Modèle bayésien pour réaliser une analyse roc avec un indicateur de risque de sclérotinia du colza

David D. Makowski, Jean-Baptiste Denis, L. Ruck

Du 29 Mai au 2 Juin 2006

38e Journées de Statistique de la SFdS

PROGRAMME DES JOURNÉES ET RÉSUMÉS DES CONFÉRENCES

SOCIETE FRANÇAISE DE STATISTIQUE

EDF

Modèle Bayésien pour réaliser une analyse ROC avec un indicateur de risque du sclérotinia du colza

David Makowski, Jean-Baptiste Denis (INRA), Laurent Ruck (CETIOM)

L'objectif de cet article est d'évaluer les performances d'un indicateur développé par les agronomes pour décider de traiter ou non les parcelles de colza infestées par le sclérotinia. Cet indicateur correspond au pourcentage de fleurs malades mesuré avant la date du traitement fongicide. La valeur de ce pourcentage est comparée à un seuil de décision et un traitement fongicide est recommandé seulement si le pourcentage est supérieur au seuil. Une base de données incluant des mesures réalisées sur 420 parcelles de colza a été utilisée pour développer 10 modèles Bayésiens reliant la valeur de l'indicateur au niveau d'attaque de la maladie, à la taille de l'échantillon de fleurs prélevées et à un effet régional. Le modèle ayant la plus faible valeur du DIC a été sélectionné et a été utilisé pour déterminer la sensibilité et la spécificité de l'indicateur, puis pour calculer l'aire sous la courbe ROC en tenant compte d'un effet régional et de la taille de l'échantillon de fleurs. Les résultats montrent que la performance de l'indicateur est légèrement supérieure pour la région Centre et qu'elle dépend fortement de la taille de l'échantillon de fleurs.

Analyse de la variabilité spatiale sur 2 maladies des arbres en milieu naturel

Benoît Marçais (INRA)

Deux exemples seront utilisés pour illustrer le type de travail réalisé. Le premier concerne un agent d'oïdium, *Erisiphe alphitoides,* agent pathogène attaquant les feuilles des chênes. C'est un organisme exotique arrivé en Europe au début du siècle. Après une période initiale où les forestiers ont été très inquiétés par son impact, cette maladie a suscité peu d'intérêt du fait de son arrivée tardive en saison de végétation qui limite fortement sa nuisibilité. Le Département de la Santé des Forêts (DSF) a toutefois mentionné une présence importante et inédite de l'oïdium ces dernières années. Le travail présenté a consisté en une analyse des données de la base DSF pour identifier les conditions, les régions et les années où l'oïdium du chêne peut poser problème. Cette base est constituée de mentions de problèmes de santé des forêts faites de façon spontanée ces 15 dernières années par des forestiers entraînés au diagnostic phytosanitaire. L'analyse des données a été réalisée en comparant la répartition spatio-temporelle des mentions d'oïdium à celle d'autres problèmes de santé des chênaies utilisés pour caractériser la population cible. L'analyse montre qu'il existe en France un fort gradient NE-SO dans l'impact de l'oïdium et met en évidence l'existence de plusieurs années à forte épidémie d'oïdium dans le SO. Ces années correspondent à une arrivée très précoce de l'oïdium en saison de végétation et à un climat particulier caractérisé par des hivers très doux et des printemps humides. Une étude rétrospective du climat montre que les années présentant ce climat particulier sont devenues beaucoup plus fréquentes dans la dernière décennie du 20^{ième} siècle.
Le second exemple concerne le dépérissement à *Phytophthora* des aulnes. Il s'agit d'une maladie émergente qui, depuis le début des années 1990, a un impact très fort le long des cours d'eau européens. Afin d'évaluer l'impact à long terme de la maladie sur la démographie de l'aulne, un dispositif a été mis en place le long de la Sarre (Moselle) en 2002. Les aulnes ont été recensés et cartographiés sur 3.5 km de rivière et leur état sanitaire est suivi annuellement. Les données ont pour l'instant été utilisée pour étudier la dispersion de l'agent pathogène sur le segment de rivière. En particulier, nous voulons déterminer la part respective de la dispersion locale entre

