A Bayesian Network Model for Yellow Rust Forecasting in Winter Wheat

Xiaodong Yang[1,2(✉)], Chenwei Nie[3], Jingcheng Zhang[4], Haikuan Feng[1,2], and Guijun Yang[1,2]

[1] Key Laboratory of Quantitative Remote Sensing in Agriculture of Ministry of Agriculture P. R. China, Beijing Research Center for Information Technology in Agriculture, Beijing, China
{yangxd, fenghk, yanggj}@nercita.org.cn
[2] National Engineering Research Center for Information Technology in Agriculture, Beijing, China
[3] Institute of Remote Sensing and Digital Earth of Chinese Academy of Sciences, Beijing, China
ll57606629@qq.com
[4] College of Life Information and Instrument Engineering, Hangzhou Dianzi University, Hangzhou, China
zhangjc_rs@163.com

Abstract. Yellow rust (YR) is one of the most destructive diseases of wheat. We introduced the Bayesian network analysis as a core method and develop a large-scale YR forecasting model based on several important meteorological variables that associate with disease occurrence. To guarantee an effective model calibration and validation, we used multiple years (2010–2012) of meteorological data and the ground survey data in Gansu Province where the YR intimidated most severely in China. The validation results showed that the disease forecasting model is able to produce a reasonable risk map to indicate the disease pressure across the region. In addition, the temporal dispersal of YR can also be delineated by the model. Through a comparison with some classic methods, the Bayesian network outperformed BP neutral network and FLDA in accuracy, which thereby suggested a great potential of Bayesian network in disease forecasting at a regional scale.

Keywords: Yellow rust · Meteorological factor · Bayesian network Forecasting model

1 Introduction

Yellow rust (YR) is one of the most important epidemic diseases of wheat. It can cause a significant loss of wheat at a global scale [5, 9]. In the year 2002, over 6.7 million hm2 wheat was infected by YR in China, which resulted in a production loss around 10 billion kg [4]. It is of great importance to predict the YR effectively at an early stage, since it can provide critical information to agriculture plant protection departments to facilitate timely spray recommendation. So far, a series of studies had been conducted to forecast YR over a long time based on meteorological and agronomy data around the

Published by Springer Nature Switzerland AG 2019
D. Li and C. Zhao (Eds.): CCTA 2017, IFIP AICT 545, pp. 65–75, 2019.
https://doi.org/10.1007/978-3-030-06137-1_7

world. Hu et al. modeled a BP neutral network to predict YR in Hanzhong city, Shaanxi Province. The forecast results were highly consistent with the actual situation [3]. Chen et al. predicted YR severities at a seasonal time step in both Maerkang county and Tianshui city using discriminant analysis, with rewind accuracy and cross-validation accuracy greater than 78% [6]. Coakley et al. developed an improved method to predict YR [27]. Wang et al. (2012) conducted a study to develop a stable neutral network for predicting YR [31].

To date, it should be noted that there were few attempts made in forecasting YR at a regional scale with a short time step (7 days). Instead, efforts were made on forecasting seasonal severities of YR using spores counts data and meteorological observations. These models can achieve high accuracy at a local site. However, in most regions where studied, Puccinia striiformis can survive through winter. It is difficult to apply these models in the region where the spores counts data are not available. Considering that the YR is a multi-cycle disease, which distributed over large areas in the world. It is necessary to develop a multi-temporal YR forecasting model at a large spatial scale. However, such forecasting models lack recently.

Several critical weather factors associating the occurrence of YR on winter wheat were reported, which were temperature (T), humidity (H), precipitation (P), sunshine (S) [30]. It is important to relate YR occurrence with meteorological factors building the developing YR forecasting model. Bayesian network is a probabilistic graphical model that based on probability and statistics theory. The characteristics of the Bayesian network include rigorous reasoning process, clear semantic expression, data learning ability, etc. It is an efficient method for uncertainty reasoning and data analysis [15]. And it has been widely used in many fields since the 1980s. In this study, the Gansu province, which is a typical wheat planting region that suffers YR in China, was selected as our study area. Based on a continuous YR field survey data and corresponding meteorological data from 2010 to 2012, the potential of Bayesian network in disease forecasting was examined. In addition, a forecasting model of YR was developed to facilitate disease management at a regional scale.

