Potato Late Blight Occurrence Period Forecasting Based on Polynomial Regression Analysis

Yu qunxiu
College of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, Yunnan, China
E-mail: qingdaoyqx@163.com

Zhang heng
College of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, Yunnan, China
E-mail: 489638065@qq.com

Zhang shouming*
College of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, Yunnan, China
E-mail: qingdaoyqx@163.com
*Corresponding Author

Wang chao
College of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, Yunnan, China
E-mail: wangchao19890516@163.com

Abstract—Application of 2003-2013 in Malone disease data of potato late blight, relationship to the incidence of late blight in Malone area and the local meteorological conditions were cluster analysis; Select closely associated with the potato late blight and plant pathology has obvious significance for the return of the independent variable meteorological factors, the establishment of origin of potato late blight forecasting model using polynomial regression analysis. Check back on the historical data, found that the prediction model is accurate. In this paper, on the basis of predecessors' research, combined with the physiological characteristics of potato late blight fungus, we use correlation analysis to analyze the temperature, relative humidity, rainfall, wet period and meteorological conditions, such as the relationship between infection rates of potato late blight in Malone area. According to the correlation coefficient, to find out the effect of meteorological conditions on potato late blight infection rate, rejecting the main meteorological conditions. It has a great important meaning of the relationship between the rate and the main meteorological factors of the subsequent analysis of potato late blight infection.

Keywords-Cluster analysis; Regression; analysis; Late blight; Prediction model

I. INTRODUCTION

Potato late blight is caused by a devastating disease caused by the pathogenic fungus, widespread in the world, very serious harm, has been listed as the world's largest food crop diseases. Yunnan is one of China's potato producing areas, all the year round all the potato planting, due to the unique climate conditions, Yunnan area summer rainy, cold and humid climate, provide favorable conditions for the occurrence and spread of late blight of potato. Since the 1990s, Yunnan province late blight disease is very serious, with an average cut of 30%, some 80% or even never produced. In Yunnan, the potato late blight has become an enemy of potato growth, seriously restricting the development of potato industry.

Because of serious harm to the important meaning and the potato late blight of potato production in the social and economic aspects of the production of potato late blight, caused by the close attention of scholars. Such as: Sun Maolin, using the International Potato Center late blight forecast software CASTOR in the Yunnan area were tested, it was found that the Ulrich model of the software can accurately predict the occurrence of various ecological types of Yunnan potato planting area of late blight[1]. Zhang Shouming, the Danish NegFry potato late blight warning system for experiments in the Malone area of Yunnan, predicted results are consistent with the actual situation [2]. Su-hua Chen thinks such as temperature and rainfall in June and July is a key factor in high rates of potato late blight [3]. Yu-bi yao thinks such as weather conditions are a major factor in the occurrence and development of potato late blight [4]. Zhao hong thinks potato late blight caused by potato species and pathogens to temperature adaptability [5].These studies mainly is the introduction or use the foreign mature early warning systems work, for the study of late blight forecast model is not much, to adapt to the establishment of the local potato late blight occurrence period forecasting model, has important significance for the prevention and control of disaster. This paper will occur late blight in Yunnan Malone area of factors according to the classification of meteorological conditions such as air temperature, relative humidity [6], wet period, according to the factors to select the samples cluster analysis, select the appropriate variable factor from, polynomial regression analysis method to establish the forecasting model, offer a kind of thinking and practices for potato late blight at period forecast.

II. THE GENERAL SITUATION OF RESEARCH AREA

Malone County is located in the eastern Yunnan Province, territory surrounded by mountains, belonging to the Wu meng Mountains, is located in low latitude plateau monsoon climate, winter and spring drought, summer wet season, dry and clear, with plenty of rainfall. According to the measured data, the annual average sunshine 1985 hours, the average annual rainfall of 1032 mm, the average annual temperature of 13.4 ℃, frost free period

241 days. Photo thermal and soil conditions are suitable for planting potatoes, potato robust growth throughout the growing season, infection rate is low, the degradation degree low, often can get high yield, but the area also has many rainy days, humidity and other factors, to late blight epidemic[7].

