8

Are Bayesian approaches useful in plant pathology?

Jonathan Yuen[#] *and Asimina Mila*[##]

Abstract

Bayesian methods are seldom seen in the context of plant pathology. However, they offer a number of possibilities in data analysis and decision theory. Most decision makers utilize a Bayes-like methodology to combine prior information with new information, and in this context the acceptance or failure of predictive systems can be itself dependent on prior information. In the context of complex systems, a Bayesian approach to data analysis using MCMC methods offers flexibility beyond that encountered in a frequentist approach. Frequency distributions of parameters become available, and the effect of the certainty of the prior distribution can also be determined.

Keywords: plant pathology, sensitivity, specificity, Gibbs sampling

Introduction

Plant pathologists, like many biologists, use statistics to help them draw conclusions from collected data. In addition, plant pathology, like many agricultural sciences, derives its statistical methods from agronomical practices. Thus, early on, plant pathologists came to rely on methods such as ANOVA, often adapting field designs, whether they were suitable or not. These methods may be sufficient when one is evaluating large effects such as those from pesticide application or use of host-plant resistance, but this approach may not be optimal if effects are smaller.

Much of the earlier work in plant pathology was descriptive, or testing of different control measures. Today’s control strategies require much more precision. In addition, much of the practical side of plant pathology involves making predictions about the future. Whether pests will occur, when, to what extent and so on, are much more modern components of plant pathology. Environmental considerations generally rule out routine application of pesticides in today’s agriculture. Most systems that can make predictions about the future are imperfect, and methods that can accommodate this uncertainty are necessary. Biological systems are often complex, and the relationships between explanatory variables and the resulting outcomes are also complex. Bayesian methods have been proven appropriate in handling such issues and could be used in plant pathology, but are not used as much as they could be. This is in part historical, but also in part because much could be done without them.

[#] Department of Ecology and Crop Production Sciences, Swedish University of Agricultural Research, SE 750 07 Uppsala, Sweden. E-mail: Jonathan.Yuen@evp.slu.se

[##] Department of Plant Pathology, University of California-Davis, Kearney Agricultural Center, Parlier, CA 93648, USA. E-mail: mmila@uckac.edu

In this paper, we discuss applications in plant pathology or plant protection that rely on Bayesian methods. The first is relatively trivial, but essential for understanding how predictive systems interact with the primary producers that make decisions (farmers). In the second, the modelled biological system increases in complexity, with a necessity for using MCMC Bayesian methods to find solutions.

A simple example

A predictive system published in the literature does not necessarily guarantee that it is good, correct, or of any use to the decision-maker. One needs to examine the correct predictions that it makes, as well as the incorrect predictions.

Farmers often have to make yes-no decisions (Bernoulli variables) that entail the use of pesticides or not. In this context, errors and correct decisions can be summarized in a 2 x 2 table (Table 1), using methods borrowed from (human) clinical epidemiology (Yuen, Twengstrom and Sigvald 1996). The proportion of correct predictions when the pest is actually present is the true positive rate or sensitivity and is equal to $\frac{A}{A+B}$. Likewise, the proportion of correct decisions when the pest is absent is called the specificity, $\frac{D}{C+D}$, and one minus the specificity, $\frac{C}{C+D}$, is the false-positive rate.

Table 1. Definition of true- and false-positive rates based on recommendations and actual outcomes

	Spray	Don’t spray	
Disease present	A	B	$\frac{A}{A+B}$ True positive, sensitivity
Disease absent	C	D	$\frac{C}{C+D}$ False positive
			$\frac{D}{C+D}$ Specificity

A numerical example was presented by Yuen and Hughes (2002) (based on data from Jones (1994)) and reproduced in Table 2. In this example, the sensitivity of the predictor proposed by Jones (1994) is $\frac{28}{41}$, the specificity is $\frac{7}{17}$, and the false positive rate is $\frac{10}{17}$.

Table 2. Eyespot predictor presented by Yuen and Hughes (2002)

		Predictor		
		Apply treatment	Withhold treatment	Total
True	Treatment justified	28	13	41
Status	Treatment not justified	10	7	17

If a predictive scheme has a continuous or almost continuous variable as an output, varying the cut-off level where the control measures are to be applied, referred to here as the decision threshold (Hughes, McRoberts and Burnett 1999) will affect both sensitivity and specificity. Such a system predicted the occurrence of economically damaging levels of *Sclerotinia* stem rot in oilseed rape (Yuen, Twengstrom and Sigvald 1996).

