

II. RELATED WORK

In general models can be classified in different ways. Some of the authors classified models into three types: descriptive, predictive and conceptual. Descriptive models provide hypotheses or generalize experimental results, but they do not usually reveal the mechanisms underlying the processes. Predictive models, which are also descriptive, allow the prediction of the occurrence and the severity of epidemics. Both descriptive and predictive models use mathematical tools, such as simple or complex functions, regression and differential equations, or simple decision models. The conceptual models, also known as explanatory or analytical models, allow the identification of problems by distinguishing cause from effect and quantify the effects of specific events on pathogenic development [3].

In one research authors outline common methodologies that are used to quantify and model spatio-temporal dynamics of plant diseases, with emphasis on developing temporal forecast models and on quantifying spatial patterns [4]. Several models in cereal crops are described, including one for *Fusarium* head blight.

In [5] author demonstrated how analytical models written as differential equations could be integrated and used to quantify the various parameters associated with disease progress. In another research authors stated that the LDE (*Linked Differential Equation*) models are good for the SEIR (*Susceptible, Exposed, Infectious and Removed*) type, which is the standard modeling approach in human disease epidemiology, and is also widely used in plant disease epidemiology [6].

Authors in [7] predicted yields based on climatic factors using mathematical modeling. Factors such as water deficit, solar radiation, maximum and minimum temperatures which play a vital role at floral initiation were taken into consideration in the process of model creation.

Authors in [8] developed a generic mathematical model framework in which a compartmental SIR (*Susceptible-Infected-Removed*) model for plant-pathogen dynamics was coupled with biological control agent pathogen dynamics. This study was extended for further numerical computations, and computation take in consideration four main biological control agent mechanisms: induced resistance, antibiosis, competition and mycoparasitism. Each mechanism had their key parameter values in order to replicate the mechanism involved.

In another example authors formulated a simple mathematical model to find the optimal solution to control an invasive plant species. The objective functional is defined to minimize the area covered by the invasive species at the end of the time period which is coupled with the control variable. The term for the control variable was chosen to be in a quadratic form for simplicity of the analysis [9].

The optimization problem of plant disease control on a lattice with a spatial model for the spread of a plant pathogen over an agricultural region is presented in [10]. Author minimized the total level of infection during a single

agricultural season, at the least possible cost treatment.

In a paper presented in [11] authors used MLR (*Multiple Linear Regression*) in order to determine periods and conditions of *Deoxynivalenol* occurrence. Their work collected relevant information from 399 farms at Ontario in Southern Canada from 1996 to 2000 for the prediction of *Deoxynivalenol* occurrence. Daily precipitation, daily minimum and maximum temperatures, and relative humidity per hour were used as weather factors, and as a result of the experiment, it has been confirmed that time points of wheat growth, rainfall, and temperature are related with *deoxynivalenol* occurrence.

Authors in [12] used SPSS software to perform multiple regression analysis to develop the disease prediction models for rice blast prediction. In these models leaf blast severity was used as the dependent variable and the weekly average of various weather variables (maximum temperature, minimum temperature, maximum relative humidity, minimum relative humidity, rainfall and rainy days per week) one week prior to disease assessment were used as independent variables. Actual prediction accuracy of these regression models was thus, determined on the basis of coefficient of determination and percent mean absolute error of the actual values.

In one of the papers authors created model for weather forecast based on the use of simple mathematical equation using MLR equations that can be easily understood by a medium educated farmer [13]. Weather data are recorded from meteorological station. The weather parameters like maximum temperature, minimum temperature and relative humidity have been predicted using the calculated features depending upon the correlation values in the weather data series over different periods from the weather parameter time-series itself. For the prediction of relative humidity authors use time series of maximum and minimum temperature and rainfall. From the other side rainfall have been estimated using features of maximum and minimum temperature and relative humidity. The development phase of the model is to obtain MLR equations using input set and output parameter [13]. The coefficients of these regression equations have been used to estimate the future weather conditions.

