0.814**		
MODELE 3			
	AUC	**AUC ac CV**	**AIC**
modele nul	0.500		18 775
1 logit	0.889	0.905	11 263
1 probit	0.887	0.904	11 376
1 cauchit	0.890	**0.906**	**11 251**
MODELE 3 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	0.897	0.879	
1 probit	0.896	0.877	
1 cauchit	0.898	**0.881**	
REFERENCE			
J connu…	0.814	0.844	

SITUATION 4 : J=J+1 i'=i	estimé sur toutes les 67782 feuilles		
VARIABLE INDICATRICE			
	AUC		
septoLIS i'	0.807		
stade i'	0.690		
age i'	**0.816**		
résistance i'	0.496		
précocite i'	0.520		
altitude i'	0.434		
contmin J i'	0.789		
MODELE 4			
	AUC	**AUC ac CV**	**AIC**
modele nul			92 896
1 logit	0.943	0.852	**40 639**
1 probit	0.942	**0.854**	41 208
1 cauchit	0.943	0.847	41 204
MODELE 4 performance sur les F3:			
	AUC	**AUC ac CV**	
modele nul	0.5		
1 logit	0.963	0.804	
1 probit	0.962	**0.807**	
1 cauchit	0.963	0.799	
REFERENCE			
J connu…	0.789	0.786	

ANNEXE-III.2 : Les arbres pour les quatre scénarios.

SCENARIO 1 :

A représentation de l'arbre CAS 1



B) Commentaire de l'arbre:

node), split, n, loss, yval, (yprob)

*** denotes terminal node**

1) root 7471 3066 0 (0.58961317 0.41038683)

2) stade< 38 2599 123 0 (0.95267411 0.04732589) *

3) stade>=38 4872 1929 1 (0.39593596 0.60406404)

6) Septo.Int.Spat.Pred< 9.455289 3010 1482 0 (0.50764120 0.49235880)

12) region=1ileDF,3nord,4est 1644 680 0 (0.58637470 0.41362530) *

13) region=2ouest,5sud 1366 564 1 (0.41288433

C) Surrogates :

- Node 1:

stade < 38 to the left, improve=1050.6710, (0 missing)
Septo.Int.Spat.Pred < 0.7288156 to the left, improve= 838.1676, (0)
Age_f < 27.5 to the left, improve= 783.8678, (0 missing)

- Node 3 :

Septo.Int.Spat.Pred < 9.455289 to the left, improve=196.54910, (0 missing)
region splits as LRLLR, improve= 96.53788, (0 missing)
stade < 54 to the left, improve= 68.05467, (0 missing)

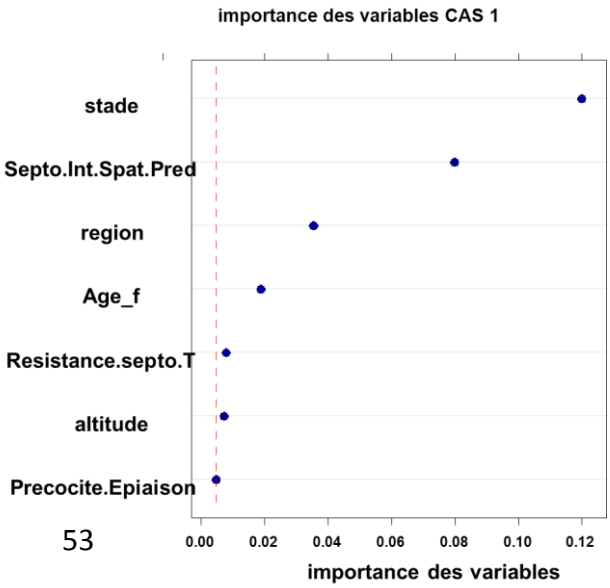
- Node 6:

region splits as LRLLR, improve=44.91245, (0 missing)
stade < 67 to the right, improve=15.25582, (0 missing)
Resistance.septo.T < 6.75 to the right, improve=11.97729, (0 missing)

	D) matrice des confusions		
		Predictions	
		0	1
réalité	0	3440	965
	1	803	2263

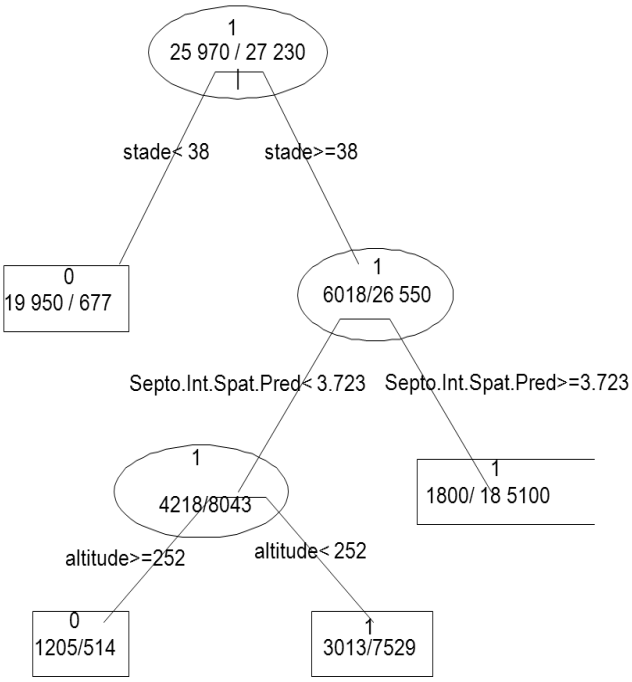
sensib	0.70
spécif	0.81
taux d'erreur	0.24

E) importance des variables.	MeanDecreaseGini
Septo.Int.Spat.Pred	943
stade	847
Age_f	584
altitude	551
region	214
Precocite.Epiaison	180
Resistance.septo.T	164



SCENARIO 2:

B représentation de l'arbre CAS 2



B) Commentaire de l'arbre.

n= 53204

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 53204 25973 1 (0.48817758 0.51182242)

2) stade< 38 20632 677 0 (0.96718689 0.03281311) *

3) stade>=38 32572 6018 1 (0.18475992 0.81524008)

6) Septo.Int.Spat.Pred< 3.72278 12261 4218 1 (0.34401762 0.65598238)

12) altitude>=252 1719 514 0 (0.70098895 0.29901105) *

13) altitude< 252 10542 3013 1 (0.28580914 0.71419086) *

7) Septo.Int.Spat.Pred>=3.72278 20311 1800 1 (0.08862193 0.91137807) *

C) Surrogate:

NO1 : stade < 38 to the left, improve=15465.33, (0 missing)

Septo.Int.Spat.Pred < 0.1767312 to the left, improve=13201.63, (0 missing)

Septo.FreqD < 1 to the left, improve=12619.63, (0 missing)

No3 : Septo.Int.Spat.Pred < 3.72278 to the left, improve=997.4012, (0 missing)

stade < 52 to the left, improve=726.9359, (0 missing)

Age_f < 39.5 to the left, improve=573.3120, (0 missing)

No6 : altitude < 252 to the right, improve=509.5367, (0 missing)

Resistance.septo.T < 4.5 to the right, improve=314.1939, (0 missing)

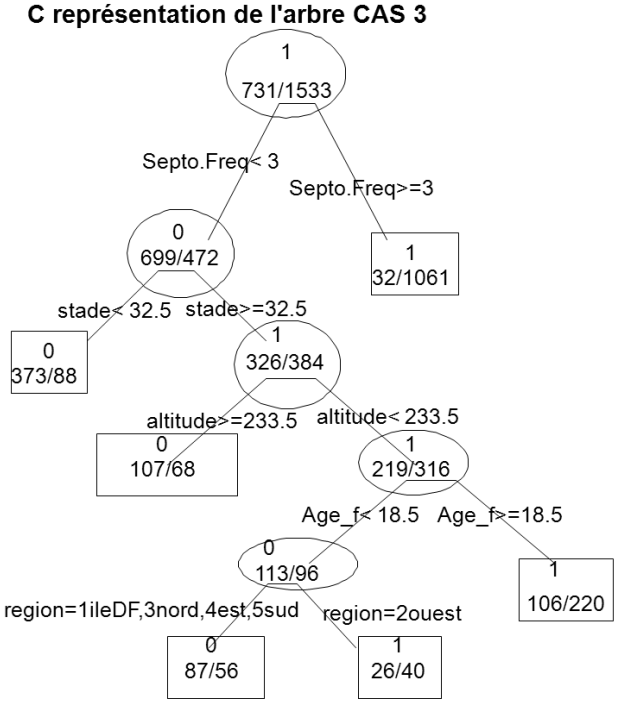
Precocite.Epiaison < 6.25 to the right, improve=232.2085, (0 missing)

	D) matrice des confusion		
		Predictions	
		0	1
réalité	0	21160	4813
	1	1191	26040

	sans CV
sensibilité	0.84
spécificité	0.95
taux d'erreur	0.14

SCENARIO 3:

C représentation de l'arbre CAS 3



B) Commentaire de l'arbre :

n= 2264

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 2264 731 1 (0.32287986 0.67712014)
2) Septo.Freq< 3 1171 472 0 (0.59692570 0.40307430)
4) stade< 32.5 461 88 0 (0.80911063 0.19088937) *
5) stade>=32.5 710 326 1 (0.45915493 0.54084507)
10) altitude>=233.5 175 68 0 (0.61142857 0.38857143) *
11) altitude< 233.5 535 219 1 (0.40934579 0.59065421)
22) Age_f< 18.5 209 96 0 (0.54066986 0.45933014)
44) region=1ileDF,3nord,4est,5sud 143 56 0 (0.60839161 0.39160839) *

C)Surrogate:

Node 1 : Septo.Freq < 3 to the left, improve=364.3255, (0 missing)

stade < 35 to the left, improve=278.5918, (0 missing)
Age_f < 18.5 to the left, improve=276.3513, (0 missing)

Node 2: stade < 32.5 to the left, improve=68.46341, (0 missing)
- Age_f < 15.5 to the left, improve=62.96486, (0 missing)
- Septo.Int.Spat.Pred < 0.0001875653 to the left, improve=48.09893, (0 missing)

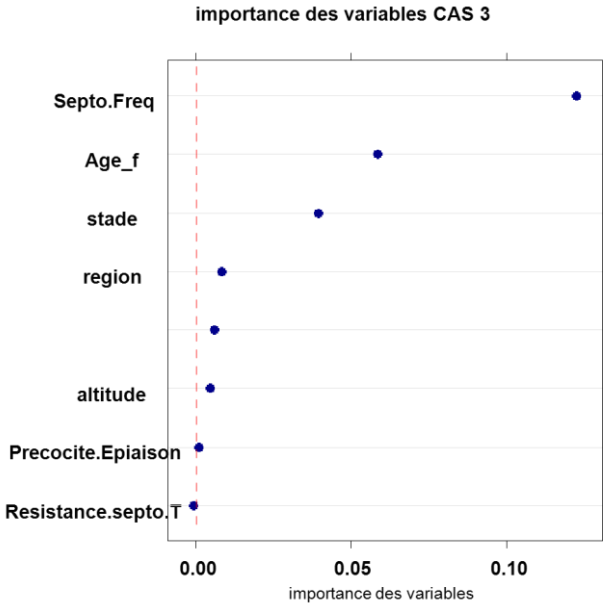
Node 5 : altitude < 233.5 to the right, improve=10.770160, (0 missing)
- Age_f < 18.5 to the left, improve=10.354450, (0 missing)
- region splits as LRLLL, improve= 9.854672, (0 missing)

Node 11 :Age_f < 18.5 to the left, improve=11.830440, (0 missing)
- Septo.Int.Spat.Pred < 0.0001875653 to the left, improve= 7.604649, (0 missing)
- Septo.Freq < 1 to the left, improve= 4.803518, (0 missing)

Node 22 :region splits as LRLLL, improve=4.153601, (0 missing)
- altitude < 111 to the right, improve=3.564521, (0 missing)
- Septo.Freq < 1 to the left, improve=2.886686, (0 missing)