From this analysis, factors that led to increased risk gave increasing numbers of points. Thus, in this system a lower decision threshold will raise the sensitivity of the predictor but also decrease the specificity. A decision threshold of zero, for example, will lead to prediction of pests in all fields. Such a predictor has a high sensitivity (i.e. all fields that need a control measure receive one) but a poor specificity.

If we continue to use this particular example, we see that the opposite situation occurs with an extremely high decision threshold. If this is very high then the specificity is good (we wouldn't spray those fields that don't need it) but sensitivity is poor (we miss fields that needed spraying).

By varying the decision threshold and plotting the true-positive rate as a function of the false-positive rate, one can produce a receiver-operating characteristic (ROC) curve (Metz 1978). An ROC curve from a recalibrated predictor for *Sclerotinia* stem rot is presented in Figure 1. Extremely low decision thresholds generate the points in the upper right-hand corner, and the extremely high decision thresholds generate the points in the lower left-hand corner. The best ROC curves appear pushed to the upper left-hand corner, and ideally would have a high true-positive rate (high sensitivity) but a low false-positive rate (high specificity).

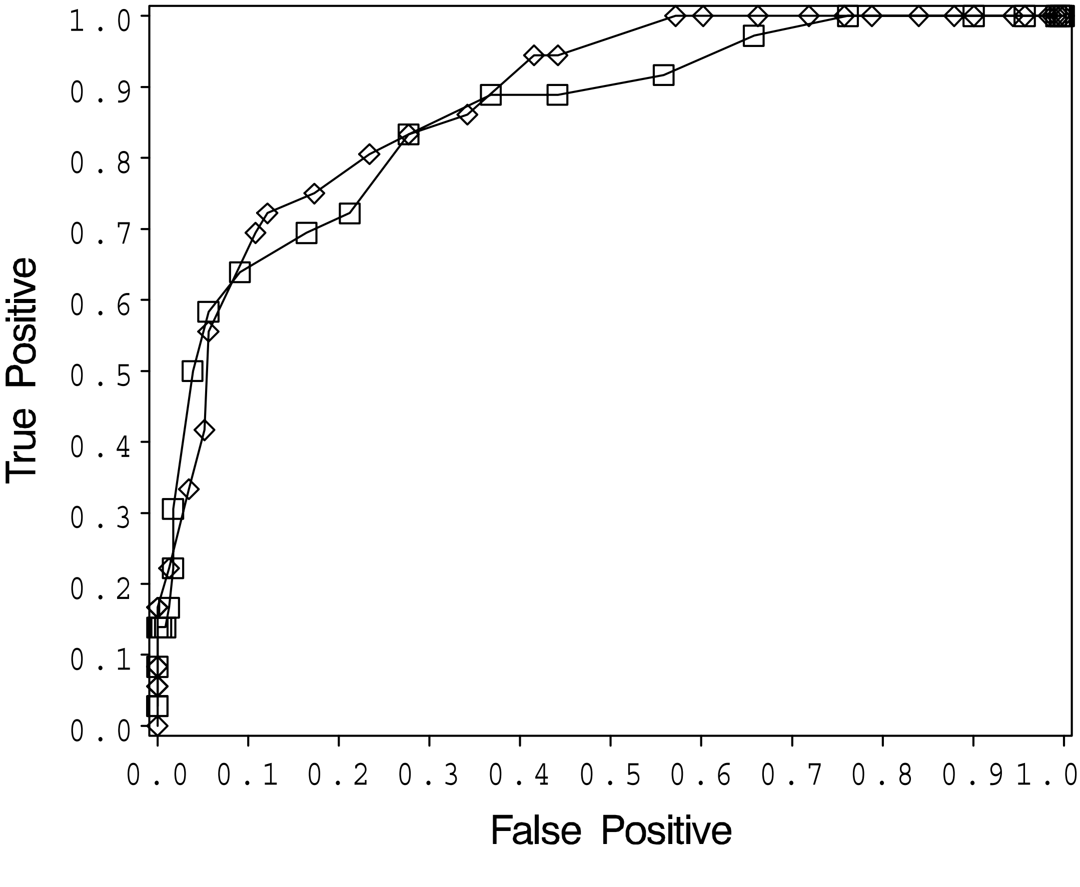

Figure 1. Receiver-operating characteristic curves from the original disease-forecast algorithm for *Sclerotinia* stem rot (—□—) and the recalibrated algorithm after logistic regression (—◊—), from Yuen, Twengstrom and Sigvald (1996)