2 Materials and Methods

2.1 Yellow Rust Survey Data

The YR survey data is collected by Gansu Provincial Protection Station. During 2010 to 2012, a weekly field survey was conducted across southern area of Gansu province (Fig. 1). The climate of the study region is characterized by high humidity and rainfall, and YR disease occurs almost every year. The surveyed data include the initial date of disease occurrence and the infected area. A total of 45, 18, 47 sites were surveyed in 2010, 2011 and 2012, respectively. The distribution of survey points is demonstrated in Fig. 1. The investigation ranged from the beginning of March to the end of July in each year. For model calibration and validation, the surveyed data were randomly split into 60% versus 40% in each year.

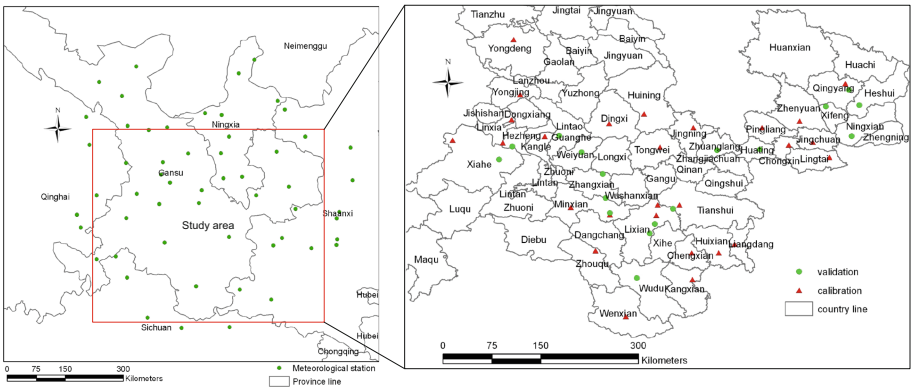


Fig. 1. Distribution of YR survey sites and meteorological stations in the study area

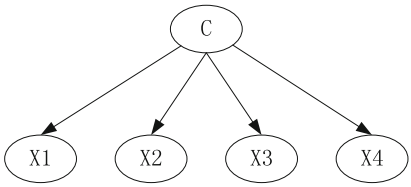
2.2 Meteorological Data

In this study, according to the research results of Cooke [9], four meteorological factors were chosen as input variables, including average temperature, average humidity, precipitation and sunshine duration. The daily data of these meteorological factors from a total of 54 weather stations around the study area was acquired from Chinese Meteorological Data Sharing Service System. The time range of the data is from a week before YR occurrence (based on the investigation data) in spring to its mature stage in each year. There are 3 steps to process meteorological data, including removal of abnormal value, averaging of meteorological factors on a weekly basis, and interpolation of each factor to a resolution of 30 m*30 m. Considering some meteorological data have a strong relationship with altitude, the DEM (Digital Elevation Model) data was used the adjust the spatial maps of meteorological factors by interpolating the fitted residue across the region [11, 14]. As for interpolation methods, the normality of the distribution of each meteorological factor was examined by Kolmogorov-Smirnov method. For those meteorological factors have a P-value > 0.05, a kriging method is used to conduct interpolation. Otherwise, an inverse distance weighted method is adopted.

2.3 Yellow Rust Forecast Based on Bayesian Network

2.3.1 The Bayesian Network Theory

Suppose there is a finite set X = {X1, X2, …, Xn} of discrete random variables, and each variable Xi can take on values from a finite set, denoted by Val(Xi). We use capital letter X to denote set of variables Xi, and lower-case letter x to denote specific values taken by those variables. A Bayesian network for X, the Bayesian network is B = <G, Θ>. The first component, G, is a directed acyclic graph whose vertices correspond to the random variables X1, X2, …, Xn, and whose edges represent direct dependencies between the variables.