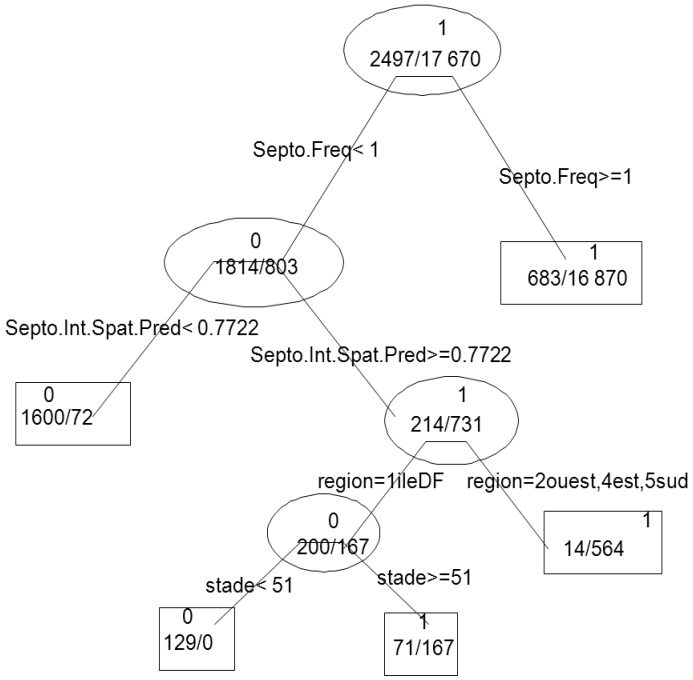
-	**D) matrice des confusion :**					
		Predictions				sans CV
		0	1		sensibilité	0.86
réalité	0	567	164		spécificité	0.78
	1	212	1321		taux d'erreur	0.17

E) importance des variables :	MeanDecreaseGini
Septo.Freq	202
Age_f	166
stade	147
Septo.Int.Spat.Pred	124
altitude	123
region	43
Precocite.Epiaison	42
Resistance.septo.T	37



SCENARIO 4 :

D représentation de l'arbre CAS 4



B) Commentaire de l'arbre :

n= 20171

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 20171 2497 1 (0.12379158 0.87620842)
2) Septo.Freq< 1 2617 803 0 (0.69316011 0.30683989)
4) Septo.Int.Spat.Pred< 0.7721911 1672 72 0 (0.95693780 0.04306220) *
5) Septo.Int.Spat.Pred>=0.7721911 945 214 1 (0.22645503 0.77354497)
10) region=1ileDF 367 167 0 (0.54495913 0.45504087)
20) stade< 51 129 0 0 (1.00000000 0.00000000) *
21) stade>=51 238 71 1 (0.29831933 0.70168067) *

C) Surrogates:

Node 1 : Septo.Freq < 1 to the left, improve=1949.719, (0 missing)
stade < 38 to the left, improve=1677.484, (0 missing)
Septo.Int.Spat.Pred < 0.05161568 to the left, improve=1291.858, (0 missing)
Node 2: Septo.Int.Spat.Pred < 0.7721911 to the left, improve=644.3388, (0 missing)
Age_f < 26.5 to the left, improve=456.6808, (0 missing)
stade < 35 to the left, improve=333.2772, (0 missin
Node5 : region splits as LR-RR, improve=121.73910, (0 missing)
Precocite.Epiaison < 6.25 to the right, improve= 46.18975, (0 missing)
stade < 47.5 to the right, improve= 45.20776, (0 missing)
Node10: stade < 51 to the left, improve=82.37769, (0 missing)
Septo.Int.Spat.Pred < 3.774069 to the left, improve=82.37769, (0 missing)

	D) matrice des confusion		
		Predictions	
		0	1
réalité	0	1178	1319
	1	623	17051

sensibilité	0.95
spécificité	0.49
taux d'erreur	0.11

ANNEXE III.3 : Tableau des coefficients des modèles et odds ratios au dessous:

	modèle cas 1:		modèle cas 2:		modèle cas 3:		modèle cas 4:	
(Intercept)	-1.97E+00	***	-3.07E+00	***	-0.920253	*	-0.880654	***
feuilleF2	5.24E-01	***	7.68E-01	***	0.4863105	***	0.5877011	***
feuilleF3	1.11E+00	***	1.38E+00	***	1.29E+00	***	8.96E-01	***
feuilleF4	1.18E+00	***	1.67E+00	***	9.73E-01	***	5.17E-02	
feuilleF5	1.62E+00	***	2.24E+00	***	1.52E+00	***	4.97E-01	***
feuilleF6	1.88E+00	***	2.36E+00	***	6.75E-01	**	6.30E-01	***
stade	3.25E-02	***	5.91E-02	***	1.74E-02	***	5.32E-02	***
Resistance.septo.T	-5.80E-02	***	-1.33E-01	***	4.00E-02	***	5.51E-02	***
Precocite.Epiaison	-1.30E-01	***	-1.98E-01	***	-1.20E-02		1.22E-02	*
Septo.Int.Spat.Pred	7.53E-03	***	7.87E-03	***	-7.55E-04	***	3.16E-04	.
altitude	-6.38E-04	***	-6.77E-05	.	2.64E-01	**	8.44E-01	***
annee2009	-4.53E-01	***	-1.19E-01	***	-6.20E-01	***	-1.55E+00	***
annee2010	-9.56E-01	***	-4.98E-01	***	-1.58E-01	*	1.02E-01	**
annee2011	-1.19E+00	***	-8.54E-01	***	1.70E-01	*	1.63E+00	***
Age_f	2.99E-02	***	2.26E-02	***	-6.77E-01	***	2.28E-01	***
region2ouest	2.87E-01	***	5.06E-01	***	-1.39E+00	***	-1.15E+00	***
region3nord	-4.06E-01	***	-1.65E-01	***	-1.80E+00	***	-1.29E+00	***
region4est	-7.47E-02	*	6.57E-03		-6.88E-02	**	-3.05E-01	***
region5sud	2.07E-01	***	5.88E-01	***	-1.96E-01	***	-4.98E-01	***
Septo.FreqD			4.98E-02	***	4.26E-01	***	6.16E-01	***

	modèle cas 1:		modèle cas 2:		modèle cas 3:		modèle cas 4:	
(Intercept)	7.18	***	21.52	***	2.51	*	2.41	***
feuilleF2	0.59	***	0.46	***	0.61	***	0.56	***
feuilleF3	0.33	***	0.25	***	0.27	***	0.41	***
feuilleF4	0.31	***	0.19	***	0.38	***	0.95	
feuilleF5	0.20	***	0.11	***	0.22	***	0.61	***
feuilleF6	0.15	***	0.09	***	0.51	**	0.53	***
stade	0.97	***	0.94	***	0.98	***	0.95	***
Resistance.septo.T	1.06	***	1.14	***	0.96	***	0.95	***
Precocite.Epiaison	1.14	***	1.22	***	1.01		0.99	*
Septo.Int.Spat.Pre	0.99	***	0.99	***	1.00	***	1.00	.
altitude	1.00	***	1.00	.	0.77	**	0.43	***
annee2009	1.57	***	1.13	***	1.86	***	4.71	***
annee2010	2.60	***	1.65	***	1.17	*	0.90	**
annee2011	3.29	***	2.35	***	0.84	*	0.20	***
Age_f	0.97	***	0.98	***	1.97	***	0.80	***
region2ouest	0.75	***	0.60	***	4.02	***	3.16	***
region3nord	1.50	***	1.18	***	6.04	***	3.62	***
region4est	1.08	*	0.99		1.07	**	1.36	***
region5sud	0.81	***	0.56	***	1.22	***	1.65	***
Septo.FreqD			0.95	***	0.65	***	0.54	***

En bleu : si la variable augmente d'une unité, alors on augmente la probabilité d'être contaminée >4/20 ou au contraire en rouge cela diminue le le risque d'être contaminée.

ANNEXE III.4 : Résultats des Anova des 4 modèles.

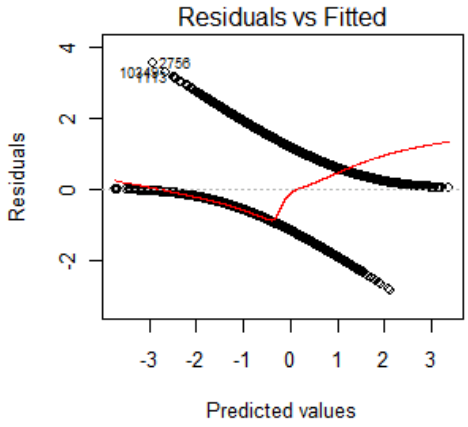
cas 1:						
	Df	Deviance	Resid. Df	Resid. De	P(>|Chi|)	
NULL	26541	32339				
feuille	5	2015.1	26536	30324	<2e-16	***
stade	1	3751.9	26535	26572	<2e-16	***
Resistance.	1	80.3	26534	26492	<2e-16	***
Precocite.E	1	1.9	26533	26490	0.1659	
Septo.Int.S	1	642.4	26532	25848	<2e-16	***
altitude	1	131.6	26531	25716	<2e-16	***
annee	3	1923	26528	23793	<2e-16	***
Age_f	1	369.1	26527	23424	<2e-16	***
region	4	446.7	26523	22977	<2e-16	***

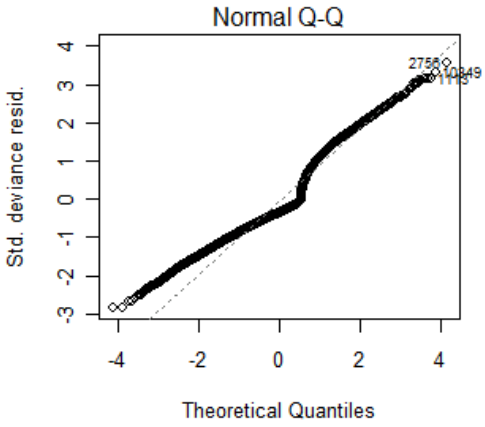
cas 2:						
	Df	Deviance	Resid. Df	Resid. D	P(>|Chi|)	
NULL	302768	382341				
feuille	5	16951	302763	365390	2.20E-16	***
stade	1	79942	302762	285447	2.20E-16	***
Resistance	1	1988	302761	283460	2.20E-16	***
Precocite.E	1	142	302760	283317	2.20E-16	***
Septo.Int.S	1	4622	302759	278695	2.20E-16	***
altitude	1	1625	302758	277069	2.20E-16	***
annee	3	16899	302755	260170	2.20E-16	***
Age_f	1	2454	302754	257716	2.20E-16	***
region	4	8368	302750	249348	2.20E-16	***
Septo.Freq	1	7495	302749	241853	2.20E-16	***

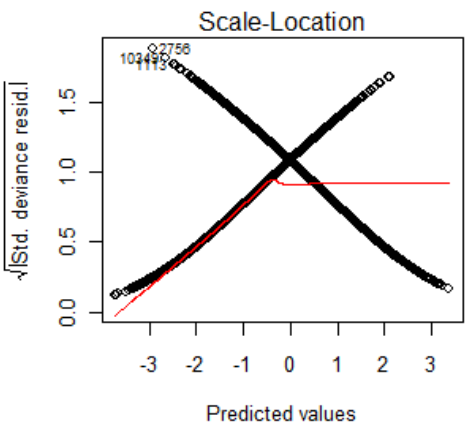
cas 3						
	Df	Deviance	Resid. Df	Resid. De	P(>|Chi|)	
NULL	14338	18773				
feuille	5	1910.3	14333	16863	2.20E-16	***
stade	1	1230.34	14332	15632	2.20E-16	***
Age_f	1	206.31	14331	15426	2.20E-16	***
Septo.Int.S	1	76.65	14330	15349	2.20E-16	***
altitude	1	102.05	14329	15247	2.20E-16	***
region	4	699.83	14325	14548	2.20E-16	***
annee	3	966.88	14322	13581	2.20E-16	***
Resistance.	1	2.54	14321	13578	0.1109	
Precocite.E	1	62.95	14320	13515	2.12E-15	***
Septo.Freq	1	2292.19	14319	11223	2.20E-16	***

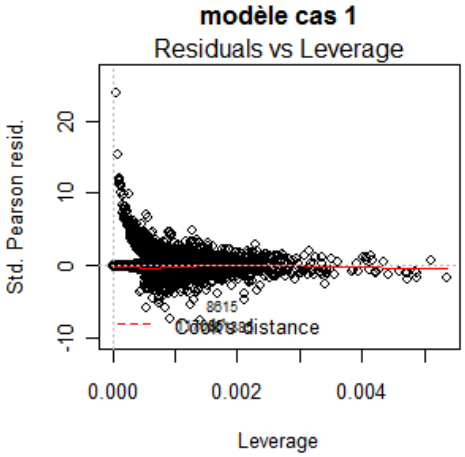
cas 4:						
	Df	Deviance	Resid. Df	Resid. D	P(>|Chi|)	
NULL	67781	92895				
feuille	5	16577.2	67776	76317	2.20E-16	***
stade	1	12516.2	67775	63801	2.20E-16	***
Age_f	1	798.6	67774	63003	2.20E-16	***
Septo.Int.S	1	266.2	67773	62736	2.20E-16	***
altitude	1	1384.9	67772	61351	2.20E-16	***
region	4	4560	67768	56791	2.20E-16	***
annee	3	1233	67765	55558	2.20E-16	***
Resistance	1	838.6	67764	54720	2.20E-16	***
Precocite.E	1	1224	67763	53496	2.20E-16	***
Septo.Freq	1	12897	67762	40599	2.20E-16	***

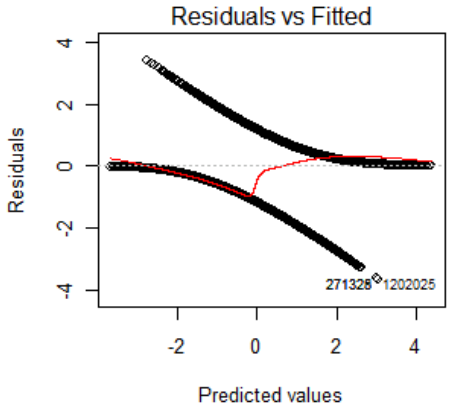
ANNEXE III.5 : résidus des modèles 1,2,3 et 4..