We can use Bayes's theorem in a classical manner to combine information on prior probabilities with the sensitivity and specificity of a predictor to calculate posterior probabilities. We have seen that a generalized form of Bayes’s theorem can be written as

$$f() \alpha g() L(y) \quad (1)$$

where *f*() represents the posterior distribution of the quantity of interest, *g*() represents the prior distribution of the quantity of interest, and L(*y*) represents the likelihood function, based on the available data.

Medical epidemiologists often use what they call the likelihood ratio to characterize diagnostic tests (Sackett, Haynes and Tugwell 1985). The likelihood ratio for a positive test $(LR_{positive})$ is the probability that a positive test will be observed in an affected person, compared to the probability that the same would be observed in a non-affected person (Knottnerus, Van Weel and Muris 2002). After converting this to a plant-protection scenario and some substitution in Table 1, we recognize the numerator of this ratio is the sensitivity, whereas the denominator is merely 1 minus the specificity.

Thus, $(LR_{positive})$ can be written as

$$LR_{positive} = \frac{sensitivity}{1 - specificity} \quad (2)$$

Likewise, the likelihood ratio for a negative prediction $(LR_{negative})$ is the ratio of the probability of a negative test in an affected person, compared to the probability of a negative test in an unaffected person. This can also be written as

$$LR_{negative} = \frac{1 - sensitivity}{specificity} \quad (3)$$

Using these values for the likelihood ratios and the odds of event occurrence where

$$\text{odds}(event) = \frac{\Pr(event)}{1 - \Pr(event)} \quad (4)$$

we can write Bayes’s theorem as

$$odds_{posterior} = odds_{prior} \times LR \quad (5)$$

where LR represents either $LR_{positive}$ or $LR_{negative}$.

Yuen and Hughes (2002) presented an example for *Sclerotinia* stem rot in oilseed rape where the likelihood ratios for positive and negative predictions were derived with varying sensitivities and specificities. These could then be coupled with varying prior probabilities via Bayes’s theorem to see the effect of the predictors.

For pests that are neither common nor rare, even predictive systems with moderate performance (sensitivity or specificity) could change the probabilities to the extent that the behaviour of the decision-maker might be affected. For rare or common pests, however, a likelihood ratio needs to be extremely large or small in order to affect substantially the prior probabilities. Yuen and Hughes (2002) pointed out that even a predictor with specificity and sensitivity of 0.9 would have moderate performance in such a situation. Using these values, an $LR_{positive}$ would change a prior probability of 0.1 to 0.5, whereas an $LR_{negative}$ would change a prior probability of 0.9 to 0.5. Thus, it is not clear that such a predictor would be helpful for common or rare pests.

It is helpful to compare these simple exercises with the thought processes that a decision-maker (possibly a farmer or farm manager in this case) has to go through. In most cases, there is some prior information about whether the pest will occur. This might be entirely historical, or could also be based on the experiences of nearby farms or other production areas. This information is then modified by the information in the predictive system. While decision-makers may not consciously use Bayes's theorem, and are certainly not able to do so unless we supply information on sensitivity and specificity, they certainly use a similar process to combine prior information and predictive systems.

Our Bayesian analysis allows us to see how this information can be combined. If the prior probabilities indicate *pest will occur* or *pest is very rare*, then the predictive system must have extremely high $LR_{positive}$ or low $LR_{negative}$ in order to affect the behaviour of the decision maker.

A more complicated example

In the previous example, the predictor itself was determined with logistic regression. This in turn allowed determination of the sensitivity and specificity of the predictor and subsequently the likelihood ratio to use in calculating the posterior probabilities. An alternate method to derive the relationship between crop management practices, environmental variables and the occurrence of disease is to use a Bayesian methodology. This was done in a study of *Sclerotinia* stem rot in soybean by Mila, Yang and Carriquiry (2003). Their use of a Bayesian approach allows more flexibility and additional analyses.