III. MATHEMATICAL MODELING

A. Building models

At the beginning of the model creation objectives need to be cleared. These determine the future directions of the project in two ways. Firstly, the level of detail included in the model depends of the purpose for which the model will be used. Secondly, division between the system that need to be modeled and its environment must be determined. This division is well made if the environment affects the behavior of the system, but the system does not affect the environment. For example in the process of modeling the pathogenic infections on the fruit it is advisable to treat weather as part of the environment. In this context weather directly effects on the development of pathogens, but from the other side pathogens do not effect on the weather conditions. After

determination which system will be modeled the basic framework of the model need to be constructed. This framework reflects beliefs about how the system operates. These beliefs can be stated in the form of underlying assumptions. Future analysis of the system treats these assumptions as being true, but the results of such an analysis are only as valid as the assumptions. If the assumptions are sufficiently precise, they may lead directly to the mathematical equations governing the system.

Mathematical equations that will describe the system must be chosen carefully, because they may have unforeseen effects on the behavior of the model. Generated mathematical equations could be solved analytically or numerically. Full analytical solution for a stochastic model involves finding the distribution of outcomes. In some cases system can be described with differential equation or with a system of differential equations [14]. If the model consists of just one differential equation, then there is a good chance that it has already been studied. However, when models contain nonlinearity, analytic results are typically harder to obtain than for the corresponding deterministic system. If the model is more complicated, and especially if the structure of the model is likely to be changed, then it is hardly worth even trying to find an analytic solution.

If the resolving of the equations with the use of analytical methods is complex, numerical methods can be used to obtain approximate solutions. Although they can never have the same generality as analytical solutions, they can be just as good in any particular instance. For differential equations, numerical solution is exact, since rules laid down in the equations follow the evolution of the system. With a stochastic model, outcomes can be repeatedly simulated using a random number generator, and combine a large number of simulations to approximate the distribution of outcomes. For the numerical resolving of differential equations Euler's method or fourth order Runge-Kutte method could be used.

B. Studying models

In this step different quantitative methods can be distinguished. One of these, sensitivity analysis, can be used in order to find out how dependent outcomes are on the particular values of chosen parameters [14]. The goal of sensitivity analysis is to differs model parameters and assess the associated changes in model outcomes. This method is particularly useful for identifying weak points of the model. It can be used for identification of outliers in the data sets. For the simple models based on this analysis it is possible to differentiate the outcome with the respect to each parameter in turn, which is not the case with more complex models.

Evaluation of complex models can be done in two ways. The first is to develop some approximation by simplifying, or summarizing the model mathematically. In this case so called closure equations can be used to summarize the statistics of a stochastic model. In the practice closure equations are differential equations that need to be resolved with numerical methods. The second approach is both simple and more general, but may run against the grain. In this approach

predictors need to be obtained from the model for the carefully chosen values of the control variables. Outcomes need to be considered as the results of a designed experiment, and empirical response surface need to be fitted to it. In that way fitted surface will be used for model outcomes estimation in the future.

C. Testing models

Once the model is studied it is time to be tested against observations from the physical system which it represents. This process is usually called validation. The most convincing way of testing a model is to use it to predict data which has no connection with the data used to estimate model parameters. In this way the chances of obtaining a spuriously good match between model predictions and data are reduced to a minimum. The standard test that can be used for measuring the predictive accuracy is a Cross Validation test. Cross Validation measures the performance of the prediction system in a self-consistent way by systematically leaving out a few datasets during the training process, and testing the trained prediction system against those left-out datasets. Compared to the test on independent dataset, Cross Validation has less bias and better predictive and generalization power [12]. If the prediction process is based on just one dataset, with the use of cross validation prediction model can be tested. For example in *k-fold* cross-validation, the original sample is randomly partitioned into *k* subsamples. Of all the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k-1* subsamples are used as training data. The cross-validation process is then repeated *k* times, with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds are averaged to produce a single estimation.

If for one system multiple models are available, comparison can be used as testing method. Comparison will always contain an element of subjectivity, since there are many different aspects on which the decision can be based. If the comparison is performed well at the end of validation process system that shows the best performance can be selected as candidate for practical using.