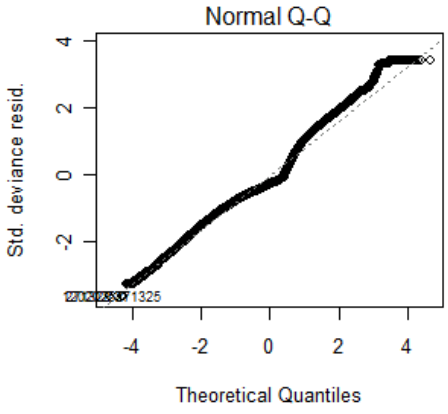


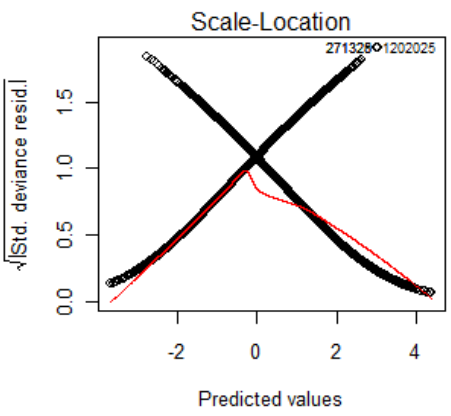


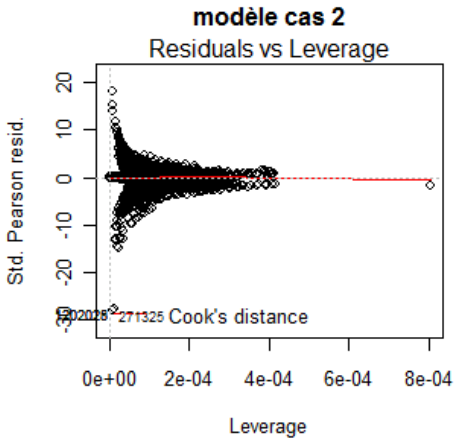


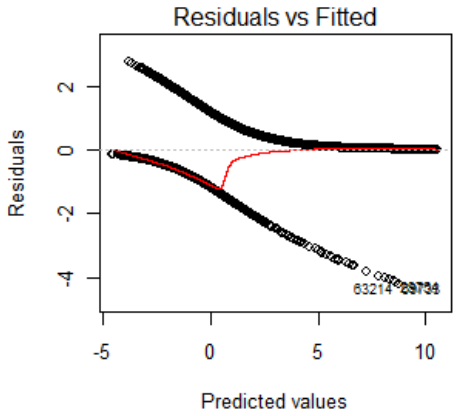


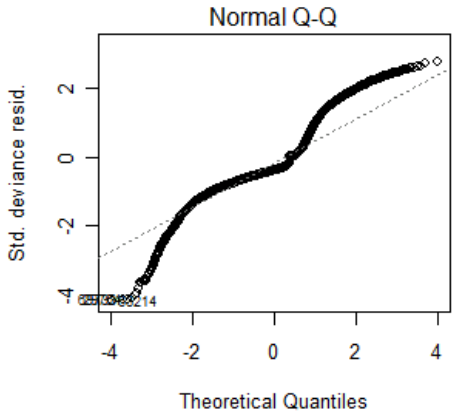


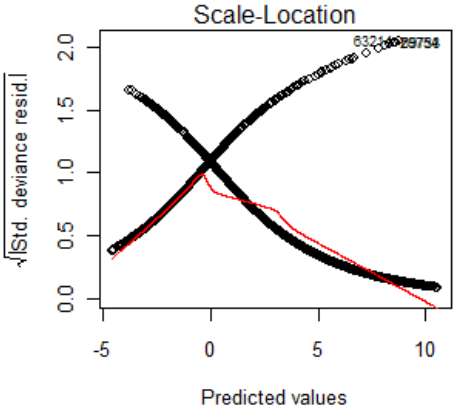


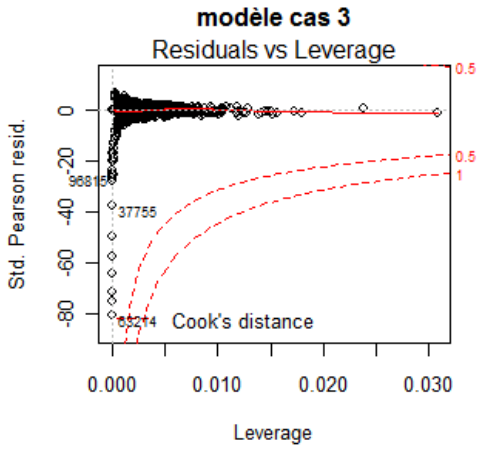


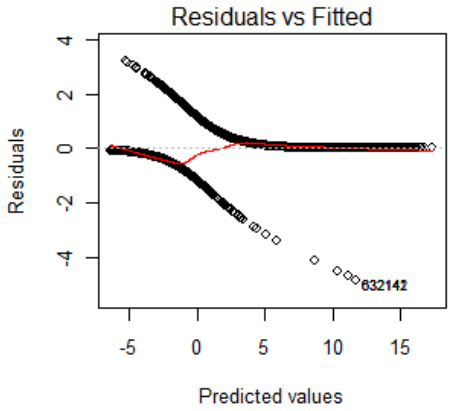


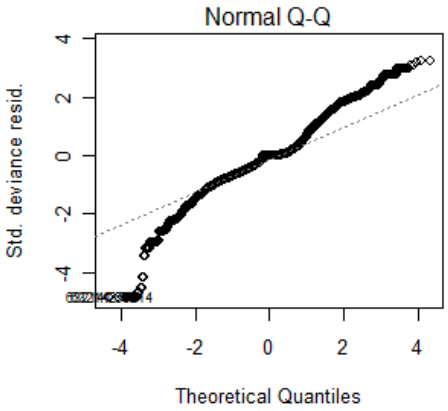

Residuals vs Fitted
Residuals
Predicted values
63214
29754

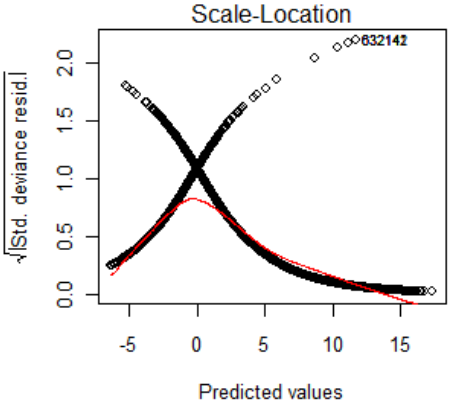

Normal Q-Q
Std. deviance resid.
Theoretical Quantiles

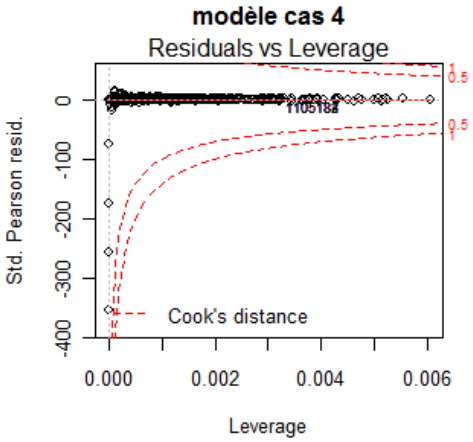

Scale-Location
√|Std. deviance resid.|
Predicted values


modèle cas 3
Residuals vs Leverage
Std. Pearson resid.
37755
63214
Cook's distance
Leverage


Residuals vs Fitted
Residuals
632142
Predicted values


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles


Scale-Location
√|Std. deviance resid.|
632142
Predicted values


modèle cas 4
Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage

ANNEXE IV :

ANNEXE IV.1. Présentation de l'amélioration des prédictions de septoLIS® à partir d'informations issues de vigicultures®.

Peut-on prédire les erreurs de prédiction et corriger SeptoLIS® ?

1) L'idée : Modéliser des erreurs de prédiction de SeptoLIS® à l'aide des variables disponibles sur Vigiculture®

1.1) Objectif :

1.2) Données :

1.3) Stratégies employées :

2) Matériel et méthodes employées.

2.1) Comment prendre en compte la dimension spatiale des erreurs de prédictions?

2.2) Comment prendre en compte la dimension temporelle des erreurs de prédictions?

2.3) Présence d'interactions potentielles entre nos différentes variables:

2.4) Sélection du meilleur modèle linéaire généralisé :

2.5) Principe des modèles mixtes et justification de leur utilisation :

2.6) Résumé des différents modèles testés.

2.7) Enoncés du modèle choisi.

2.8) Principe de la validation croisée utilisée :

3) Résultats : quelles sont les performances de notre correction ?

1) L'idée : Modéliser des erreurs de prédiction de SeptoLIS® à l'aide des variables disponibles sur Vigiculture®.

-Description de notre objectif, des données utilisées et des stratégies employées pour l'atteindre.

1.1) Objectif :

On recherche à améliorer les performances de SeptoLIS®. C'est un modèle qui se base sur les connaissances biologiques. Il décrit le développement de la plante hôte et du champignon à l'aide de nombreuses équations imbriquées. Il a été pris pour parti de ne pas le modifier. Ainsi on a travaillé à partir de ses prédictions en intensité et sur ses erreurs de prédictions.

Notre idée est d'utiliser les données mesurées en intensité et de chercher à prédire l'erreur de SpetoLIS® *(figure a et b)* à l'aide des données environnementales et culturales disponibles dans vigiculture®. Si l'on peut faire un modèle prédisant les erreurs ***(figure b)***, on pourra alors facilement rajouter à chaque prédiction SeptoLIS® l' « erreur prédite » et ainsi corriger SeptoLIS®.