MODÈLE BAYÉSIEN POUR RÉALISER UNE ANALYSE ROC AVEC UN INDICATEUR DE RISQUE DE SCLÉROTINIA DU COLZA

David Makowski[1,2], Jean-Baptiste Denis[1] & Laurent Ruck[3]

[1] Unité MIA INRA 78352 Jouy-en-Josas (david.makowski@jouy.inra.fr)
[2] UMR Agronomie INRA INA-PG 78850 Thiverval-Grignon
[3] CETIOM 78850 Thiverval-Grignon

Résumé

L'objectif de cet article est d'évaluer les performances d'un indicateur développé par les agronomes pour décider de traiter ou non les parcelles de colza infestées par le sclérotinia. Cet indicateur correspond au pourcentage de fleurs malades mesuré avant la date du traitement fongicide. La valeur de ce pourcentage est comparée à un seuil de décision et un traitement fongicide est recommandé seulement si le pourcentage est supérieur au seuil. Une base de données incluant des mesures réalisées sur 420 parcelles de colza a été utilisée pour développer 10 modèles Bayésiens reliant la valeur de l'indicateur au niveau d'attaque de la maladie, à la taille de l'échantillon de fleurs prélevées et à un effet régional. Le modèle ayant la plus faible valeur du DIC a été sélectionné et a été utilisé pour déterminer la sensibilité et la spécificité de l'indicateur, puis pour calculer l'aire sous la courbe ROC en tenant compte d'un effet régional et de la taille de l'échantillon de fleurs. Les résultats montrent que la performance de l'indicateur est légèrement supérieure pour la région Centre et qu'elle dépend fortement de la taille de l'échantillon de fleurs.

Abstract

This paper aims at evaluating an indicator for sclerotinia control in oilseed rape crops. This indicator represents the percentage of diseased flowers measured before chemical treatment. The value of the indicator is compared to a decision threshold and a treatment is recommanded when the indicator value is higher than the threshold. Ten Bayesian regression models were developed from a dataset including measurements of diseased flowers collected in 420 oildseed rape fields in France between 2002 and 2005. The model with the lowest DIC value was used to perform a ROC analysis. Sensibility, specificity, and area under the ROC curves were computed in function of a regional effect and of the size of the flower sample. Results show that the accuracy of the indicator is higher for the region Centre and highly depends on the size of the sample of collected flowers.

Introdution

Le sclérotinia (*Slecrotinia sclerotiorum, Lib., de Bary*) est un champignon qui peut être à l'origine de pertes de rendement dans les cultures de colza (*Brassica napus L.*). Cependant, en France, ces pertes de rendement ne sont significatives que pour environ 20% des parcelles agricoles. Des indicateurs de risque ont été développés par les agronomes pour aider les agriculteurs à identifier les parcelles de colza nécessitant réellement un traitement fongicide. L'utilisation de tels indicateurs pourrait éviter aux agriculteurs d'avoir à appliquer un traitement chimique systématique dommageable pour l'environnement.

Dans cet article, nous nous intéressons à un indicateur qui correpond au pourcentage de fleurs malades mesuré dans une parcelle de colza avant la date du traitement fongicide. La valeur de ce pourcentage est comparée à un seuil de décision et un traitement fongicide est recommandé seulement si le pourcentage est supérieur au seuil (Taverne et al., 2003).