D. Using models

Created model can be used as standalone prediction model in less complex systems. If the system is more complex, created model can be combined with different data mining and statistical techniques in order to get more precise prediction. In both cases data that represents independent variables collected in current time will be used for prediction of dependent variable. The software solution into which mathematical model could be incorporated need to have a sufficient level of abstraction. This means that end users (farmers or agricultural experts) with basic IT skills or without knowledge of mathematical modeling need to be able to use it. With each correct prediction data over which prediction is based, can be inserted in dataset which is in use for mathematical model creation. In this way precision of mathematical model will be improved.

IV. DATA ANALYSIS AND MODEL CREATION

In order to create and test mathematical model, dataset with independent and dependent variables is created first.

Measurements had been collected from the meteorological station placed near Prokuplje (region Toplica, in Republic of Serbia). The measurements are performed during 2014 and 2015, in the period from April to July. This particular period is selected based on the fact that in this time span observed diseases can infect the host fruits. Spore activity and visual check of infection is performed in the cherry fruit (lat. *Prunus cerasus*) orchard, which is most sensitive fruit species on these two selected pathogen diseases. The observed orchard has the surface of four hectares, with 4800 fruit trees.

Based on measured values dataset with 244 instances is created. From the initial dataset two separated datasets (training and test) are created. Into training dataset 204 instances are translated in order to be used for model creation. The remaining 40 instances are translated into test dataset, and used in the testing phase. Both datasets (training and test) consists of eight independent variables, and one dependent variable. Six of the independent variables represent weather parameters that are collected from mentioned meteorological station. Those weather parameters are: daily minimal, maximal and mean temperature, average humidity, rainfall and wind speed. Another two independent variables represent spore activity for the two observed pathogen diseases (*Monilinia laxa* and *Coccomyces hiemalis*). For all independent variables numeric representation is selected. All temperatures (minimal, maximal and mean) are presented in degrees Celsius, rainfall is presented in millimeters of water sediment per square meter, wind speed in kilometers per hour, and the average humidity is calculated by dividing the measured value by 100. Spores activity for both pathogen diseases are coded with 0 if the pathogenic spore is in passive state and with 1 if the pathogenic spore is in active state.

Dependent variable represents three possible prediction outcomes. The first outcome named *monilia* represents that the conditions for infection of *Monilinia laxa* are fulfilled. The second outcome named *coccomyces* represents that the conditions for infection of *Coccomyces hiemalis* are fulfilled. The third and the last outcome represents that there are no appropriate conditions for development either of one of two particular diseases.

For the data analysis and the model creation IBM SPSS Statistics tool is used. In order to see dependencies between each independent variable and dependent variables graph visualization is applied in MatLab and WEKA tool. The use of different tools for visualization and outlier detection provides better final results. Correlation between multiple variables is shown on Fig. 1. For this particular visualization Scatterplot matrices graph is used. The each variable is plotted against each other. Here each Y versus each X plot displays a plot for each possible Y-X combination. This type of matrix is effective when we are only interested in relationships between responses and predictors, which are entered separately.

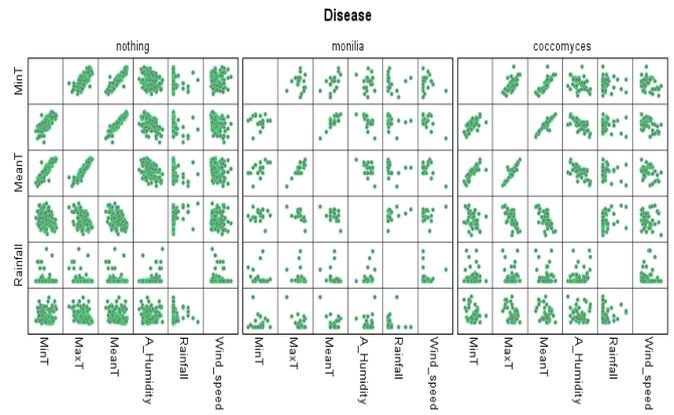


Fig. 1. Correlation between multiple variables

The similar graph, created just for one particular disease is presented on Fig. 2. This figure is separated in two groups. The group marked with 0 represents state in which spores for *Monilinia laxa* are not active. The second group marked with 1 represents state in which spores for disease *Monilinia laxa* are active. If we compare representation of independent variables based on these two categories we can see that parameters for some of the independent variable have different values.