Regression analysis can build model between explain and be explained variables. In data mining environment, when all explain and be explained variables are continuous, regression analysis is a good choice. After of a large number of experiments and research, there are a large number of historical data statistics, this paper preliminary finding out the impact factor of influence of potato late blight infection rate, so this article can use the regression analysis to set the short-term forecast model.

In the social economy life, often encounter some phenomena are explained a curve relationship between variables and explanatory variables. Analysis and prediction of this type of phenomenon generally used non-linear regression forecast. In nonlinear regression, is commonly used polynomial regression [8].

Based on cluster analysis of multivariate linear regression model, the first thing is to determine the dependent variable, the model is the purpose of the study of meteorological factors on the development of potato late blight and the influence degree of the popular, infection rate of potato late blight value can reflect the influence degree, so the selection of late blight rate as the dependent variable. Then the most core of the problem is how to determine the variable can reflect the main influencing factors. This paper USES clustering analysis to help choose which can reflect the influence of the dependent variable of the most reasonable factors, to determine the independent variable.

A. The basic equations

If the explanatory variables interacted with each other, can the regression model[9] to fit the variables data using polynomial, and regression equations were obtained by determining the parameters of the model. The basic equation has variable polynomial regression model for interaction:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \cdots + \varepsilon \quad (1)$$

Among them, β_0 is called regression constant, β_1, β_2, …, β_p known as the linear regression coefficient, β_{12}, β_{23}, … As the interaction coefficient, they are all unknown parameters. Y is known as the dependent variable, x_1, x_2, …, x_p as dependent variable, ε as the random error (residual).

B. The choice of dependent variable

The multiple linear regression model is established based on, first to determine the dependent variable, a model is designed to influence the development of popular of meteorological factors on potato late blight of potato late blight, and establishment of infection rate exactly can reflect the extent of the impact, so select the late blight infection rate as the dependent variable.

C. Variable selection

The multiple linear regression models is established based on the selection of variables, to follow the main factors of ignoring secondary factors, but the main factors to determine the existence of the problem. In this paper, by using system clustering method to select the repressor, solve the problem. Specific approach is: the late blight infection rate in daily temperature, relative humidity, wet period is divided into three categories, according to the selected sample, systematic cluster analysis on the samples, selecting main factors as a repressor.

- temperature class of variable selection: Select 24 days except the temperature factors from the historical data, other factors completely the same sample, systematic cluster analysis[10] on the samples, the result is the best temperature is divided into three categories, the spectrum diagram as follows:

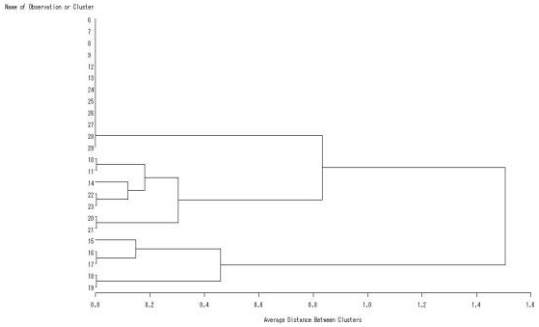


Figure 1. System cluster chart spectrum temperature

Seen from Fig .1 can be when the air temperature is 10 ℃ T<12 ℃ and 14 ℃ T<24 ℃, almost no impact on potato late blight, the temperature within each twice, to have different effects on late blight infection rate, the temperature is also added in 15 ℃ T<20 C. Therefore, the temperature T, can be divided into 10 ℃ ≤ T <12 ℃, 14 ℃ ≤ T <16 ℃, 16 ℃ ≤ T <18 ℃, 18 ℃ ≤ T <20 ℃, 20 ℃ ≤ T <22 ℃, 22 ℃ ≤ T <24 ℃ and return to the independent variable 15 ℃ ≤ T <20 ℃ temperature effects, etc. 7.