Note: C: indicate class node, X1-X4: indicate attribute nodes

Fig. 2. An example of a Bayesian network

As an example, let X1 = {X1, X2, …, Xn, C}, where variables X1, X2, …, Xn are the attributes and C is the class variable. The graph structure of this example is demonstrated in Fig. 2. given a variable set D = {x1, x2,.., xn}, and a class variable set c, according to Bayesian theory, the posterior probability of the most likely class can be estimated by [28]:

$$p(c|D) = \arg\max_{c \in C} \frac{p(D|c)p(c)}{p(D)} \tag{1}$$

where the p(D) is independent constant, the formula (1) can be written as:

$$p(c|D) = \arg\max_{c \in C} p(D|c)p(c) \tag{2}$$

Based on the rules of multiplication, p(D|c) can be expressed formulas:

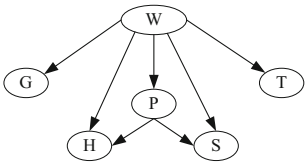
$$\begin{aligned} p(D|c) &= p(x_1|c)P(x_2|x_1, c)p(x_n|x_1, x_2, \cdots, x_{n-1}, c) \\ &= \prod_{i=1}^{n} p(x_i|x_1, x_2, \cdots, x_{i-1}, c) \end{aligned} \tag{3}$$

For each xi, if there is a set π(xi) ∈ {x1, …, xi − 1}, xi and {x1,…, xi − 1} are conditional independence given the set π(xi). Then formula (2) has the form as formula (4), and this is the classification formula of Bayesian network.

$$c(x) = \arg\max_{c \in C} p(c) \prod_{i=1}^{n} p(x_i|\pi(x_i), c) \tag{4}$$

2.3.2 Development of Bayesian Network

In this study, a Bayesian network model is developed to forecast YR with not only the four meteorological factors as mentioned above, but also the growth period, given the growth period has a significant impact on disease occurrence probability. In addition, considering the physical relationships between precipitation and humidity, and between precipitation and sunshine duration, the structure of the Bayesian network is illustrated in Fig. 3.



Note: W: represents the disease status of wheat; G: represents growth period of wheat; H: represents the average humidity; P: represents the precipitation; S: represents the sunshine duration; T: represents the average temperature

Fig. 3. Bayesian network structure

In this Bayesian network, W represents the status of YR occurrence, which is a binary variable (w1 = health, w0 = YR infected; G is the growth stage (1 = reviving stage, 2 = jointing stage, 3 = heading stage, 4 = milk stage,). While T, P, H, S denote average temperature, precipitation, average humidity, sunshine duration respectively. Each of them has 6 degrees following de Vallavieille-Pope, Cooke [9, 18], etc. The value range of each degree for all weather factors is given in Table 1.

Table 1. Grade of meteorological factor

	Min	Max	Grade of factor					
			1	2	3	4	5	6
P(mm)	0	7.52	[0, 0.1]	(0.1, 1]	(1, 2]	(2, 3.5]	(3.5, 4.5]	(4.5, 7.52]
H(%)	24.88	91.97	[24.88, 35]	(35,45]	(45,50]	(50, 60]	(60, 80]	(80, 91.97]
T(°)	−8.5	24.73	[−8.5, 0]	(0, 5]	(5, 10]	(10, 15]	(15, 20]	(20, 24.73]
S(h)	0.37	12.9	[0.37, 3]	(3, 5]	(5, 6.5]	(6.5, 8]	(8, 9.5]	(9.5, 12.9]

As the YR field surveys were conducted on a weekly basis, the meteorological data was also processed per week. Considering the possible latent effect, the independent variables were prepared to start from one week in advance to the initial YR field survey date. The conditional probability was calculated with Laplace estimate method to avoid possible zero occurrence frequency. The equations are shown in (5)–(7) [16].