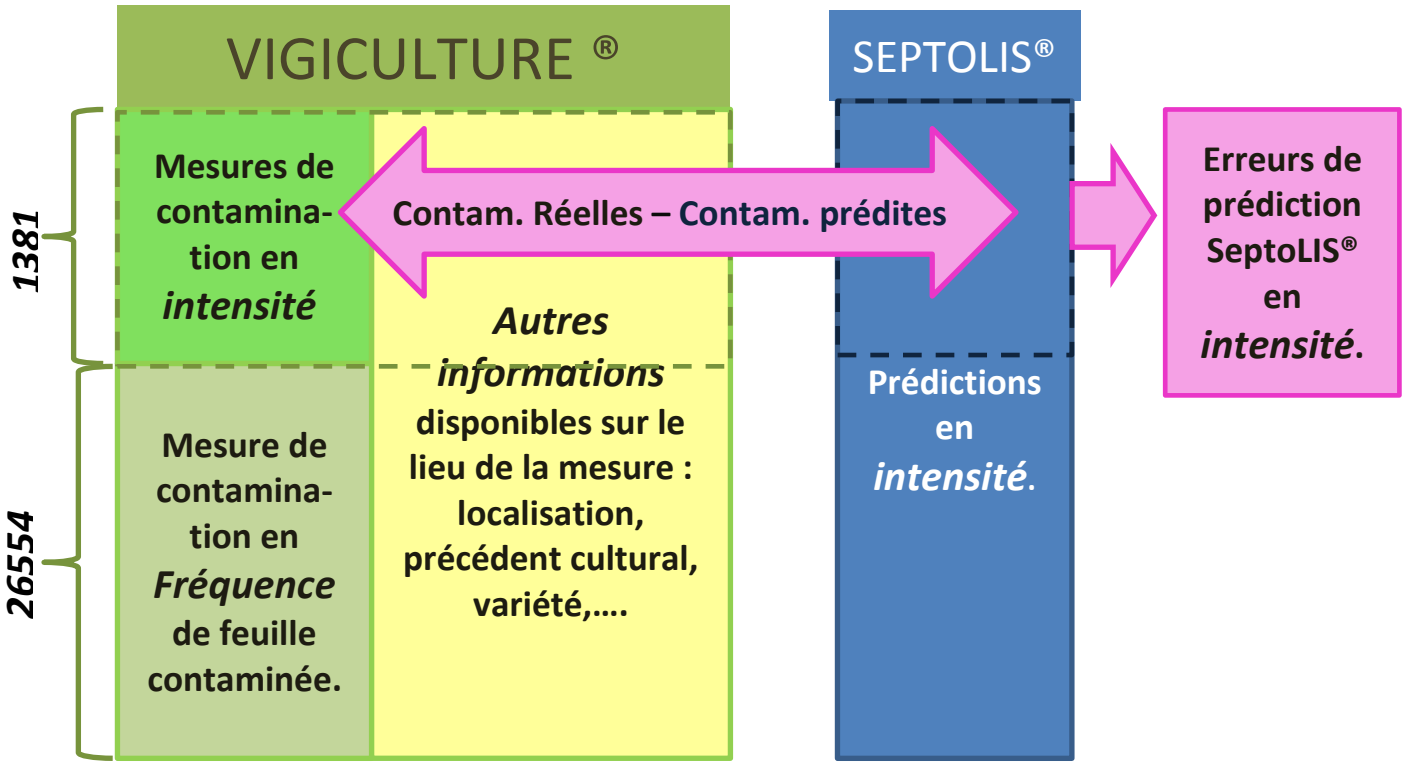


Figure a : Schéma simplifié du tableau de données.

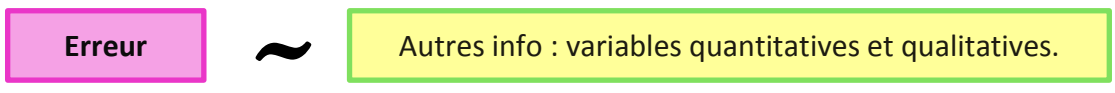


Figure b : Schéma de l'objectif.

1.2) Données :

Dans mes données je dispose des relevés de contaminations de septoriose en intensité effectués par les réseaux BSV entre 2010 et 2011. Elles sont rangées dans vigicultures®. A chaque mesure réelle de vigicultures®, est associée la prédiction que septoLIS® aurait faite pour ce lieu, dans les conditions météorologiques de l'époque et à cette date. Ce rapprochement permet de calculer l'erreur de prédiction de SeptoLIS®. Elle est définie en précisément ci-dessous. L' « erreur Y_{ijk} »pour un site i, une date j et une feuille k donnés, représente la différence entre la contamination mesurée et la contamination prédite par SeptoLIS®. Elle a pour unité « l'intensité », le pourcentage de surface de feuille occupée par les symptômes de la maladie.

$$Y_{ijk} = mesure_{ijk} - predictions_{ijk}.$$

***Y_{ijk}*: Le C'est l'erreur au site i, à la date j et sur la feuille k.**

***i* : Le site varie de 1 ,…,N.** Avec N=191 nombre de sites mesuré au total.

***j* : La date (jour/mois/année) varie de 1,…,Ji.**
Nombre de dates auxquelles on a mesuré un site i donné Ji $\leq$ 27. Au maximum pour un site donné il y a eu 27 dates ; en moyenne un site i donné est mesuré à 7 dates différentes.

***k:* La feuille varie de kij,…,kij+2.**
Avec k variant de 1 à 4, pour un site i donné à une date j donnée, on aura 3 mesures faites sur 3 étages foliaires différents : Les 3 feuilles successives kij., kij+1, kij+2.

Mes variables explicatives disponibles sont : la région, le numéro de la feuille, l'année, le précédent cultural, la note de résistance et la note de précocité de la variété.

1.3) Stratégies employées :

Nous cherchons à prédire « l'erreur de prédiction », Yijk, faite par SeptoLIS® à l'aide d'un modèle linéaire généralisé. Puis nous rajouterons cette erreur à la prédiction SeptoLIS® pour la corriger.

2) Matériel et méthodes employées.

Dans un premier temps nous avons vérifié s'il existait une corrélation des erreurs de SeptoLIS® dans l'espace à l'aide d'analyses de krigeage. Dans un second temps nous avons essayé de prendre en compte les variations temporelles et nous avons visualisé les interactions potentielles entre nos variables. Enfin une fois la phase d'observation des variables finie, nous avons évalué les capacités de différents types de modèles linéaires généralisés pour prédire l'erreur de SeptoLIS®. Finalement, nous avons testé le modèle en validation croisée et nous avons corrigé les prédictions de SeptoLIS® avant d'examiner les nouvelles performances de « SeptoLIS® corrigé ».

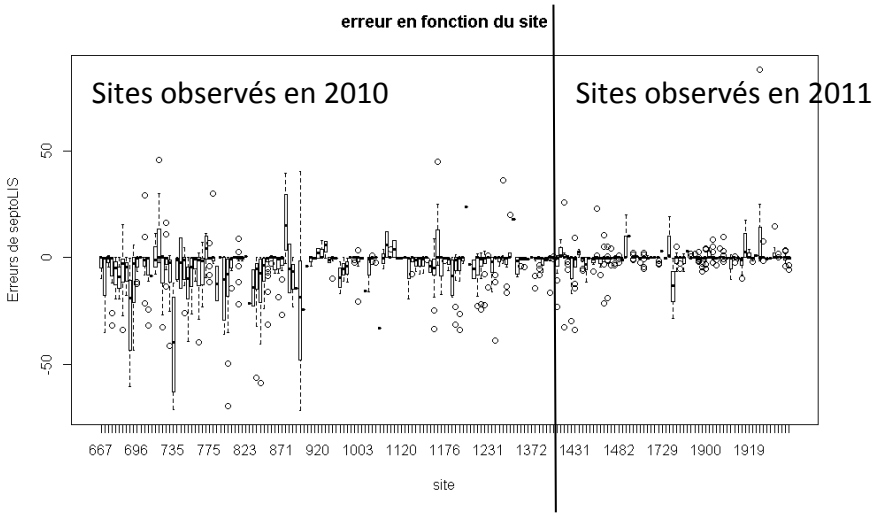
2.1) Comment prendre en compte la dimension spatiale des erreurs de prédictions?

Dans un premier temps, on a voulu vérifier si **les erreurs faites par SeptoLIS® étaient spatialisées**. Bien que le modèle prenne en compte la météo locale, on pensait pouvoir détecter des « gradients d'erreurs » différents selon la zone géographique. Grâce au krigeage, on pensait pouvoir faire un variogramme, à l'aide duquel on pourrait savoir à n'importe quel endroit la valeur probable de l'erreur. **Pour réaliser le krigeage, j'ai utilisé les packages geoR et stats.**

Bien que nous n'ayons pas réussi à détecter de corrélation spatiale en tant que telle grâce au krigeage, après observation des erreurs (fonction plot de geoR), il apparaît cependant que dans certaines zones en France, on trouve plus d'erreurs que dans d'autres, **nous avons donc créé une variable région à 5 modalité qui découpe la France en 5 zones**. ***(ANNEXE II.3)***.

Les erreurs varient beaucoup d'un site à l'autre. ***(figure c).***Donc pour essayer de nous affranchir des variations spatiales des erreurs, nous avons parfois mis en aléatoire l'effet « site de mesure » dans les modèles linéaires généralisés.

Figure c : Graphique des erreurs de SeptoLIS® en fonction du numéro du site.



2.2) Comment prendre en compte la dimension temporelle des erreurs de prédictions?

Les erreurs varient nettement d'une période à l'autre, s'aggravant au cours de la saison. Nous avons utilisé les variables âge de la feuille et stade de la plante à la date d'observation pour prendre en compte le fait que les erreurs varient au cours du temps. Dans les modèles testés, nous avons mis un effet quadratique sur ces variables (âge + âge^2 + stade + stade^2+…), car la variation des erreurs en fonction du stade ou de l'âge de la feuille ne semble pas être linéaire à première vue ***(figure d & e)***. Plus l'âge de la feuille augmente, plus la contamination potentielle est importante. On peut voir sur ces figures que plus l'âge augmente et plus l'erreur de prédiction augmente.

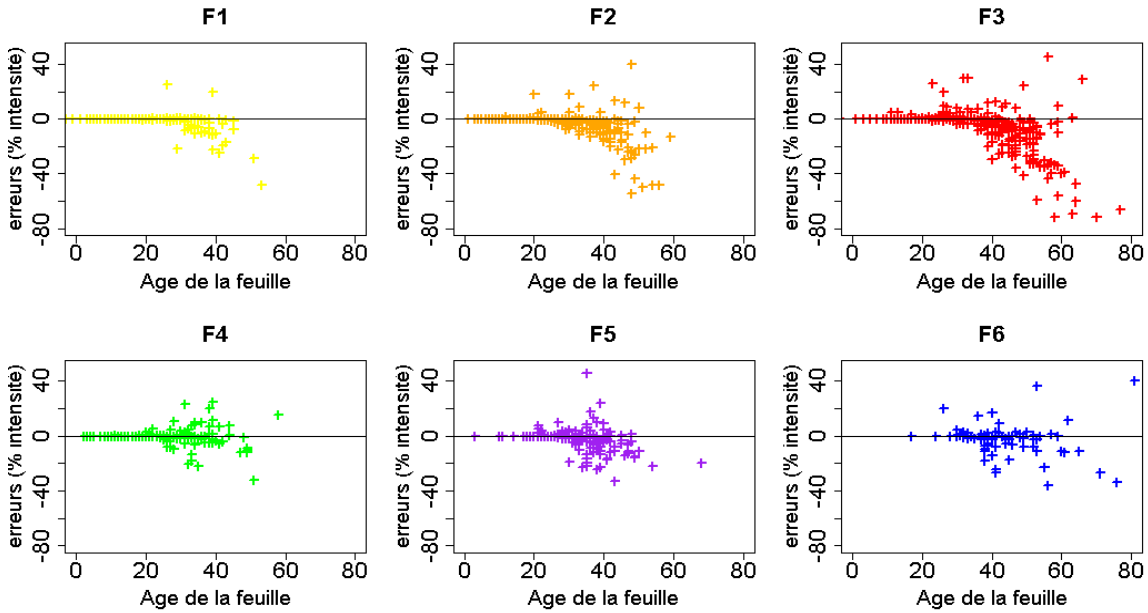


Figure d : Erreurs de SeptoLIS® en fonction de l'âge de la feuille observée.

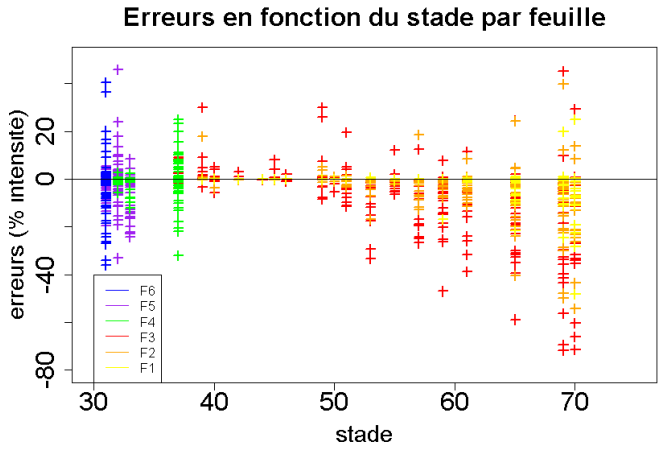


Figure e : Erreurs en fonction du stade de la plante par feuille.