- Relative humidity class variable selection: Select from the historical data for 10 days in addition to relative humidity factors, other factors are the same sample, systematic cluster analysis on the samples, the result is the relative humidity is divided into two categories, the spectrum diagram as follows:

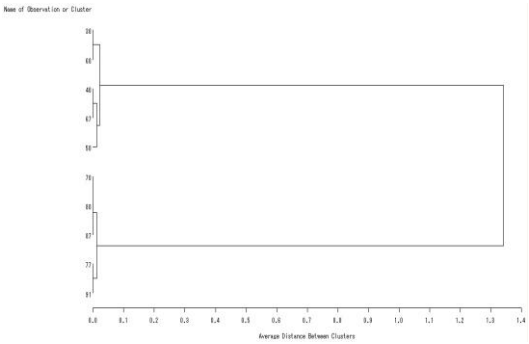


Figure 2. System cluster chart pedigree relative humidity

As can be seen from Fig .2, have a certain effect of relative humidity factors of epidemic of Potato Late Blight in the Malone area, not to form an effective wetting period, when the relative humidity is greater than or equal to 70, weak influence on the late blight, when the relative humidity is less than 70, have a certain influence on the late blight.

- Humid period class variable selection: Humid period refers to the continuous relative humidity is greater than a certain threshold (for late blight pathogen growth development of humidity) hours. Select 23 days except the humid period factors from the historical data, other factors completely the same sample, systematic cluster analysis on the samples, the result is the wet period is divided into three categories, the spectrum diagram as follows:

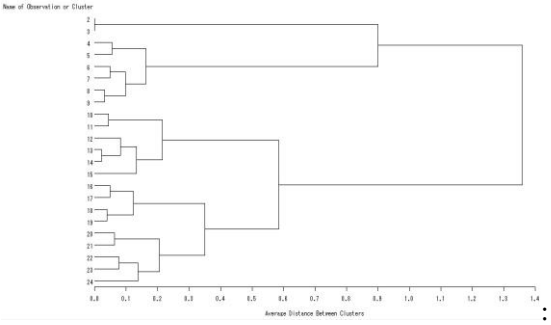


Figure 3. System cluster chart pedigree humid period

Seen from Fig .3 can be, humid period factors on potato late blight epidemic of Malone area for: When the wet period of less than 4 hours, the epidemic of late blight is very weak; When the wet period of 4~9 hours, have a certain impact on late blight epidemic, in order to facilitate the analysis, this paper put the wet period defined 4~9 hours for short humid period; When the wet period of 10~24 hours, for late blight epidemic influence is very strong, the wet period defined 10~24 hour long humid period. Therefore, regression of the humid period in W 4h ⩽ W<10h, 10h ⩽ W<24h short humid period long humid period 2 humid period.

D. Model checking

The above argument, the dependent variable into the polynomial regression model, and the polynomial regression model into linear model, significant t test and F test on it, and to examine the residual graph. If you are through, the model can be used to forecast, otherwise the model needs to re build.

III. EXAMPLES OF APPLICATION

Select the 2003~2012 Malone area of Yunnan province for 10 consecutive years of potato late blight disease data; establish the polynomial regression model, using the data of 2013 to test the accuracy of the model.