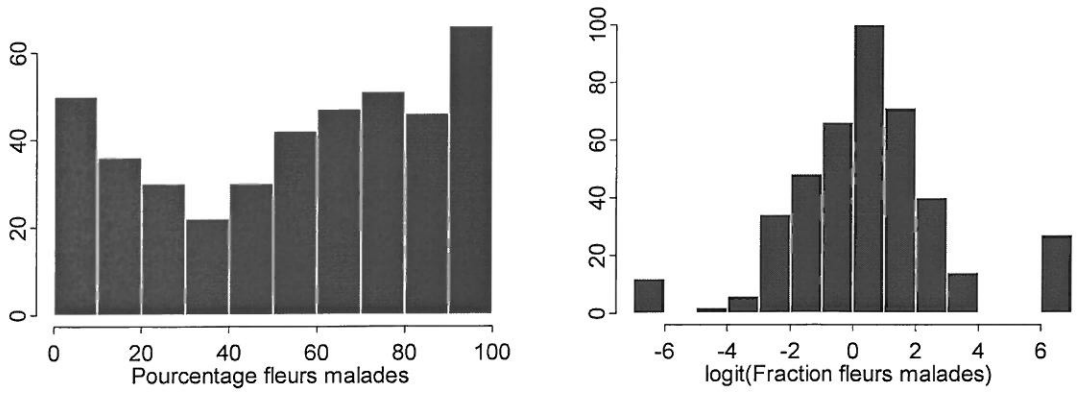
Une étude récente a montré que cet indicateur 'fleurs malades' était plus performant que des indicateurs de type 'grille de risque' utilisant des informations sur les pratiques agricoles, le type de sol, et le climat (Makowski et al., 2005). Cependant, les performances de l'indicateur 'fleurs malades' restent mal connues. Ces performances sont susceptibles de varier en fonction de la taille

de l'échantillon de fleurs prélevées pour estimer le pourcentage de fleurs malades et des caractéristiques de la région dans laquelle l'indicateur est utilisée.
O'Malley et al. (2001) ont proposé une approche Bayésienne pour analyser les performances des indicateurs utilisés pour le diagnostic médical en tenant compte de l'effet de plusieurs covariables. Cette approche est utilisée ici dans un contexte agronomique pour étudier l'effet de la région et de la taille de l'échantillon sur les performances de l'indicateur de risque de sclérotinia.

Données

Le pourcentage de fleurs malades a été mesuré à début floraison sur 420 parcelles de colza en 2002, 2003, 2004 et 2005. Soixante neuf parcelles étaient localisées dans la region Centre où les attaques de sclérotinia sont souvent assez fortes. En 2002 et 2003, le pourcentage de fleurs malades a été mesuré sur chaque parcelle à partir d'un échantillon de 80 fleurs prélevées sur 80 plantes différentes. En 2004 et 2005, la taille de l'échantillon a été réduite à 40 fleurs. Chaque fleur prélevée a été incubée en boite de pétri pendant 4 jours. La présence ou l'absence du champignon a ensuite été déterminée pour chaque boite. La distribution des mesures est présentée sur la figure 1.
Le pourcentage de plantes malades a également été mesuré sur chaque parcelle trois semaines avant la récolte sur un échantillon de 200 plantes. Le pourcentage de plantes malades a été utilisé pour classer les parcelles en deux catégories: faiblement attaquée (pourcentage de plantes malades < 10%) ou fortement attaquée (pourcentage de plantes malades ≥ 10%). Le nombre total de parcelles classées dans la catégorie 'fortement attaquée' est égal à 82 dont 20 dans la région Centre. Dix huit parcelles 'fortement attaquées' correspondaient à des parcelles où le pourcentage de fleurs malades avait été déterminé à partir d'un échantillon de fleurs de taille 80.

Figure 1. Mesures de pourcentage de fleurs malades obtenues sur 420 parcelles de colza (avec et sans transformation logit).



Modèles

L'approche Bayésienne proposée par O'Malley et al. (2001) a été utilisée pour relier le pourcentage de fleurs malades au niveau d'attaque, à la taille de l'échantillon et à l'effet de la région.
Une transformation logit a été d'abord appliquées aux mesures de fraction de fleurs malades notées Y. Les valeurs transformées, $\text{logit}(Y)$, ont été supposées distribuées selon des lois normales indépendantes dont l'espérance μ_{ATR} et la variances σ^2_{ATR} peuvent dépendre du niveau d'attaque (A), de la taille de l'échantillon (T) et de la région (R).
Dix modèles ont été ensuite définies pour relier la variable $\text{logit}(Y)$ aux trois variables (tableau 1). Tous ces modèles sont des modèles linéaires incluant un nombre de variables plus ou moins grand. Par exemple, le modèle 8 est défini par

$$\text{logit}(Y) \sim N(\mu_{ATR}, \sigma^2_{ATR})$$

$$\mu_{ATR} = \alpha_1 + \alpha_2 A + \alpha_3 R + \alpha_4 T + \alpha_5 TA$$

avec A l'effet du niveau d'attaque (A=1 si la parcelle est fortement attaquée, zéro sinon), R l'effet de la région (R=1 si la parcelle est dans la région Centre, zéro sinon), T l'effet de la taille de l'échantillon (T=1 si nombre de fleurs=80, zéro sinon). Dans ce modèle, la variance σ^2_{ATR} prend 8 valeurs différentes en fonction de A, T et R.