$$p(w) = \frac{\sum_{i=1}^{n} \delta(w_i, w) + 1}{n + n_w} \tag{5}$$

$$p(a_j|w, b) = \frac{\sum_{i=1}^{n} \delta(a_{ij}, a_j)\delta(w_i, w)\delta(b_i, b) + 1}{\sum_{i=1}^{n} \delta(w_i, w)\delta(b_i, b) + n_j} \tag{6}$$

$$p(a_j|w) = \frac{\sum_{i=1}^{n} \delta(a_{ij}, a_j)\delta(w_i, w) + 1}{\sum_{i=1}^{n} \delta(w_i, w) + n_j} \tag{7}$$

Where n is the number of samples, nw is the number of classes, nj is the number of the jth variable's values, wi is the actual class value of the ith sample, aj is the jth value of the independent variables, aij is the jth value of the independent variables in the ith sample. δ(wi, w) is a two-valued function, the value of the function is 1 when wi = w, or else, the value is 0.

The posterior probability of YR occurrence is expressed as:

$$w(x) = \underset{w \in (w_1, w_0)}{\arg\max}\, p(w) \prod_{i=1}^{5} p(x_i|\pi(x_i), w) \tag{8}$$

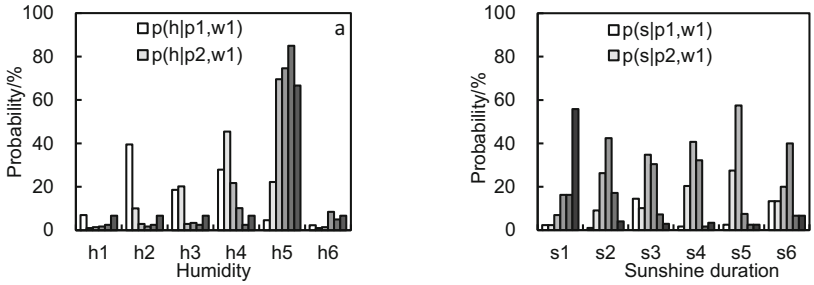
2.3.3 Evaluation of Disease Forecast Model

Based on the posterior probability that is generated from the forecasting model, a threshold is applied to convert the forecasting probability to disease occurrence status. A sample will be marked as health when the probability value is smaller than the threshold. Otherwise, it will be classified as a YR infected sample. To obtain an optimal threshold, we calculated the model accuracy under different thresholds varying from 0 to 1 with a step of 0.05. The optimal threshold can be determined when the highest model accuracy researched. To further compare the Bayesian network to other classic methods, we also compared its performance with that under BP neutral network and FLDA.

3 Results and Discussion

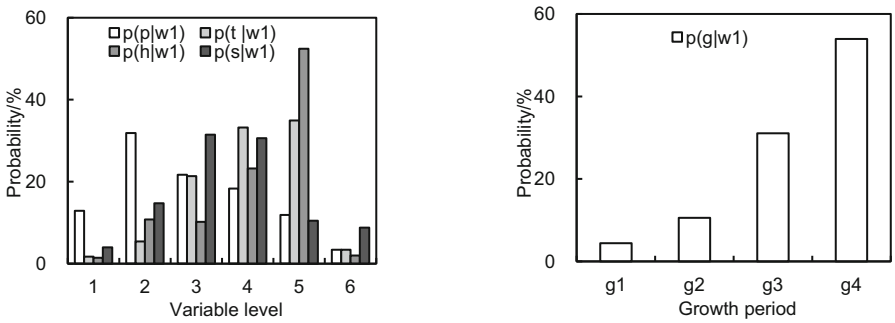
In the bayesian network, the distribution of conditional probability for each node was calculated through formulas (5)–(7) (Figs. 4 and 5). In Fig. 4, for infected sites, with an increase of precipitation, the conditional probability of humidity during in h4 and h5 have a certain increase. While the conditional probability distribution of sunshine duration is relatively uniform. In Fig. 5a, for infected survey sites, the conditional probability variation trends of T, H, P, S are similar to each other, which approaching the Gaussian distribution. In Fig. 5b, the conditional probability of growth stage rise as time goes on. This result is in agreement with the research results of Cooke [9].