In their study, the relationship between prevalence (presence or absence) of *Sclerotinia* stem rot in soybean and explanatory variables such as air temperature and precipitation in July and August, tillage and state effects was examined in a data set resulting from a survey from 1995 to 1998 in Illinois, Iowa, Minnesota and Ohio, 4 states of the North-Central region of the USA. In total 1853 soybean fields were randomly sampled during the 4-year period. From each field 20 soybean stems were sampled in a zig-zag pattern and shipped to Iowa State University for examination. Information about the tillage system implemented in each field was obtained from farmers during interviews with enumerators of the National Agricultural Survey Services (NASS). Tillage systems are generally classified in three categories according to the amount of plant residues remained on the soil surface. Weather data for each field were obtained from the nearest National Oceanic and Atmospheric Administration (NOAA) weather station. Conventional logistic regression was first used to examine the relationship between the explanatory variables and *Sclerotinia* stem-rot prevalence (Mila et al. 2003).

As a second step, a Bayesian logistic-regression model was fit to the data set. For this purpose, a 3-stage hierarchical model was fitted to the data using version 1.3 of the programme BUGS (Bayesian Inference Using Gibbs Sampling) (Spiegelhalter, Thomas and Best 1999). This programme is available at *http://www.mrc-bsu.ac.uk/bugs*. Parameter estimates were produced with an MCMC algorithm, Gibbs sampling. A graphical representation of the model appears in Figure 2.