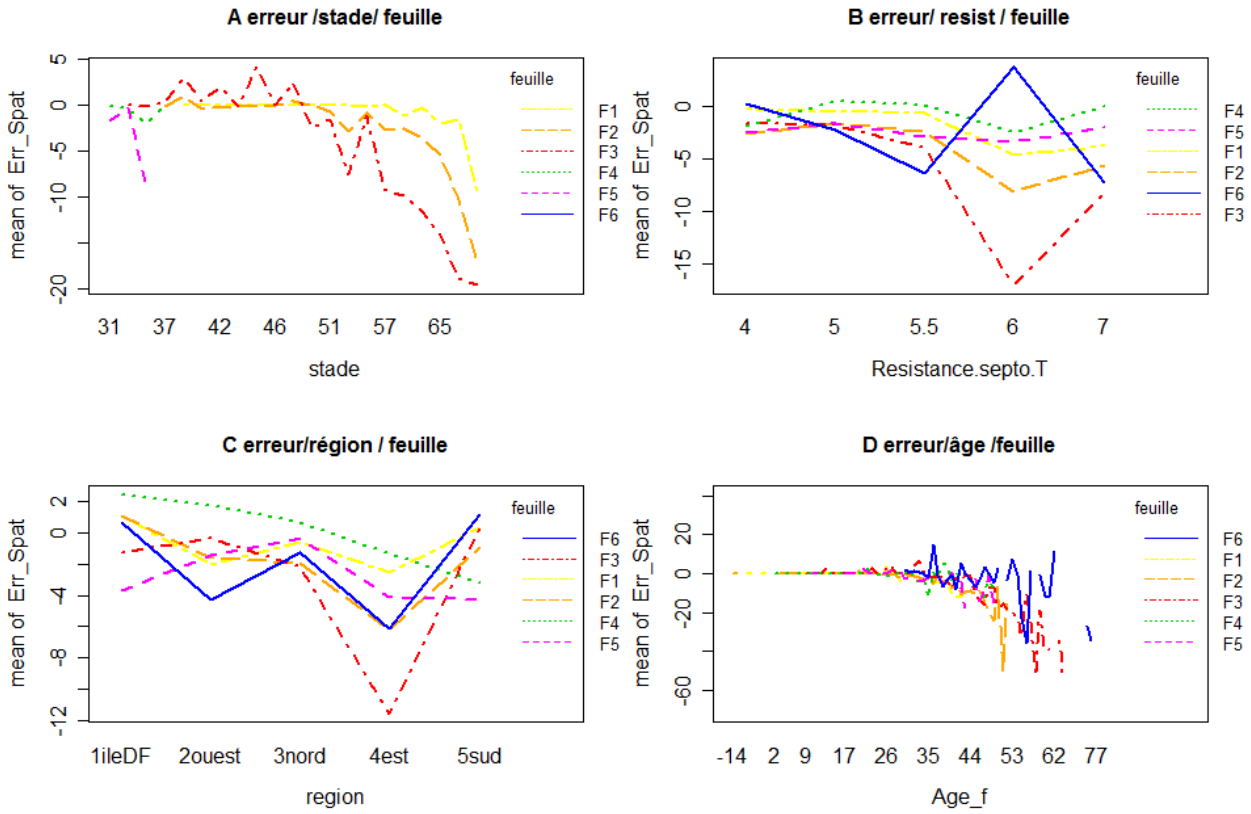
2.3) Présence d'interactions potentielles entre nos différentes variables:

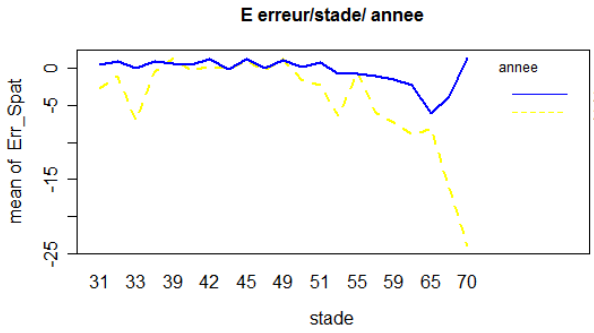
Nous avons également testé la présence d'interactions du second degré. ***(figure f)***. On voit que les données concernant les feuilles F6, F5 et F4 sont situées essentiellement pour des plantes jeunes stade 30 à 37 ***(figure f : A)***. On commet des erreurs plus lourdes sur les prédictions concernant la F3 et cela semble être le cas tout particulièrement dans les données situées à l'est ***(figure f : C)***.

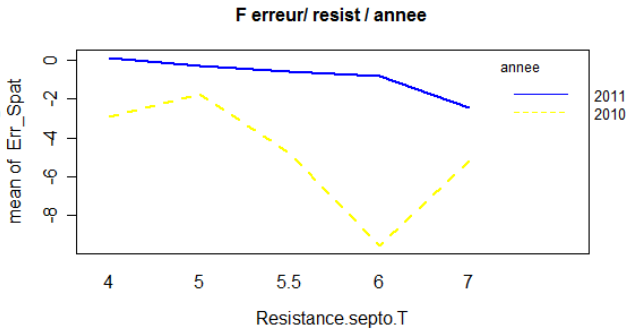
Il semble bien qu'il y ait un fort effet année ***(figure f : D,E,F,G)***. 2011 a été une année moins contaminée et nous retrouvons moins d'erreur cette année-là.

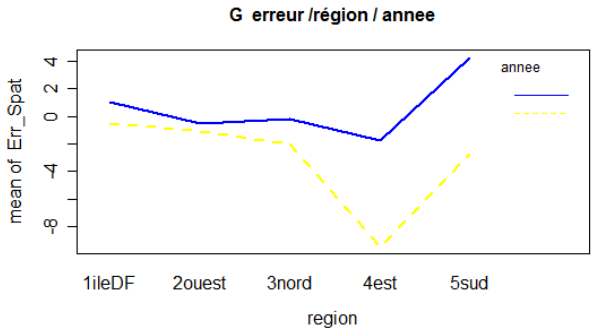
Il ne semble pas y avoir d'interaction entre précédent et nos autres variables. ***(figure f : H,I,J,K)*.**

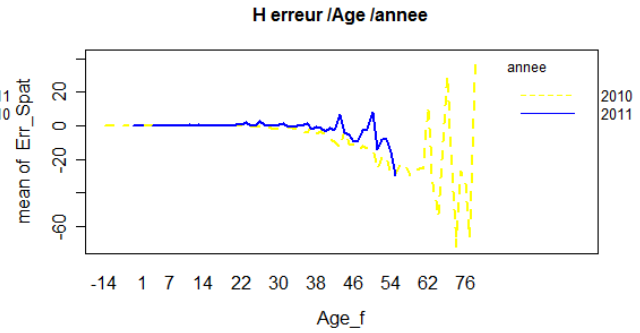
Figure f : Visualisation des interactions potentielles : graphiques représentant les erreurs de septoLIS® en fonction des différentes variables.

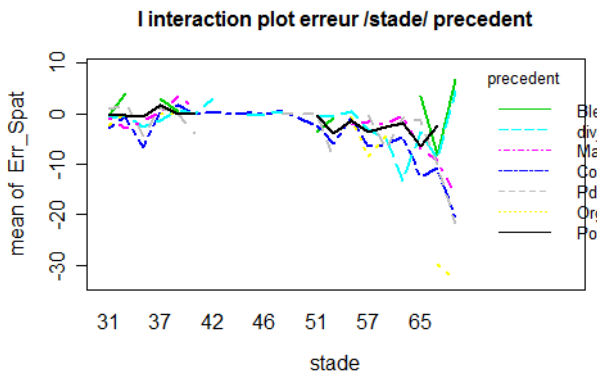


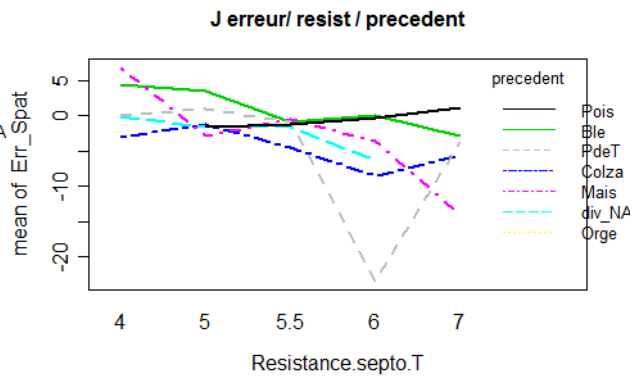

E erreur/stade/ annee
mean of Err_Spat
0
-5
-15
-25
31 33 39 42 45 49 51 55 59 65 70
stade
annee
2011
2010

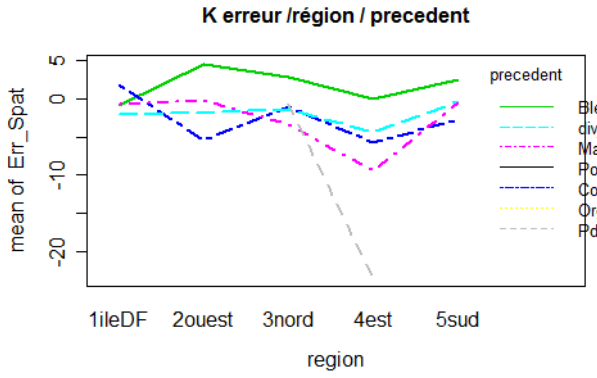

F erreur/ resist / annee
mean of Err_Spat
0
-2
-4
-6
-8
4 5 5.5 6 7
Resistance.septo.T
annee
2011
2010

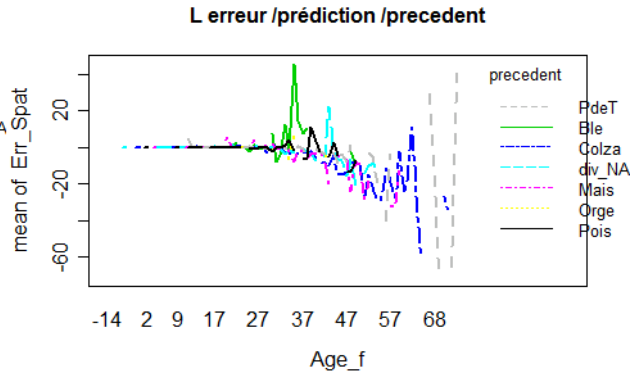

G erreur /région / annee
mean of Err_Spat
4
2
0
-4
-8
1ileDF 2ouest 3nord 4est 5sud
region
annee
2011
2010


H erreur /Age /annee
mean of Err_Spat
20
0
-20
-60
-14 1 7 14 22 30 38 46 54 62 76
Age_f
annee
2010
2011


I interaction plot erreur /stade/ precedent
mean of Err_Spat
10
0
-10
-20
-30
31 37 42 46 51 57 65
stade
precedent
Ble
div_NA
Mais
Colza
PdeT
Orge
Pois


J erreur/ resist / precedent
mean of Err_Spat
5
0
-10
-20
4 5 5.5 6 7
Resistance.septo.T
precedent
Pois
Ble
PdeT
Colza
Mais
div_NA
Orge


K erreur /région / precedent
mean of Err_Spat
5
0
-10
-20
1ileDF 2ouest 3nord 4est 5sud
region
precedent
Ble
div_NA
Mais
Pois
Colza
Orge
PdeT


L erreur /prédiction /precedent
mean of Err_Spat
20
-20
-60
-14 2 9 17 27 37 47 57 68
Age_f
precedent
PdeT
Ble
Colza
div_NA
Mais
Orge
Pois

2.4) Sélection du meilleur modèle linéaire généralisé :

En partant d'un **modèle linéaire généralisé complet avec toutes les interactions d'ordre 2, nous avons sélectionné les variables par stepwise selon le critère BIC** (direction backward/forward) ***(tableau A)***.

Tableau A : Tableau des modèles linéaires généralisés complet et sélectionné après stepwise :

	Variables :	***AIC***	***BIC***
Modèle nul	*ErreurSeptoLIS®~1*	10334	10344
Modèle complet	*ErreurSeptoLIS® ~* **(***feuille + stade +Resistance.a septoriose +* ***Precocite.Epiaison*** *+ Annee+Age + region +****precedent*** **)^2** *+ I(Age^2)+ I(stade^2)*	9428	10327
Modèle sélec-tionné	*ErreurSeptoLIS® ~ feuille + stade +* I(*stade^2*) *+Resistance.a septoriose + Annee+ Age + region + stade:Age+ annee:Age + Resistance.:Age + Age:region*	9528	9648

Hypothèses :

- $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$
- $Cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'})=0.$

Après observation des résidus, il apparaît qu'ils ne sont pas normaux ni hétérosédastiques ***(figure g & h)*****.** Nous avons cherché à prendre en compte cela en: **ajoutant des hypothèses sur la variance des résidus** : soit on a considéré que la variance des résidus était une fonction puissance(ou exponentielle) de l'âge ou de la valeur prédite. (package nlme, fonction varpower ou varexp).