In this paper, we developed a YR forecasting bayesian network with four weather factors and one phonological variable, to model the probability of YR infection a week in advance. The output of the Bayesian network model is a posterior probability. The forecasted probability of YR occurrence is compared with the number of actual infected sites according to the survey data (Fig. 6). It is noted that both the number of actual infected sites and the forecasted YR probability showed an increasing trend over time (from reviving stage to milk stage). Figure 7 demonstrated the spatial distribution



Note: h indicate humidity, p indicate precipitation, w indicate whether the wheat infected with YR(w1 indicate infected, and w0 indicate taintless), s indicate sunshine duration, p(h|p1,w1) indicate the probability under the situation that rainfall for p1, and wheat infected with YR.

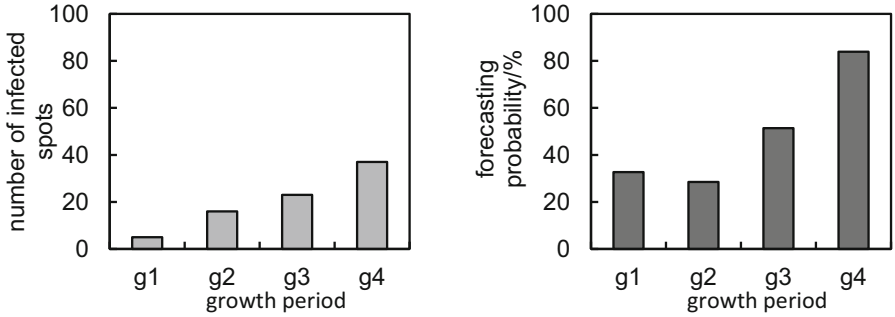
Fig. 4. Conditional probability distribution of the attribute nodes that with more than one parent nodes. a. The conditional probability of the humidity in the case of different rainfall and YR happened, b. The conditional probability of the sunshine duration in the case of different rainfall and YR happened.



Note: Attribute variables in the figure include precipitation, temperature, humidity, sunshine duration, the level of attribute variables showed in table 2; g1, g2, g3, g4 in figure b indicate reviving stage, joining the stage, heading stage, milk stage respectively

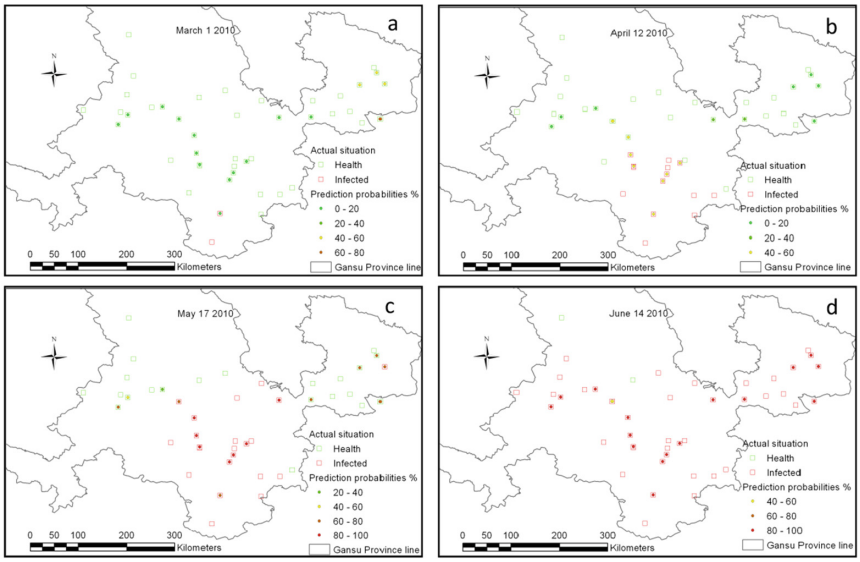
Fig. 5. The conditional probability distribution of the nodes that with single parent. a. The conditional probability of meteorological factor, b. The conditional probability of growth period

of both the forecasting results and the ground truth. The YR started to show up in the southeast of study area at an early stage (reviving stage). Then, another YR occurrence was spotted in the central region of study area in early April. After a spread process, in the middle of June, most surveyed sites were identified as infected over the study area. Such a spatial trend can be well modelled with the developed Bayesian network (Fig. 7).