A. Parameter estimation

In accordance with the polynomial regression model, parameter estimation using the least square method, the selection of optimal regression equation stepwise regression analysis results are as follows:

$$y = -0.5645 + 0.6921x_1x_9 + 0.371x_2x_9 + 0.5645x_3x_9 + 0.8968x_4x_9$$

$$+0.8398x_5x_9 + 0.5682x_6x_9 + 0.4569x_1x_{10} + 0.0731x_2x_{10} + 0.1557x_3x_{10}$$

$$+0.6738x_4x_{10} + 1.3319x_5x_{10} + 0.706x_6x_{10} + 0.1112x_7$$

Among them: where y is the potato late blight infection rate;

x_7 is a day at the temperature of 15 ℃ ⩽ T<20 ℃ hours;

x_8 is a day of relative humidity in H <70% of hours;

x_9 is the short wet period, wet period 4h ⩽ W<10h values 0, 1;

x_{10} is a long wet period, wet period 10h ⩽ W<24h values 0, 1;

$(x_1 \sim x_7)\, x_9$ is all in a short period of moist air hours;

$(x_1 \sim x_7)\, x_{10}$ for each in the long wet period temperature section of the number of hours;

β_0 regression constants, β_1 , $\beta_2 \cdots$, β_{24} regression coefficient;

ε is a random error.

B. Significant test

The result of F test to examine the regression equation is effective, the variance analysis table as follows:

TABLE 1. Polynomial regression equations of the significant test table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	8104.36161	623.41243	206.01	<.0001
Error	376	1137.82187	3.02612		
Corrected Total	389	9242.18348			

Root MSE	1.73958	R-Square	0.8769
Dependent Mean	8.20756	Adj R-Sq	0.8726
Coeff Var	21.19480		

From table 1 can be observed, f=206.01 regression statistical significance test value, P value p<0.0001, through the F test, the regression relationship. Multiple coefficient of determination Adj R-Sq=0.8726 corrected, compared with close to 1, indicating that the regression model is good for the effect of data fitting, which can describe the variation of the regression relationship observed late blight infection rate of 87.26% of the Y value.

C. Regression diagnosist

Multiple linear regression equation after the t test and F test, just show a linear relationship between independent variable and dependent variable is significant, or is given to illustrate the effectiveness of the linear regression equation, but can't guarantee the data fitting is very good, cannot be ruled out due to accidental causes made the

model is not completely reliable. Only when the associated with the residual term in the model assumes that meet, can safely use the regression model. Therefore, when data fitting was not very good, should use the residual figure help diagnose regression effect and the quality of the sample data; check whether the model is to satisfy, in order to further modify the model.

D. Residual plot test

Using the residual graph examination in the model residuals meets the relevant assumptions. To confirm the data fitting effect. The regression equation residual plots are as follows:

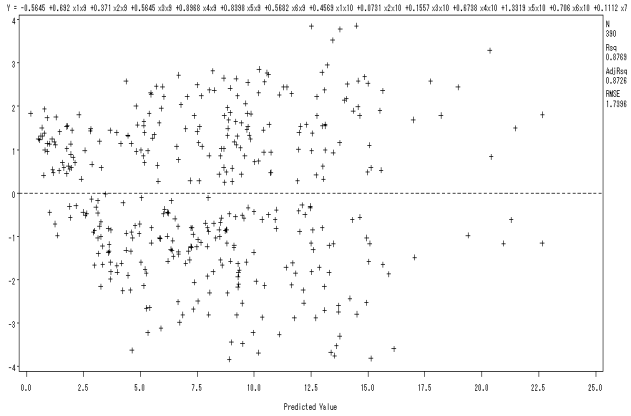
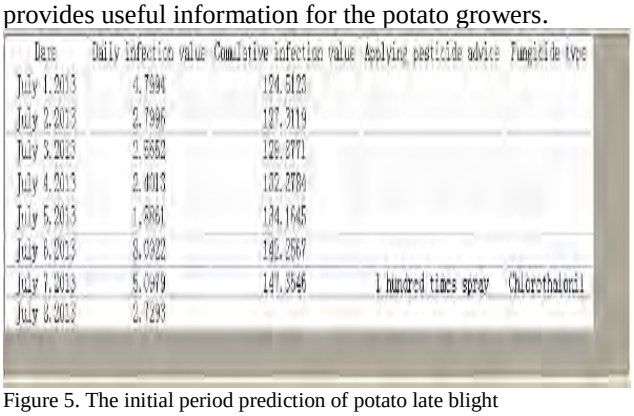


Figure 4. Multivariate polynomial regression equation residuals chart

The residuals from Fig .4 can be seen, the scatter generally in a horizontal axis as the center line of the band, basically meets the demands of normal distribution, shows that the regression model basically meet the basic hypothesis, regression equation is basically reasonable, can rest assured that the use of multivariate polynomial regression model to forecast the trend of potato late blight.