Dans tous les modèles, les lois a priori des coefficients de régression (les α s) sont des lois normales indépendantes $N(0,10^6)$ et les lois a priori des variances sont des inverses Gamma indépendantes $IG(0.001,0.001)$.

Les paramètres des modèles ont été estimés avec l'échantilloneur de Gibbs mis en oeuvre par WinBUGS. 20000 itérations ont été réalisées pour chaque modèle. Seules les 10000 dernières ont été retenues pour les analyses.

Les 10 modèles ont été comparés à l'aide du Déviance Information Criterion (DIC). Les valeurs obtenues sont présentées dans le tableau 1. Les résultats montrent que la valeur la plus faible du DIC est obtenue avec le modèle 8. Les paramètres de ce modèle sont présentés dans le tableau 2.

Tableau 1. Variables prises en compte par les 10 modèles et valeurs des DIC.

Modèle	Variables	DIC
1	-	2007,24
2	*A*	1989.00
3	*A, T*	1925.04
4	*A, T, A*T*	1919.27
5	*A, R*	1974.53
6	*A, R, A*R*	1975.88
7	*A, T, R*	1907.2
8	*A, T, R, A*T*	1902.29
9	*A, T, R, A*R*	1908.29
10	*A, T, R, A*T, A*R*	1904.23

Tableau 2. Espérances et écart-types des distributions a posteriori des paramètres du modèle 8.

Paramètre	Espérance	Ecart-type
α_1	0.54	0.15
α_2	0.82	0.4
α_3	1.56	0.34
α_4	-2.62	0.26
α_5	1.94	0.75

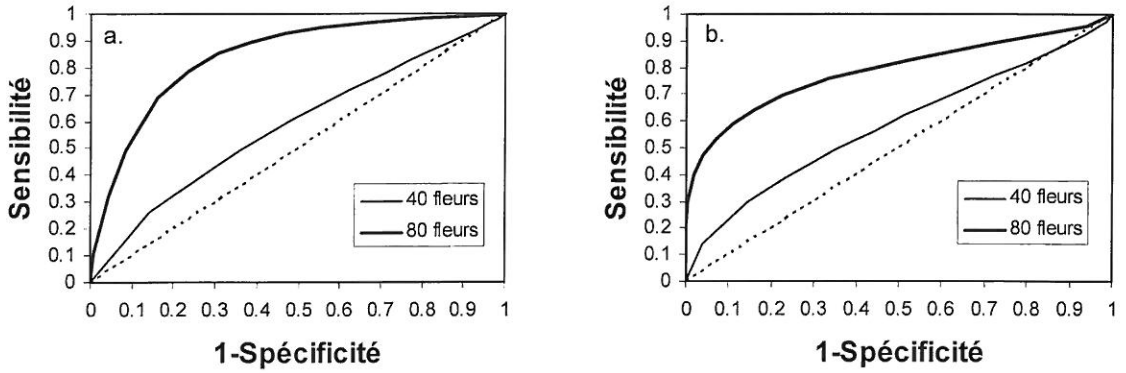
Analyse ROC

Le modèle 8 est utilisé pour réalisé une analyse ROC. Ce type d'analyse est une méthode classique pour évaluer la précision d'indicateur de diagnostic médical (*e.g.* Pepe, 1998). Cet approche est utilisée ici pour évaluer l'indicateur 'fleurs malades' pour différentes valeurs des variables T (taille de l'échantillon) et R (région).