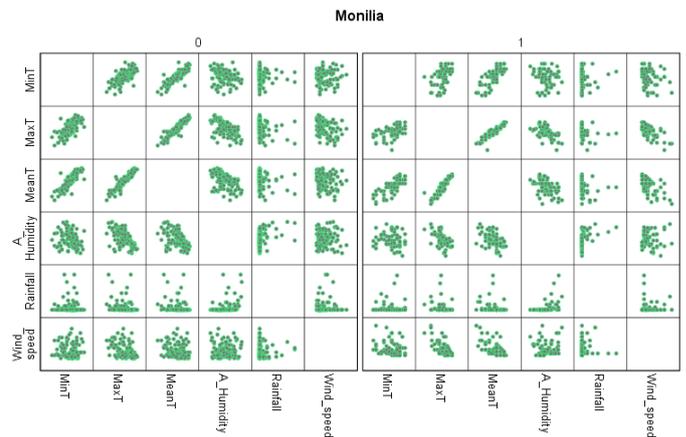


Fig. 2. Correlation between variables for particular disease divided in groups

This means that weather parameters in the time when this particular disease is in active state can be different than in the time when pathogenic spores are in passive state. This is not mandatory because there are many values that overlap. Beside visualization of mutual correlation this figure can be used for outlier's detection, in order of their reduction.

Representation of possible outcomes towards maximal temperature values is presented on Fig. 3. From this figure can be seen that all three dependent variables overlap throughout the entire range of the maximum temperature. In most of the cases the maximum temperature value is not appropriate for the development neither of one of two particular diseases. From the same figure can be seen that the majority of *Monilinia* infections were when the maximum temperature was in the range between 16°C and 22°C. The majority of *Coccomyces hiemalis* infections were when the maximum temperature was in the range between 25°C and 28°C.

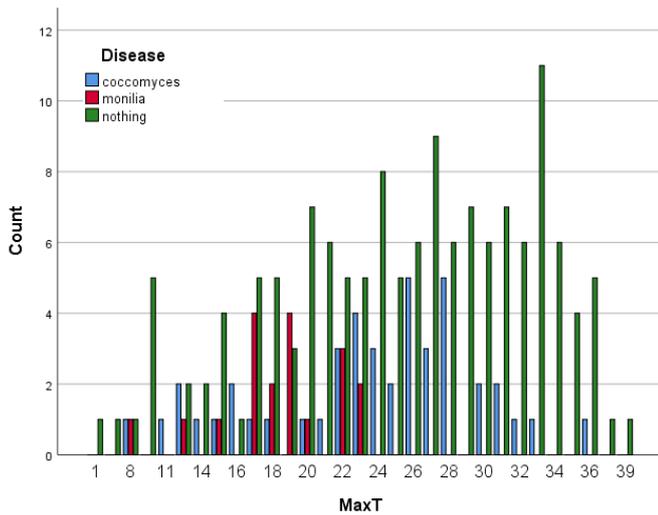


Fig. 3. Representation of possible outcomes towards maximal temperature

Based on created dataset linear regression model is created. For the model creation *SPSS Analysis* functionality is used. The outputs after the evaluation are presented in following tables. Table 1 represents basic model summary. The *R* value represents the simple correlation, and the value in the table indicates a medium degree of correlation.