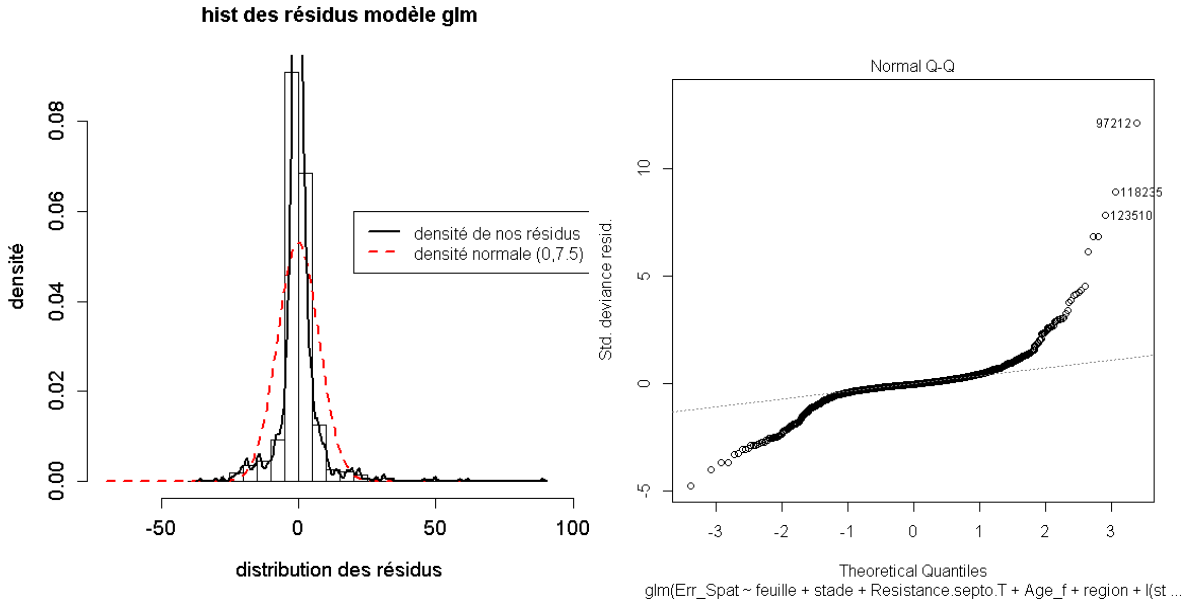


Figure g : histogramme des résidus du modèle de prédiction d'erreur (glm après sélection stepwise) (à gauche) et QQ plot (à droite).

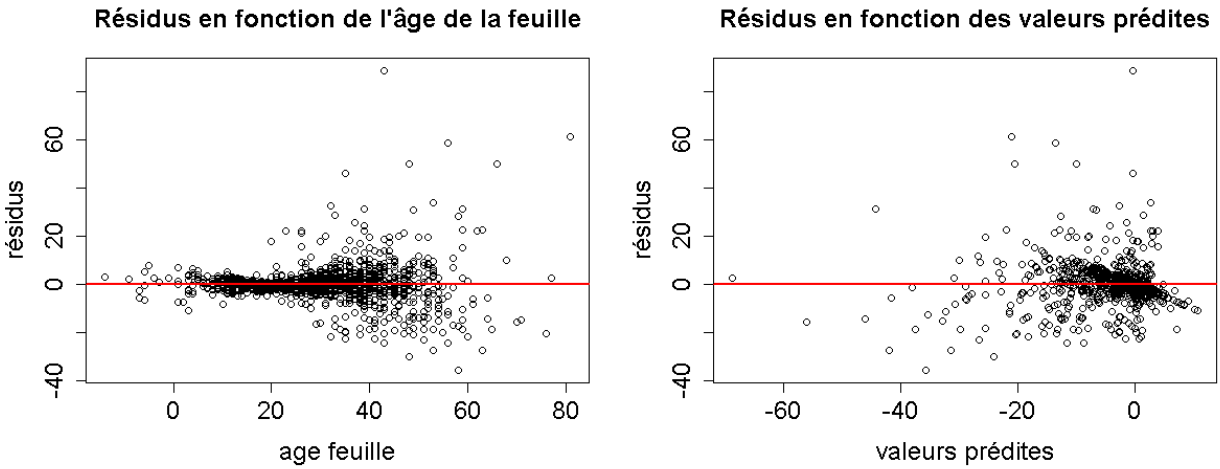


Figure h : Résidus du modèle glm sélectionné après stepwise en fonctionde l'âge (à gauche) et des valeurs prédites (à droite).

2.5) Principe des modèles mixtes et justification de leur utilisation :

Pour des données ayant des dépendances spatiales ou temporelles, il existe souvent des **corrélations intra-groupes**. Les erreurs intra-groupes ne sont alors ni indépendantes, ni homosédastiques. **Les modèles à effet mixtes** ***(figure i)*** permettent de modéliser cela, en définissant les corrélations et la structure de variance des résidus intra-groupe. **(@5,(18)).**

ANNEXE 2.8 : Présentation théorique des modèles mixtes (source (18)).

$$Y_{ij}=\underbrace{\beta_1 x_{1ij}+....+\beta_p x_{pij}}_{\text{Effets fixes}}+\underbrace{b_{i1}z_{1ij}+...+b_{iq}z_{qij}}_{\text{Effets aléatoires}}+\varepsilon_{ii}$$

$$b_{ik}\sim N(0,\psi_k^2)\ \ cov(b_k, b_{k'})=\psi_{kk'}$$

$$\varepsilon_{ii}\sim N(0, \sigma^2\lambda_{ijj})\ \ cov(\varepsilon_{ii}, \varepsilon_{ii'})= \sigma^2\lambda_{ijj'}$$

- Y_{ij} est la variable réponse pour la jième observation du groupe i.
- $\beta_1,\dots,\beta_p$ sont les coefficients des effets fixes, ce sont les mêmes quel que soit le groupe.
- $x_{1ij},\dots,x_{pij}$ sont les régresseurs des effets fixes pour l'observation j du groupe i.

  Généralement le premier correspond à la constante : $x_{1ij}=1$.
- $b_{i1},\dots,b_{iq}$ sont les coefficients des effets aléatoires pour un groupe i donné. Ils suivent une loi normale multivariée.
- les variances ψ_k^2 et les covariances $\psi_{kk'}$ entre effets aléatoires sont supposés constantes d'un groupe à l'autre.
- $z_{1ij},\dots,z_{qij}$ sont les régresseurs des effets aléatoires.
- ε_{ii} Erreurs intra groupe i pour l'observation j. elles suivent une loi normale multivariée.
- $\sigma^2\lambda_{ijj'}$ sont les covariances entre les résidus du groupe i. Si les observations au sein d'un groupe sont indépendantes les unes des autres et ont une variance constante, alors on aura $\lambda_{ijj}=\sigma^2$ et $\lambda_{ijj'}=0$ pour $j\neq j'$, si par contre les mesure au sein d'un même groupe sont des données mesurées sur un même individu au cours du temps, alors on peut imaginer que les λ devront refléter l'autocorrélation des erreurs.

Dans notre cas, un site correspond à une commune sur laquelle plusieurs mesures ont été faites à différents stades de développement de la plante. Les mesures faites sur un même site ne sont vraisemblablement pas indépendantes *(tableau B)*. **Les erreurs moyennes varient certainement d'un site à l'autre, nous avons donc mis l'effet site en aléatoire et ainsi attribué un intercept différent à chaque site**.

Tableau B : Extrait de notre tableau de données : les erreurs de prédictions à un stade j sur une feuille k du site i=837.

Erreur SeptoLIS	Stade	Age	feuille	site	résis-tance	précé-dent	région
Yijk	j	(lié à la feuille)	k	i	(lié au site)	(lié au site)	(lié au site)
0	32	26	F4	837	7	Colza	Nord
0	32	13	F3	837	7	Colza	Nord
…	…	…	…	837	7	Colza	Nord
-5.538	46	30	F3	837	7	Colza	Nord
-1.154	46	19	F2	837	7	Colza	Nord
0	46	6	F1	837	7	Colza	Nord
-14.529	50	36	F3	837	7	Colza	Nord
-5.268	50	25	F2	837	7	Colza	Nord
-1.004	50	12	F1	837	7	Colza	Nord
…	…	…	…	837	7	Colza	Nord
-80.195	65	51	F3	837	7	Colza	Nord
-72.496	65	40	F2	837	7	Colza	Nord
-52.404	65	27	F1	837	7	Colza	Nord

2.6) Résumés des différents modèles testés :

Si l'on teste toutes les combinaisons possibles de modèles, cela revient à tester 20 modèles différents. A chaque fois, on note l'AIC, le BIC et le RMSEP (obtenu sans et avec « validation croisée » comme décrite en 2.5) ***(tableau C)***.

Tableau C : différents modèles testés et leurs performances.

	modèle stat	variables	effet aleatoi.	hétérosed. Resid.	AIC	BIC	RMSEP (sans CV)	RMSEP (CV 7 jours)
1	glm	0	0	0	10334	10345	10.22	10.51
2	gls	0	0	varExp(fitted(.))	10337	10353	10.22	10.51
3	gls	0	0	varExp(stade)	8033	8048	10.61	10.66
4	gls	0	0	varPower(fitted(.))	10337	10353	10.22	10.51
5	gls	0	0	varPower(stade)	7880	7896	10.58	10.62
6	glm	select.	0	0	9528	9648	**7.51**	**8.40**
7	gls	select.	0	varExp(fitted(.))	9142	9267	**8.10**	NA
8	gls	select.	0	varExp(stade)	7931	8056	9.76	10.18
9	gls	select.	0	varPower(fitted(.))	NA	NA	NA	NA
10	gls	select.	0	varPower(stade)	7970	8095	9.76	10.28
11	lme	0	site	0	10181	10196	8.61	NA
12	lme	0	site	varExp(fitted(.))	NA	NA	NA	NA
13	lme	0	site	varExp(stade)	7897	7918	10.55	NA
14	lme	0	site	varPower(fitted(.))	NA	NA	NA	NA
15	lme	0	site	varPower(stade)	7872	7893	10.58	NA
16	lme	select.	site	0	9511	9636	**6.68**	**7.13**
17	lme	select.	site	varExp(fitted(.))	NA	NA	NA	NA
18	lme	select.	site	varExp(stade)	7904	8035	9.76	9.49
19	lme	select.	site	varPower(fitted(.))	NA	NA	NA	NA
20	lme	select.	site	varPower(stade)	7836	7967	10.02	9.70

2.7) Enoncés du modèle choisi et observation des prédictions:

Ci-dessous nous décrivons le modèle linéaire généralisé mixte N°16 ***(tableau C)***.

Modèle 16:

$$Y_{ijk} = (\beta_0 + b_{i0}) + Feuille_k + Région_i + Annee_j + \beta_{1.1}.\underline{stade}X_{1ij} + \beta_{1.2}.\underline{stade}X_{1ij}^2 + \beta_2(1 + R_i + A_j).\underline{age}X_{2k} + \beta_3.\underline{Resist}X_{3ij} + \beta_4\,(ageX_{2k}.stadeX_{1ij}) + \beta_5\,(ageX_{2k}.ResistX_{3ij}) + \varepsilon_{ijk}$$

- $\mathbf{Y_{ijk}}$ est l'erreur de prédiction faite par SeptoLIS®, à un site i, une date j et une feuille k.
- $\boldsymbol{\beta_0}$ est l'erreur de prédiction moyenne globale. $\mathbf{b_{i0}}$ est l'erreur de prédiction moyenne spécifique à un site i donné, effet aléatoire.
- $\mathbf{F_k}$ est **l'effet feuille** k sur l'erreur de prédiction. $\mathbf{R_i}$ est **l'effet région** du site i sur l'erreur de prédiction

- $\beta_{1.1}$ est le coefficient de **l'effet stade** j de la plante du site i X_{1ij}, $\beta_{1.2}$ est le coefficient du **carré de l'effet stade** X_{1ij}^2
- β_2 est le coefficient de **l'effet âge de la feuille** X_{2k} . L'effet de l'âge sur l'erreur de prédiction varie en fonction de la région considérée (**interaction entre la région et l'âge**) et est aussi différent d'une année à l'autre (**interaction année-âge**).
- β_3 est le coefficient **la résistance de la plante à la septoriose** X_{3ij}**. L'effet de l'âge sur la résistance est différent selon la note de résistance de la plante**, cela est pris en compte dans l'interaction âge-résistance**.** β_5 est le coefficient de **cette interaction**.
- β_4 est le coefficient de **l'interaction** entre le stade de développement de la plante et âge de la feuille. Si une feuille est âgée de 10 jours sur une plante stade précoce, il y a généralement une moindre erreur de prédiction (Z30), que sur une feuille de 10 jours sur une plante développée (Z45) *(figure 8)*.
- ε_{ijk} sont les résidus intra-sites du modèles.