E. The model prediction

Through the test, the fitting precision of the polynomial regression model reached 87%, late blight occurrence period forecasting system using polynomial regression model for the initial period of prediction for future one growing season in study area. Fig . 5 shows the initial period, potato late blight forecast research area for July 7th, the actual date of July 2nd, the error for 5 days. This model can forecast late blight in the initial period, provides useful information for the potato growers.



Date	Daily infection value	Cumulative infection value	Applying pesticide advice	Fungicide type
July 1,2013	4.7994	124.5125		
July 2,2013	2.7996	127.3119		
July 3,2013	2.9552	126.2771		
July 4,2013	2.4013	132.2784		
July 5,2013	1.8961	134.1745		
July 6,2013	3.0822	145.2567		
July 7,2013	5.8949	147.3546	1 hundred times spray	Chlorothalonil
July 8,2013	2.7293			

Figure 5. The initial period prediction of potato late blight

IV. DISCUSSION AND CONCLUSION

A. The sick rate and seed potato late blight, disease resistant variety, cultivation technology and other relevant factors, when these factors are the same, the meteorological factors are the main factors of potato late blight epidemic. Establishment of potato late blight occurrence period forecasting model based on meteorological factors in Malone area of Yunnan Province, good prediction effect.

B. The data measured for the Malone area, where dew, foggy and other factors, led to the analysis results and the relative humidity has a strong correlation with rainfall, but lose relevance, so we should further expand the test area, increasing the amount of data, the model has more adaptability and universality.

REFERENCES

[1] Sun Maolin, Li Shulian, Zhao Yongchang. The application of CASTOR software to establish early warning system of potato late blight [J]. Yunnan agricultural science and technology, 2003 supplement: 154-157.

[2] [2] Zhang Shouming, He Huilong, Li Canhui. Early warning system [J]. Yunnan agricultural science and technology forecasting the potato late blight, 2003 (B12):131-135.

[3] Chen suhua, Zhang xiaolei, Yang Wanglin. Weather forecast method of potato late blight in middle Inner Mongolia study [J]. China's agriculture, 2012.28 (12) : 198-201.

[4] Yao biyu, Zhang cunjie, Wan xin. Research on the loess plateau of potato late blight occurrence and development relation with meteorological conditions [J]. Journal of plant protection, 2008, 34(4): 90-92.

[5] Zhao hong, Sun guowu. The potential effect of environment change on agricultural pollution [J]. Journal of meteorological drought, 2004, 22 (1) : 69-73.

[6] Sun Maolin, Zhao yongchang, Li Shulian. The popularity of potato late blight in Yunnan and warning system research [J]. Journal of Yunnan university, 2008, 30 (S1): 47-52.

[7] Peng jiuhui, Guo lixia, Ma xiuling. Chengde city, hebei province potato late blight meteorological condition analysis and prediction [J]. Journal of meteorological drought, 2009, 27 (1) : 57-60.

[8] RuanJing. SAS statistical analysis from entry to master [M]. Beijing: people's posts and telecommunications publishing house, 2009.4:141-143.

[9] Samprit Chatterjee, Ali S.Hadi. Stata [M]. Zheng Zhongguo, translated by Xu Jing. Beijing: Mechanical Industry Press, 2013.8:46-61.

[10] Michael Steinbach, Vipin Kumar. Introduction to data mining [M]. Fan zeming. Beijing: people's posts and telecommunications publishing house, 2011.1:355-358.