TABLE I
MODEL SUMMARY

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.685 ^a	0.470	0.448	0.615

a. Predictors (Constant), MaxT, Coccomyces, Rainfall, Wind_speed, Monilinia, A_Humidity, MinT, MeanT

The *R Square* value in second column indicates how much of the total variation in the dependent variable, can be explained by the independent variable. In our case the value of

0.448 represents 44.8% of the variability of the dependent variable, which may be explained by independent variables, so that the bond strength is strong. Value for *Adjusted R Square* in practice is always lies between 0 and 1. A value of 1 indicates a model that perfectly predicts values in the target field, and a value that is less than or equal to 0 indicates a model that has no predictive value. *Std. Error of the Estimate* represents the standard deviation of the error term and the square root of the *Mean Square* for the *Residuals* in the Table 2.

Data in the Table 2 report how well the regression equation fits the data. This table indicates that the regression model predicts the dependent variable significantly well.

TABLE II
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	65.206	8	8.151	21.574	0.000 ^b
Residual	73.671	195	0.378	/	/
Total	138.87	203	/	/	/

b. Predictors (Constant), MaxT, Coccomyces, Rainfall, Wind_speed, Monilinia, A_Humidity, MinT, MeanT

The *Regression* row displays information about the variation accounted for created model. The *Residual* row displays information about the variation that is not accounted for created model. From the values for *F* and *Sig.* we can say that all independent variables are indeed different from each other and that they affect the *Disease* in a different manner.

The values in column *Sig.* indicates the statistical significance of the regression model that was run. Here, $p < 0.005$, which is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts the outcome variable.

TABLE III
COEFFICIENTS

	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
(Constant)	-1.087	0.587		-1.850	0.066	-2.245	0.072
MeanT	0.007	0.037	0.045	0.182	0.856	-0.066	0.079
A_Humidity	1.620	0.522	0.261	3.104	0.002	0.591	2.649
Rainfall	0.075	0.016	0.259	4.547	0.000	0.042	0.107
Wind_speed	0.005	0.014	0.025	0.366	0.715	-0.022	0.32
Monilinia	0.208	0.115	0.117	1.812	0.072	-0.018	0.434
Coccomyces	0.698	0.108	0.408	6.484	0.000	0.486	0.911
MinT	0.024	0.024	0.148	1.022	0.308	0.023	0.072
MaxT	-0.011	0.022	-0.093	-0.488	0.626	-0.53	0.032

a. Dependent Variable: Disease

In Table 3 we can see all necessary information in order to predict *Disease* from independent variables, as well as determine whether each of the variables contributes statistically significantly to the model. Furthermore, values in the *B* column under the *Unstandardized Coefficients* column can be used for creation of regression equation as in (1).

$$\begin{aligned} Disease = & -1.087 + 0.007(MeanT) + \\ & 1.620(A_Humidity) + 0.075(Rainfall) + \\ & 0.005(Wind_Speed) + 0.208(Monilia) + \\ & 0.698(Coccoomyces) + 0.024(MinT) - 0.011(MaxT). \end{aligned} \quad (1)$$

In the same way regression equation can be created in order to predict outcome for just one particular disease. For example, regression equation for prediction of *Monilinia* presents in (2). Equations like this can also be used for accuracy confirmation, obtained from equations involving several dependent variables. From other side if the accuracy of equations such is equation (1) is good, then the use of such equations requires less time consumption, in comparison with a larger number of equations (one for each disease).

$$\begin{aligned} Monilinia = & -0.286 + 0.000(MeanT) + \\ & 0.462(A_Humidity) + 0.006(Rainfall) - \\ & 0.006(Wind_Speed) + 0.333(Monilia) + \\ & 0.009(MinT) - 0.003(MaxT). \end{aligned} \quad (2)$$

By standardizing the variables before running the regression, all of the variables can be put on the same scale, and magnitude of the coefficients can be compared to see which one has more of an effect. Larger betas are associated with the larger *t*-values and lower *p*-values.

In order to test accuracy of the created mathematical model test dataset is used. Instances in test dataset (all 40) have not been previously used for model creation. Practically, from the initial dataset (with 244 instances), test dataset instances were randomly selected to cover entire period (from April to July). Based on the fact that outcome of dependent variable in all instances was known in advance accuracy of predicted outcomes was calculated. Values of the independent parameters from each of the test instances were put in the created mathematical equation, and the outcomes of dependent variable were calculated. Calculated outcomes and the known outcome values have coincided in 89% of the cases. This means that this regression equation is good representation of the system. Created equation can be used for the future predictions.