Le principe est de déterminer la sensibilité et la spécificite de l'indicateur pour différents seuils de décision. La sensibilité est définie par $P(Y \geq Seuil | A = 1, T, R)$ et la spécificité est définie par $P(Y < Seuil | A = 0, T, R)$. Ces deux probabilités représentent respectivement la probabilité de

recommander un traitement fongicide lorsque la parcelle est fortement attaquée et la probabilité de ne pas recommander de traitement lorsque la parcelle est faiblement attaquée.
Une courbe ROC est une courbe reliant la sensibilité à 1 - spécificité. Une courbe passant par le point (0, 1) indique que l'indicateur *Y* est performant. Inversement, si la courbe est proche de la bissectrice, l'indicateur est très peu informatif. L'aire sous la courbe ROC (ASC) est un critère souvent utilisé pour analyser la performance de l'indicateur (e.g. Pepe, 1998). Une aire proche de 1 indique un indicateur performant, une aire proche de 0.5 indique un indicateur inutile.
La sensibilité et la spécificité sont estimées ici à l'aide du modèle 8 avec une série de valeurs de *Seuil* comprises entre 0 et 1 pour les deux valeurs de *T* et les deux valeurs de *R*. Au total, quatre courbes ROC sont obtenues en utilisant les 10000 dernières valeurs de paramètres générées par WinBUGS (Figure 2).
Les résultats montrent que la région n'a pas une très forte influence sur la performance de l'indicateur. Les valeurs d'aire sous la courbe ROC (ASC) obtenues pour la région Centre (*R*=1) et les autres régions (*R*=0) sont assez similaires. L'ASC est cependant un peu plus élevée pour la région Centre (ASC=0.85 pour le Centre contre 0.79 pour les autres régions, avec 80 fleurs).
Par contre, les résultats montrent que l'ASC est nettement plus élevée lorsque *T*=1 (nombre de fleurs=80) que lorsque *T*=0 (nombre de fleurs=40). Ainsi, pour la région Centre, ASC=0.85 avec 80 fleurs et ASC=0.59 avec 40 fleurs. Ces résultats indiquent que la taille de l'échantillon peut avoir une influence importante sur la précision de l'indicateur et donc sur la qualité de décisions de traitement. Il est cependant possible que cet effet 'taille de l'échantillon' soit confondu avec un effet année. En effet, les mesures basées sur des échantillons de 80 fleurs ont toutes été réalisées en 2002 et 2003. Par ailleurs, il est important de noter que seulement 18 parcelles appartiennent à la fois à la catégorie 'fortement attaquée' et à la catégorie 'échantillon de 80 fleurs'. Les valeurs de sensibilité sont donc peut-être mal estimées.

Figure 2. Courbes ROC obtenues avec le modèle 8 pour la région Centre (a) et pour les autres (b).



Conclusion

Les résultats montrent que les modèles Bayésien peuvent être utilisés pour analyser les performances d'indicateurs de maladie en fonction de plusieurs covariables. Cette approche a été utilisée ici pour étudier l'effet de la région et l'effet de la taille de l'échantillon sur la performance d'un indicateur utilisé pour décider de traiter contre le sclérotinia du colza. Il serait utile d'étudier l'effet d'autres covariables, par exemple le type de sol ou le climat. Une approche similaire pourrait être utilisée pour évaluer des indicateurs développés par les agronomes pour raisonner le traitement d'autres maladies.

Bibliographie

[1] Makowski, D., Taverne, M., Bolomier, J. et Ducarne, M. (2005). Comparison of risk indicators for sclerotinia control in oilseed rape. *Crop protection*, 24, 527-531.

[2] O'Malley, A.J., Zou, K.H., Fielding, J.R. et Tempany, C.M.C. Bayesian regression methodology for estimating a Receiver Operating Characteristic curve with two radiologic applications. *Acad Radiol*, 8, 713-725.
[3] Pepe, M.S. (1998). Three approaches to regression analysis of Receiver Operating Characteristic curves for continuous test results. *Biometrics*, 54, 124-135.
[4] Taverne, M., Dupeuble, F. et Penaud, A. (2003). Evaluation of a diagnostic test for Sclerotinia on oilseed rape at flowering. *Proceedings of the 11th International Rapeseed Congress*, Copenhagen.