Note: g1, g2, g3, g4 indicate reviving stage, jointing stage, heading stage, milk stage respectively

Fig. 6. Prediction probability and infected spots: a. Trend of prediction probability b. Trend of actually happened number of infected spots



Note: ● indicating forecasting probability of YR sample points; □indicating the actual happened probability of YR sample points

Fig. 7. Forecasting of YR and physical truth distribution from 2010 to 2012. a. The distribution of prediction probability on March 1, 2010. b. The distribution of prediction probability on April 12, 2010. c. The distribution of prediction probability on May 17, 2010. d. The distribution of prediction probability on June 14, 2010.

Through an optimization of threshold that was mentioned in Sect. 2.4.2, the probability of 0.4 was used to convert the forecasted probability to a binary disease occurrence result. Table 2 summarized the forecasted results of the Bayesian network,

BP neutral network and FLDA. The result suggested that Bayesian network and FLDA produced more accurate forecasts than BP neutral network. For Bayesian network and FLDA, the Bayesian network outperformed FLDA at both heading stage and milk stage, which are important time points for prevention.

Table 2. Accuracy indices of three tested methods

	Reviving stage (%)	Jointing stage (%)	Heading stage (%)	Milk stage (%)
FLDA	82.98	72.50	52.75	88.02
BP neutral network	52.13	47.50	57.69	80.65
Bayesian network	62.92	63.18	79.48	94.75

4 Conclusions

The Bayesian network was successfully used to develop a forecast model of YR occurrence probability across vast area in this paper. The performance of the model was evaluated against a weekly survey data during wheat's key growth stages from 2010 to 2012. The results confirmed that the disease forecasted results are able to reflect the spatio-temporal development and distribution pattern of YR. Further, superior performance of the Bayesian network in comparing with BP neutral network and FLDA also demonstrated that the Bayesian network is of great potential in forecasting crop diseases at a regional scale.

Acknowledgments. This work was supported by the National Key R&D Program (2016YFD0300602) and the National Natural Science Foundation of China (41101395, 41601346).

References

1. Zeng, S.M.: Interregional spread of wheat Yellow Rust in China. Acta Phytopathol. Sin. **18** (4), 219–223 (1988)
2. Yang, Z.W., Shang, H.S., Pei, H.Z., Xie, Y.L.: Dynamic forecasting of stripe rust of winter wheat. Sci. Agric. Sin. **24**(6), 45–50 (1991)
3. Hu, X.P., Yang, Z.W., Li, Z.Q., Deng, Z.Y., Ke, C.H.: Prediction of wheat stripe rust in Hanzhong area by BP network. Acta Agric. Boreali-Occidentails Sin. **9**(3), 28–31 (2000)
4. Wan, A.M., Zhao, Z.H., Wu, L.R.: Reviews of occurrence of wheat stripe rust disease in 2002 in China. Plant Prot. **29**(2), 5–8 (2003)
5. Zeng, S.M.: Macro-phytopathology. Agriculture Press of China, Beijing (2005)
6. Chen, G., Wang, H.G., Ma, Z.H.: Forecasting wheat stripe rust by discrimination analysis. Plant Prot. **32**(4), 24–27 (2006)
7. Liu, R.Y., Ma, Z.H.: The prediction methodology of wheat stripe rust using combination model based on GM (1,1). J. Biomath. **22**(2), 343–347 (2007)