V. CONCLUSION

Accurate prediction of appropriate time for chemical protection provides number of benefits for farmers and chemical companies. In the same time prediction of the occurrence of diseases is complex job. This process can be

improved with mathematical modeling. Created mathematical model provides information about dependencies between weather parameters, active or passive pathogenic spores from one side, and occurrence of two particular diseases from other side. Beside this, model provides good basis for prediction of future occurrence of this two fruit diseases. The accuracy of prediction depends of selected independent variables, and of the accuracy of input data on which training process is based.

In the future authors will create software application on the basis of this model. The idea is that the entire prediction process, from data collection to farmer's notification be automated. This means that the system will collect data from the network of the automatic meteorological stations. After that in advance determined time spans, system will start prediction. If the appropriate conditions for diseases infection are fulfilled notification will be sent to the farmers.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of Republic of Serbia within the projects TR 32023 and TR 35026.

REFERENCES

- [1] R. N. Strange, P. R. Scott, "Plant Diseases: A Threat to Global Food Security", *Annual Review of Phytopathology*, vol. 43, pp. 83-116, Jul. 2005.
- [2] R. Gebbers, V. I. Adamchuk, "Precision Agriculture and Food Security", *Science*, vol. 327 no. 5967, pp. 828-831, 2010. DOI: 10.1126/science.1183899.
- [3] J. Kranz, D. J. Royle, "Perspectives in mathematical modeling of plant disease epidemics", In: *Scott PR and Bainbridge A (eds) Plant Disease Epidemiology*, London, England, Blackwell Scientific Publications, Oxford, 1978, pp. 111-120.
- [4] A. van Maanen, X. M. Xu, "Modelling Plant Disease Epidemics", *European Journal of Plant Pathology*, vol. 109, pp. 669-682, 2003.
- [5] V. der Plank, "*Plant Diseases: Epidemics and Control*", New York, London, USA: Academic Press, 1963.
- [6] M. J. Jeger, "Theory and plant epidemiology", *Plant Pathology*, vol. 49, pp. 651-658, 2000.
- [7] S.F. Foong, "An improved weather model for estimating oil palm fruit yield", International Conference on Oil Palm in Agriculture in the Eighties. Session-Prediction, Prospect and Forecast, Kuala Lumpur, Malaysia, pp. 235-261, 17-20 Jun 1981.
- [8] X. M. Xu, N. Salama, P. Jeffries, M. J. Jeger, "Numerical studies of biocontrol efficacies of foliar plant pathogens to the characteristics of a biocontrol agent", *Phytopatology*, vol. 100, pp. 814-821, 2010.
- [9] A. J. Whittle, S. Lenhart, L. J. Gross, "Optimal control for management of an invasive plant species", *Mathematical Biosciences and Engineering*, vol. 4, pp. 101-112, 2007.
- [10] M. L. Ndeffo Mbah, C. A. Gilligan, "Optimal control of disease infestations on a lattice", *Mathematical Medicine and Biology*, vol. 31, no. 1, pp. 87-97, 2013.
- [11] D. C. Hooker, A. W. Schaafsma, L. Tamburic-Ilicic, "Using weather variables pre-and post-heading to predict deoxynivalenol content in winter wheat", *Plant Disease*, vol. 86, no. 6, pp. 611-619, Jun, 2002.
- [12] R. Kaundal, A. S. Kapoor, G. Raghava, "Machine learning techniques in disease forecasting: a case study on rice blast prediction", *BMC Bioinformatics*, vol. 7, no. 1, pp. 485-501, Nov. 2006.
- [13] P. Mathur, S. Mathur, "A simple weather forecasting model using mathematical regression", *Indian Research Journal of Extension Education*, Special Issue, vol. 1, pp. 161-168, January, 2012.