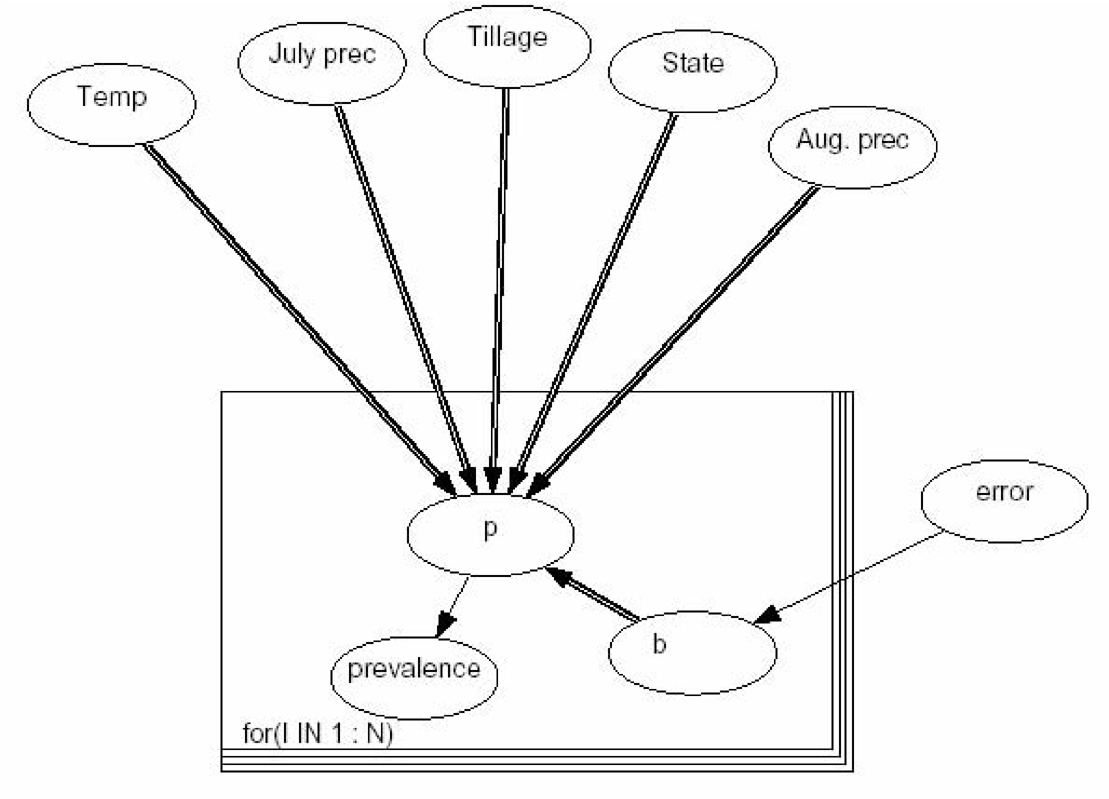

Figure 2. Graphical model for prediction of *Sclerotinia* stem rot in soybean

In the first part of the Bayesian analysis, it was assumed that little was known about the parameter values and thus relatively non-informative priors *g*() were used in equation 1. In the second part of the analysis the relatively non-informative priors were replaced with informative priors to investigate if the information in the data set was sufficient to produce reliable estimates (Table 3). For that reason the following assumptions on the parameter values were made. The parameter for average air temperature of July and August should be negative (since *Sclerotinia* stem rot is a

Table 3. Means and variances of prior distributions used in a Bayesian analysis of soybean *Sclerotinia* stem rot (Mila, Yang and Carriquiry 2003). See text for additional details.

	Prior distributions[a]	
Parameter	Relatively non-informative	Informative
Intercept	0, 1000000	0, 1000000
Meant T July and August	0, 1000000	0[b], 1000000
July precipitation	0, 1000000	0.4[c], 1000000
(No tillage)* Mean T	0, 1000000	-0.5[b], 1000000
(Minimum till)* Meant T	0, 1000000	2[c], 1000000
Iowa (IA)	0, 1000000	2[c], 1000000
Minnesota (MN)	0, 1000000	3[c], 1000000
Ohio (OH)	0, 1000000	0.1, 1000000
August precipitation	0, 1000000	0.25[c], 1000000

[a] Values consist of the mean and variance of a normal distribution.
[b] Distribution only takes negative values.
[c] Distribution only takes positive values.

cool-temperature disease). Parameters for July and August precipitation should be positive (prolonged periods of moisture are favourable for the disease development). Specifically, the July precipitation effect was assumed to be equal to the temperature effect and the August precipitation effect half of the July precipitation effect. Since no previous survey on *Sclerotinia* stem rot in the North-Central region is available we used Empirical Bayes to form informative prior distributions of the state effect. Finally, the effect of using informative priors was investigated to see if the information in the data set was sufficient to produce reliable estimates.

Conventional logistic regression and a Bayesian analysis with relatively non-informative priors yielded similar estimates (Table 4). In addition, the Bayesian analysis allowed calculation of a distribution of the possible values for the different parameters.

Table 4. Point estimates of logistic-regression analysis and posterior-distribution summaries of the parameters used to explain soybean *Sclerotinia* stem rot (Mila, Yang and Carriquiry 2003)

Parameter	Conventional	Bayesian Mean	Bayesian SD
Intercept	5.71	5.58	0.59
Meant T July and August	-0.4	-0.47	0.103
July precipitation	0.029	0.039	0.021
(No tillage)* Mean T	-0.011	-0.0104	0.022
(Minimum till)* Meant T	0.01	0.0137	0.0123
Iowa (IA)	0.52	0.69	0.531
Minnesota (MN)	0.98	1.181	0.532
Ohio (OH)	-0.82	-1.09	0.945
August precipitation	-	0.0104	0.025

Use of informative prior distributions had an effect on the posterior distributions of intercept, July and August precipitation, and the state variables of Iowa (IA) and Ohio (OH) (Table 5). The most remarkable changes occurred at August precipitation and OH parameter estimates. With an informative prior, the posterior distribution of the August precipitation parameter was confined in the positive space and the posterior mean shifted from 0.0104 (with relatively non-informative prior distribution) to 0.1138 (with an informative prior distribution). Similarly, the posterior distribution for the OH parameter significantly changed shape with a shift of the posterior mean from negative (with a relatively non-informative prior) to positive (with an informative prior) and large reduction of the corresponding standard error, indicating that an

Table 5. Posterior-distribution summaries of the parameters used to explain soybean *Sclerotinia* stem rotusing very informative priors (Mila, Yang and Carriquiry 2003)

Parameter	Bayesian Mean	Bayesian SD
Intercept	2.86	0.528
Meant T July and August	-0.432	0.111
July precipitation	0.1685	0.0173
(No tillage)* Mean T	-0.0134	0.0136
(Minimum till)* Meant T	0.0052	0.0136
Iowa (IA)	0.901	0.289
Minnesota (MN)	1.353	0.43
Ohio (OH)	0.044	0.0357
August precipitation	0.1138	0.021

informative prior may have an increased influence on the posterior distribution for the OH parameter. Thus, analysis based on the sample data alone may not be sufficiently informative on the effect of these parameters on *Sclerotinia* stem rot prevalence in the North-Central region of US.