Hypothèses :

- $b_{i0} \sim N(0, \sigma_0^2)$ et $cov(b_{i0}, b_{i'0})=0$. (les variances σ_0^2 et les covariances entre effets aléatoires sont supposés constantes d'un sites à l'autre)
- $\varepsilon_{ijk} \sim N(0, \sigma_{\varepsilon i}^2)$ et $Cov(\varepsilon_{ijk}, \varepsilon_{ij'k'})=0$. $\sigma_{\varepsilon i}^2$ la variance des résidus intra site est supposée constante dans le cas du modèle 16. On suppose aussi qu'il n'y a aucune corrélation entre les errueurs d'un même site.

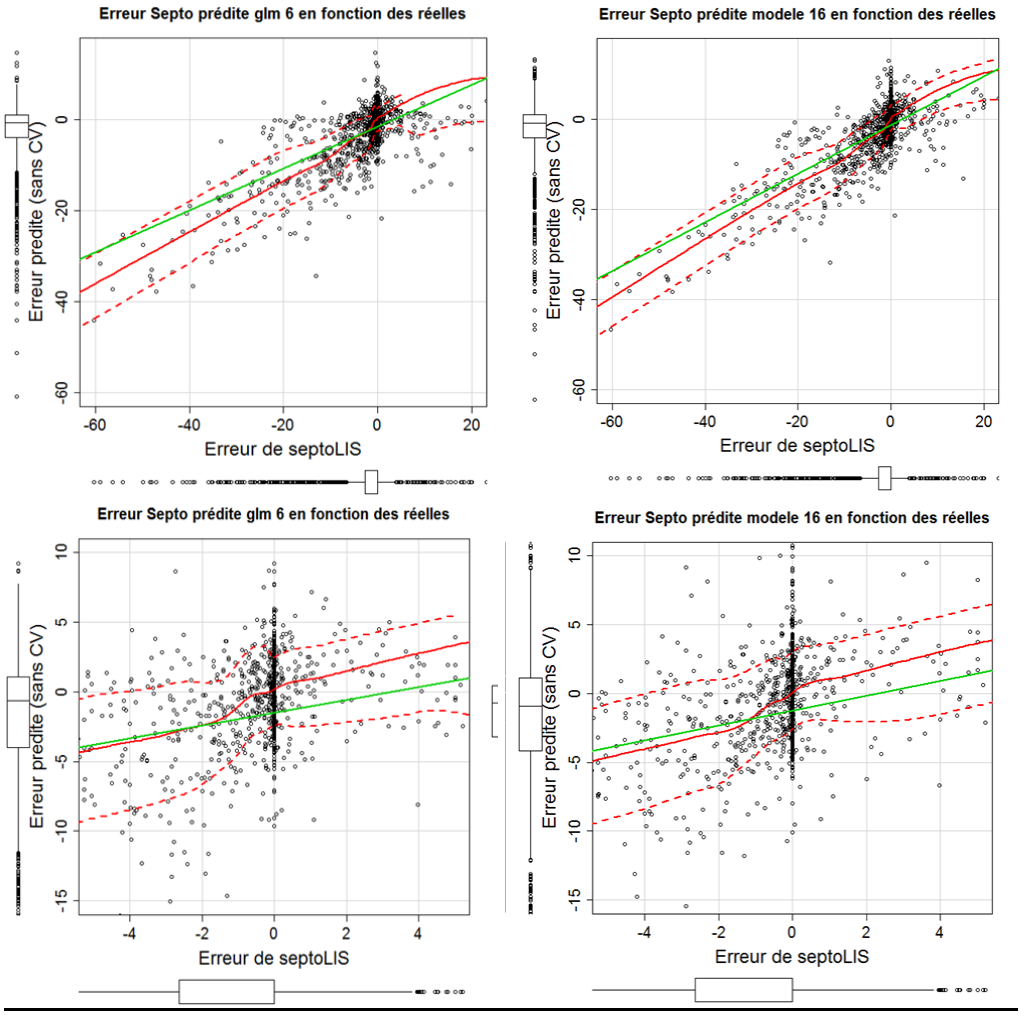


Figure j : observation des erreurs prédites en fonction des erreurs réelles de SeptoLIS®, on peut voir le modèle 6 colonne de gauche et le 16 à droite. En haut on observe la totalité des erreurs et en dessous on a un zoom autour des petites erreurs.

Si l'on regarde la ***figure j***, les modèles 6 et le 16 font d'assez bonnes prédictions sur les erreurs importantes, mais ils sont plutôt mauvais quand il s'agit de prédire une petite erreur de SeptoLIS® ou quand SeptoLIS® ne fait pas d'erreur. Le modèle 6 ainsi prédit des erreurs allant de -5 à 5%, alors que SeptoLIS® n'en a pas fait (x=0) et le 16 lui prédit lui des erreurs allant de -5 à 10%, alors qu'il n'y en a pas.

2.8) Principe de la validation croisée utilisée:

Afin de tester les performances des modèles créés pour prédire les erreurs de septoLIS® (objectif 1) ou les fréquences de contamination (objectif 2), il faut se mettre en conditions réelles d'utilisation. Dans la mesure où nos données sont chronologiques et où l'on souhaite prédite une date ultérieure, une validation croisée classique n'est pas la méthode la plus adaptée. ***Dans cette partie je décris comment j'ai réalisé les validations croisées.***

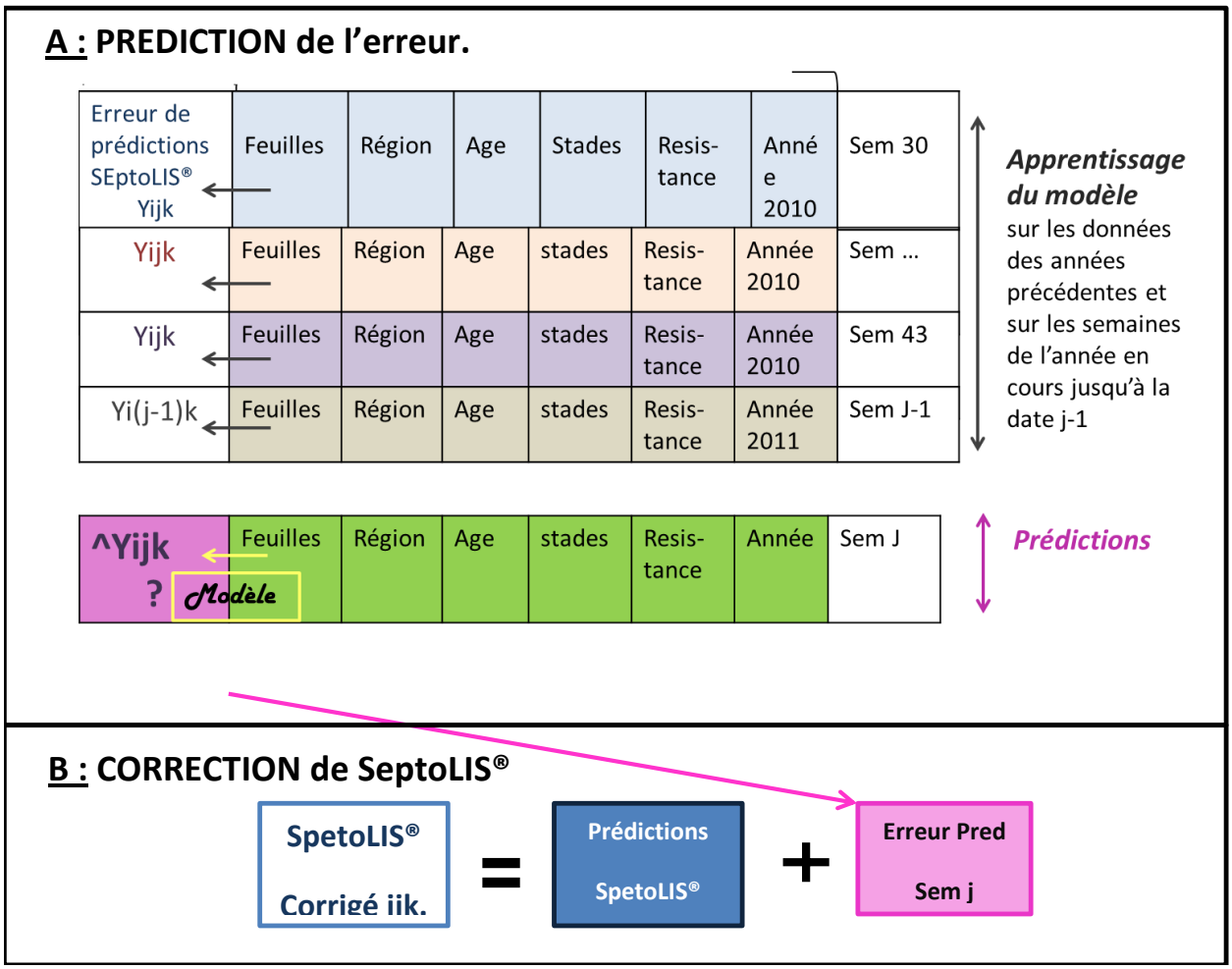


Figure k: Illustration de la validation croisée utilisée pour prédire l'erreur de SeptoLIS® (A) et Utilisation de cette « erreur de septoLIS® prédite » pour corriger SeptoLIS® (B).

La *figure k A* illustre la stratégie utilisée pour tester les capacités prédictives du « modèle d'erreur de spetoLIS® ». Pour estimer les paramètres du modèle j'ai utilisé les données observées sur une année entière + une semaine de l'année à prédire au minimum (par exemple on utilise tout 2010 et la semaine d'observation numéro 1 de 2011). Je prédis ensuite l'erreur de SpetoLIS® pour la semaine suivante (semaine 2 de 2011). Je peux ensuite comparer mon erreur prédite et l'erreur réelle connue. Pour prédire l'erreur probable en semaine 3, j'utilise 2010 et les deux premières semaines de 2011 connues pour ajuster mon modèle, ensuite les informations sur les varaibles de la semaine 3 me servent pour prédire.

Une fois que l'on a prédit une erreur de prédiction que septoLIS® fait à la semaine J, on va vouloir la rajouter à la prédiction SeptoLIS® J, afin de la corriger *(figurek B).* J'ai donc rajouté mon « erreur prédite » par validation croisée, à la sortie de SeptoLIS® pour obtenir une « prédiction septoLIS® corrigée ». A la suite de cela on peut recalculer les performances de « SeptoLIS® corrigé » et voir si l'on a amélioré nos performances *(tableau 2).*

3. Résultats :

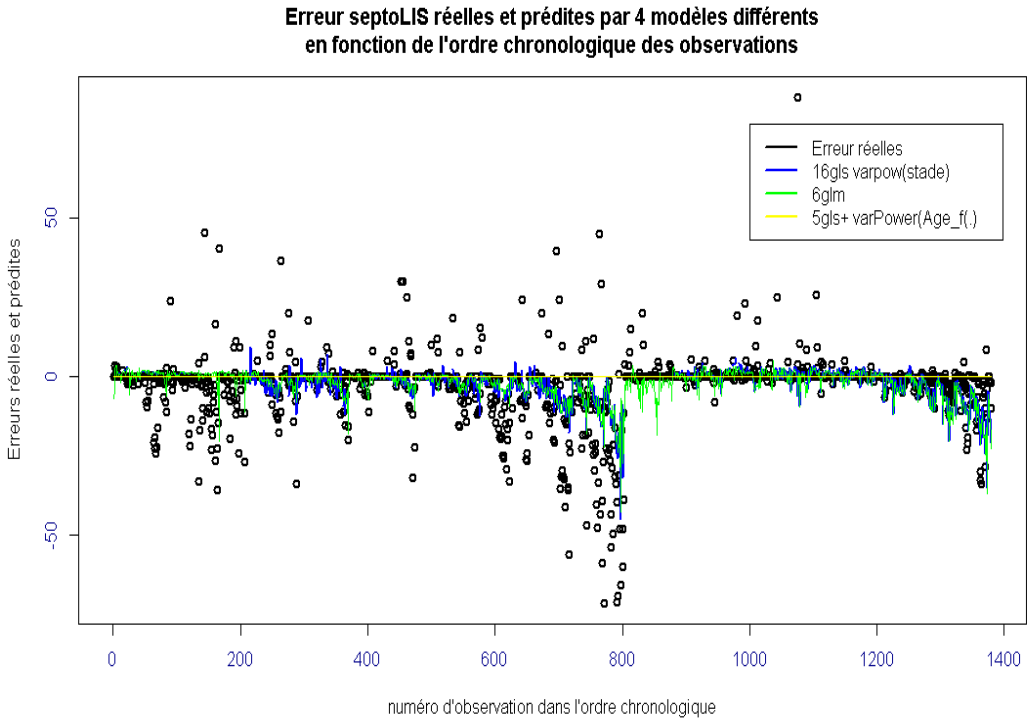


Figure l :Erreurs de prédictions de SeptoLIS® rélles superposées aux prédictions faites par validation croisées à l'aide de nos 3 modèles.