8. Yuan, L., Li, S.Q.: Prediction of wheat stripe rust by wavelet neural network. Microcomput. Inf. **25**(12–2), 42–43 (2009)
9. Cooke, B.M., Jones, D.G., Kaye, B.: The Epidemiology of Plant Diseases. Springer, Netherlands (2006). https://doi.org/10.1007/1-4020-4581-6
10. Xu, Y.P., Yao, X.H., Wang, C.S., An, W., Duan, Y.L.: Meteorological prediction of formation and development of winter-wheat stripe rust in Tianshui City, Gansu Province. J. Nat. Disasters **20**(1), 142–148 (2011)
11. Pan, Y.Z., Gong, D.Y., Deng, L., Li, J., Gao, J.: Smart distance searching-based and DEM-informed interpolation of surface air temperature in China. Acta Geogr. Sin. **59**(3), 366–374 (2004)
12. Li, J.L., Zhang, J., Zhang, C., Chen, Q.G.: Analyze and compare the spatial interpolation methods for climate factor. Pratacultural Sci. **23**(8), 6–11 (2006)
13. Wang, H.X., Liu, X.N., Ren, Z.C., Wei, J.Q., Pan, D.R., Hou, J.J.: Spatial interpolation of a precipitation-a case of Gansu Province. Grassl. Turf **32**(5), 12–16 (2012)
14. Wang, Z., Shi, Q.D., Chang, S.L., Wu, Y.J., Liang, F.C.: Study on spatial interpolation method of mean air temperature in Xinjiang. Plateau Meteorol. **31**(1), 201–208 (2012)
15. Zhang, L.W., Guo, H.P.: Introduction to Bayesian Networks. Science Press, Beijing (2006)
16. Jiang, L.X.: Research on naive Bayes classifiers and its improved algorithms. China University of Geosciences, Wuhan (2009)
17. Van Maanen, A., Xu, X.M.: Modelling plant disease epidemics. Eur. J. Plant Pathol. **109**, 669–682 (2003)
18. de Vallavieille-Pope, C., Huber, L., Leconte, M., Goyeau, H.: Comparative effects of temperature and interrupted wet periods on germination, penetration, and infection of Puccinia recondita f.sp.tritici and P.striifornis on wheat seedlings. Ecol. Epidemiol. **85**(4), 409–415 (1994)
19. de Vallavieille-Pope, C., Huber, L., Leconte, M., Bethenod, O.: Preinoculation effects of light quantity on infection efficiency of Puccinia striiformis and P.triticina on wheat seedlings. Phytopathology **92**(12), 1308–1314 (2002)
20. Madden, L.V.: Rainfall and the dispersal of fungal spores. Adv. Plant Pathol. **8**, 39–79 (1992)
21. Madden, L.V., Yang, X., Wilson, L.L.: Effects of rain intensity on splash dispersal of Colletotrichum acutatum. Phytopathology **86**, 864–874 (1996)
22. Madden, L.V.: Effects of rain on splash dispersal of fungal pathogens. Can. J. Plant Pathol. **19**, 225–230 (1997)
23. Madden, L.V., Wilson, L.L., Ntahimpera, N.: Calibration and evaluation of an electronic sensor for rainfall kinetic energy. Phytopathology **88**, 950–959 (1998)
24. Rapilly, F.: Yellow rust epidemiology. Annu. Rev. Phytopathol. **17**, 59–73 (1979)
25. Coakley, S.M., Line, R.F.: Quantitative relationships between climatic and stripe rust epidemics on winter wheat. Ecol. Epidemiol. **71**(4), 461–467 (1981)
26. Coakley, S.M., Boyd, W.S., Line, R.F.: Development of regional models that use meteorological variables for predicting stripe rust disease on winter wheat. J. Clim. Appl. Meteorol. **23**, 1234–1240 (1984)
27. Coakley, S.M., Line, R.F., McDaniel, L.R.: Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data. Ecol. Epidemiol. **78** (5), 543–550 (1988)
28. Friedman, N.: Bayesian network classifiers. Mach. Learn. **29**, 131–163 (1997)
29. Ferreiro, S., Sierra, B., Irigoien, I., Gorritxategi, E.: A Bayesian network for burr detection in the drilling process. J. Intell. Manuf. **23**, 1463–1475 (2012)

30. Te Beest, D.E., Paveley, N.D., Shaw, M.W., Van Den Bosch, F.: Disease-weather relationships for powdery mildew and yellow rust on winter wheat. Phytopathology **98**(5), 609–617 (2008)
31. Wang, H., Ma, Z.: Prediction of wheat stripe rust based on neural networks. In: Li, D., Chen, Y. (eds.) CCTA 2011. IAICT, vol. 369, pp. 504–515. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27278-3_52