The study of Mila, Yang and Carriquiry (2003) allowed exploring the effect of the prior probabilities in the analysis. If priors are uncertain or relatively non-informative, then the results of the Bayesian analysis are similar to those of a classical analysis. With a Bayesian analysis, however, they could also examine the amount of information that was available in the data set. If the parameter estimates are overly sensitive to the nature of the priors, then this may be due to lack of sufficient information in the data set.

Conclusions

Plant protection and plant pathology could benefit from increased awareness of Bayesian methods. While decision-makers may not actively use Bayes's theorem in making practical plant-protection decisions, they use methods that are quite similar in order to combine new information with previous knowledge.

Even as scientists we have a need to combine new knowledge with what we already know. Bayes's theorem is one way in which this can be done, although the subjectivity that surrounds the choice of prior distributions is, unfortunately, unescapable.

Bayesian approaches in more complex modelling situations offer benefits compared to traditional statistics. If little prior information is available, the results are quite similar to those of traditional statistics. Additional benefits include the ability to examine the distributions of parameters and being able to test the effects of changing the priors themselves. This in turn can lead to a reinterpretation of the results of a study. A traditional statistical approach with a given data set may show non-significant effects of a variable, and in our pursuit of simplicity we usually state that it has no effect. A Bayesian approach would, on the other hand, indicate that the data could not change our prior beliefs, thus leaving us informed (or misinformed) as we were before the study.

Although not specifically discussed in this paper, Bayesian approaches can also be used in the context of complex decision-support systems, especially where decisions are being made under uncertainty. Such decision-support systems, in an agricultural production context, can include plant-disease management.

An excellent example of this type of approach, where the decisions for management of mildew in winter wheat were dealt with using both dynamic-simulation models and Bayesian networks, is the thesis of Jensen (1995).

References

Hughes, G., McRoberts, N. and Burnett, F.J., 1999. Decision-making and diagnosis in disease management. *Plant Pathology*, 48 (2), 147-153.

Jensen, A.L., 1995. *A probabilistic model based decision support system for mildew management in winther wheat*. Ph.D. Thesis, Aalborg University, Dina Report 39.

Jones, D.R., 1994. Evaluation of fungicides for control of eyespot disease and yield loss relationships in winter wheat. *Plant Pathology*, 43 (5), 831-846.

Knottnerus, J.A., Van Weel, C. and Muris, J.W.M., 2002. Evidence base of clinical diagnosis: evaluation of diagnostic procedures. *British Medical Journal,* 324 (7335), 477-480.

Metz, C.E., 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine,* 8 (4), 283-298.

Mila, A.L., Carriquiry, A.L., Zhao, J., et al., 2003. Impact of management practices on prevalence of soybean *Sclerotinia* stem rot in the north-central United States and on farmers' decisions under uncertainty. *Plant Disease,* 87 (9), 1048-1058.

Mila, A.L., Yang, X.B. and Carriquiry, A.L., 2003. Bayesian logistic regression of soybean *Sclerotinia* stem rot prevalence in the US north-central region: accounting for uncertainty in parameter estimation. *Phytopathology,* 93 (6), 758-764.

Sackett, D.L., Haynes, R.B. and Tugwell, P., 1985. *Clinical epidemiology: a basic science for clinical medicine*. Little, Brown and Company, Boston.

Spiegelhalter, D.J., Thomas, A. and Best, N.G., 1999. *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.

Yuen, J. and Hughes, G., 2002. Bayesian analysis of plant disease prediction. *Plant Pathology,* 51 (4), 407-412.

Yuen, J., Twengstrom, E. and Sigvald, R., 1996. Calibration and verification of risk algorithms using logistic regression. *European Journal of Plant Pathology,* 102 (9), 847-854.