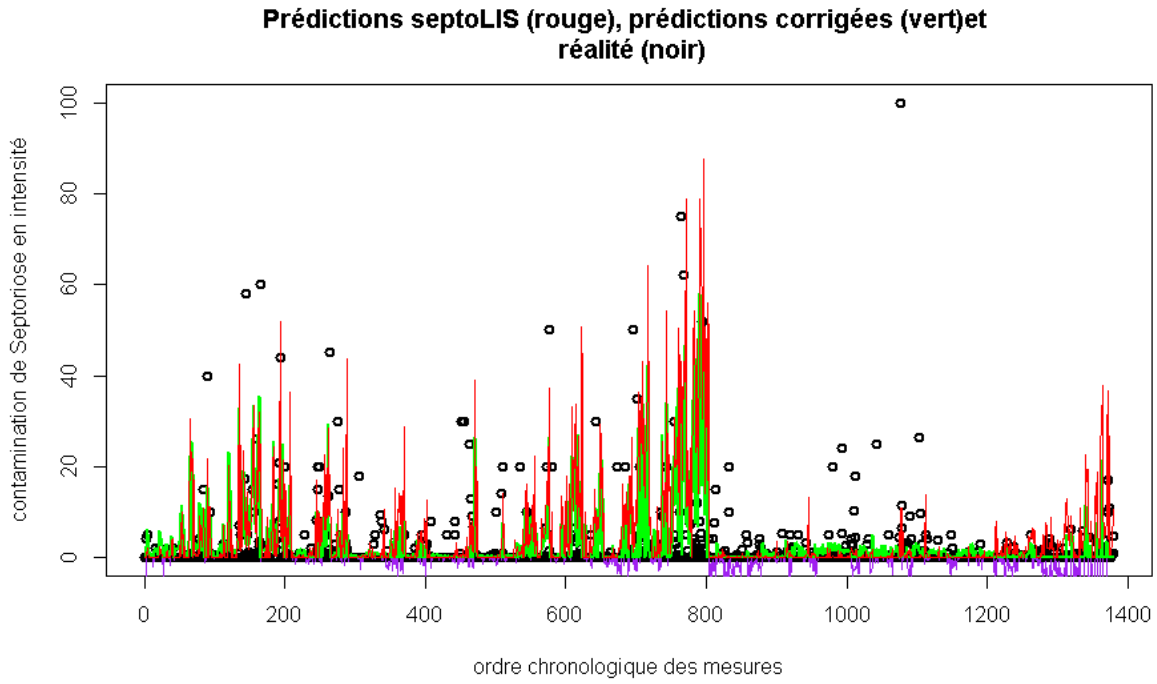


Figure m : Contaminations réellement observées (en noir), superposées aux Prédictions de contamination de SeptoLIS® avant notre correction en rouge , et après en vert.

Nous avons prédit les erreurs, mais les erreurs jusqu'à une certaine limite. Quand celle-ci dépassent 50%, nous n'arrivons pas à les prédire parfaitement comme on peut le voir en ***figure l***. Une fois le modèle d'erreur réalisé, nous avons ajouté les erreurs prédites par validation croisée aux prédictions septoLIS® initiales. Nous avons ensuite visualisé les nouvelles prédictions, que l'on peut voir dans le grphique ***figure m***. On peut lire ci- dessous les performances initiales avant correction et après correction. Nous avons diminué l'erreur moyenne de prédiction de 11% à 8.5% (***figure n***).

$$\textbf{RMSEP intial}=\sqrt{\frac{1}{1100}\sum(Y_{ijk})^2}=10.58\ \%$$

$$\textbf{RRMSEP post corrections (ac glm6)}=\sqrt{\frac{1}{1100}(\text{"}Yijk\ corrigées\text{"})2}=8.55\%$$

Figure n : les erreurs de prédictions de SeptoLIS®avant et après correction.

On transforme maintenant les données de manière à ce que toutes les contaminations supérieures à 3% soient considérées comme appartenant à la classe contaminée « 1 »et en dessous à la classe des saines « 0 ». On peut voir que notre correction a baissé la sensibilité et la spécificité par rapport aux performances de SeptoLIS® sans correction.
On détecte donc moins bien les plantes saines. On pourrait imaginer créer une correction en deux étapes, une étape de détection de la contamination et si c'est le cas, une correction à l'aide de notre modèle glm6.

Tableau D: Matrices de confusion de SeptoLIS® initial en haut et de SeptoLIS® corrigé en bas.

		PREDICTION	
	Minitial	1	0
REALITE	1	293 61%	181
	0	245	661 72%

	M6	1	0
REALITE	1	246 52%	228
	0	328	578 64%

BIBLIOGRAPHIE.

SITES WEB :

@1 *Chambre d'agriculture*, Agriculture et territoire. Consulté le 04 21, 2012, sur **http://www.chambre-agriculture.fr/thematiques/ecophyto-2018/tous-les-bulletins-de-sante-du-vegetal-par-region/**

@2 *La France agricole*, les productions végétales. Consulté le 02 05 2012, sur **http://www.lafranceagricole.fr/l-agriculture/productions-vegetales-artFa/cereales-19871.html#1**

@3 *ARVALIS*, Outil septoLIS®. Consulté le 02 05, 2012, sur **http://www.septolis.arvalisinstitutduvegetal.fr/**

@4 National Center for Atmospheric Research, *Hazardous weather Testbed Model Evaluation*. Consulté le 7-05, 2012, sur **http://verif.rap.ucar.edu/eval/hwt/2011/traditional_eval.php**

@5 *INSA Toulouse et L'institut de mathématiques de Toulouse*, Modèles pour données répétées. Consulté le 21-05, 2012, sur **wikistat.fr**

@6 *Bouveris N., Brochard M.,Cholet D, Délos M., Legros S., Eychenne N., Savary C., Simonneau D., Soularue A., Weissenberger A.*,Harmonisation des protocoles du réseau d'épidémiosurveillance –Généralités (Domaine des Grandes Cultures) . Consulté le 22 08 2012, sur **http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCIQFjAA&url=http%3A%2F%2Fwww.afpp.net%2Fapps%2Faccesbase%2Fbindocload.asp%3Fd%3D6589%26t%3D0%26identobj%3DyCqgcPnk%26uid%3D57305290%26sid%3D..%26idk%3D1&ei=3dg5UJmjBoW70QXdrICgCw&usg=AFQjCNF1qURNKVcDLgoqBzyCQJmgDjUvqQ&sig2=TMepLjWeqdoEOVMfXaUAOQ**

ARTICLES ET AUTRES TYPES D'OUVRAGES **:**

(1) *ARVALIS – institut du végétal*, (2003). **Stades du blé**, 68 p.

(2) *Suffert F., Sache I. et Lannou C.* (2011). **Early stages of septoria tritici blotch epidemics of winter wheat: build-up, overseasoning, and release of primary inoculum**. Plant Pathology, N° 60, pp.166–177.

(3) *Robert C., Bancal M-O., Nicolas P., Lannou C. and Ney B.* (2004). **Analysis and modelling of effects of leaf rust and Septoria tritici blotch on wheat growth. Journal of Experimental Botany**, Vol. 55, N° 399, pp. 1079-1094.

(4) *Gouache D., Bensadoun A., Brun F.; Pagé C., Makowski D., Wallach D.* (en cours de publication). **Modelling climate change impact on Septoria tritici blotch (STB) in France: accounting for climate model and disease model uncertainty.**

(5) *McKendry A.L., Henke G.E.et Finney P.L.* (1995). **Effects of Septoria Leaf Blotch on Soft Red Winter Wheat Milling and Baking Quality.** Cereal Chemistry. Vol. 72 , N°2, pp.142-146.

(6) *EyalZ., A.L.Scharen, J.M.Prescott et M. van Ginkel (1987).* **The septoria diseases of wheat : Concept and method of disease management.** CIMMYT, Mexico, 46 p.

(7) *Baccar R., Fournier C., Dornbusch T., Andrieu B., Gouache D. et Robert C.* (2011). **Modelling the effect of wheat canopy architecture as affected by sowing density on Septoria tritici epidemics using a coupled epidemic–virtual plant model.** Annals of Botany. Vol 108, pp. 1179–1194.

(8) *Gouache D. et Couleaud G.* (2009). **Le positionnement des traitements fongicides : enjeu pour la septoriose et interet du modele « septolis »**. In AFPP – 9ème conférence internationale sur les maladies des plantes, Tours, 8 et 9 décembre 2009**.**

(9) *Gouache D.* (2010). **Fongicides céréales : positionnement des traitements ciblant la septoriose : un enjeu de 5 q/ha.** Perspectives agricoles, N°365, pp.42-45.

(10) *Gibert C., Gouache D., Oumouhou N., Brun F., Piraux F., Aubertot J-N., Wallach D.* (en cours de publication). **Calibration and evaluation of a model of Septoria leaf blotch for use in a decision support tool.** Soumis à Phytopathologie.

(11) *Simonneau D., Taupin P., Couleaud G., Mauffras J-Y., Robin* N. (2011). **Vigicultures ® Mode opératoire observation Blés d'hiver**. Version N°9.

(12) *van Maanen A. et Xu X-M.* (2003). **Modelling plant disease epidemics.** European Journal of Plant Pathology, Vol. **109, pp.** 669–682.

(13) *Audsley E., Milne A. & Paveley N.* (2005). **A foliar disease model for use in wheat disease management decision support systems.** Annals of Applied Biology, Vol.147, pp. 161–172.

(14) *Crombez J.*, (2007). **Epidemiologie de la septoriose sur ble : premiers éléments pour la construction d'un modele predictif.** Mémoire du diplôme d'ingénieur agronome, Agrocampus Rennes, Rennes, 76p.

(15) *M. El Jarroudi, P. Delfosse, H. Maraite, L. Hoffmann et B. Tychon* (2009). **Assessing the Accuracy of Simulation Model for Septoria Leaf Blotch Disease Progress on Winter Wheat.** Plant Disease, Vol. 93, No. 10, pp.983-992.

(16) *Gate* (1995). **Ecophysiologie du blé.** T. Lavoisier, Ed.,429 p.

(17) *R Development Core Team* (2011)**. R : A language and Environment for statistical computing. R foundation for statistical computing,** Vienna Austria, http://www.R-project.org.

(18) *Pinheiro J.C., Douglas B.M.* (2000)*.* **Mixed-effects Mehtods and Classes for S and S-PLUS :lme and nlme,** Springer New York Berlin Heidelberg, 528p..

(19) *LaDeau S. L., Glass G. E., Hobbs T. N., Latimer A and Ostfeld R. S.* (2011).**Data-model fusion to better understand emerging pathogens and improve infectious disease forecasting**, Ecological Applications, Vol. 21, No.5, pp.1443-1460.

(20) *Makowski D., Monod H.* (2011). **Analyse statistique des risques agro-environnementaux : Etude de cas.** Springer New York Berlin Heidelberg, collection statistique et probabilités appliquées, 162p..

(21) *Agresti A.* (1990) **Categorical data analysis.** John Wiley & Sons*.*

(22) *Breiman L., Friedman J.H., Olshen R., Stone C.J.*, (1984). **Classification and regression trees**. Belmont CA, Wadsworth.

(23) *Cornillon P-A, Guyader A., Husson F., Jégou N., Josse J., Kloareg M., Matner-Lober E., Rouvière L.* (2010). **Statistique avec R, 2ème édition augmentée**. Pratique de la statistique, presses universitaire de Rennes.

(24) *Hugues H.,McRoberts N., Burnett F.J.*(1985). **Descision-making and diagnosis in disease management.** Plant pathology, 49:140-145.