13 | Research paper I Оригинални научни рад
ISSN 2566-4484 I Doi 10.7251/STP2014127P
pages 127-141

Bojan Popović, bojanpop94@uns.ac.rs, Faculty of Technical Sciences, University of Novi Sad

Igor Ruskovski, rus_igor@uns.ac.rs, Faculty of Technical Sciences, University of Novi Sad

Stevan Milovanov, s.milovanov@uns.ac.rs, Faculty of Technical Sciences, University of Novi Sad

Miro Govedarica, miro@uns.ac.rs, Faculty of Technical Sciences, University of Novi Sad

Dušan Jovanović, dusanbuk@uns.ac.rs, Faculty of Technical Sciences, University of Novi Sad

ANALYSIS OF THE INFLUENCE OF CLIMATE FACTORS ON DIFFERENT CROPS USING MULTIVARIABLE ANALYSIS AND REMOTELY SENSED DATA

Abstract:

Phenology modeling of the most common agricultural cultures based on time series of high spatial resolution satellite imagery and vegetation indices was conducted for several crop types. This data was used as an indicator of crops state in the multivariable correlation model as dependent variables. The influence of climate factors (temperature, air pressure, precipitation, insolation, cloudiness and humidity) on crops state was determined using multivariable correlation. This model allows prediction of delayed influence of climate factors on plant health and state through the values of NDVI.

Keywords: remote sensing, multivariable analysis, NDVI, climate factors

АНАЛИЗА УТИЦАЈА КЛИМАТСКИХ ФАКТОРА НА РАЗЛИЧИТЕ УСЕВЕ ПОМОЋУ МУЛТИВАРИЈАБИЛНЕ АНАЛИЗЕ И ДАЉИНСКИ ДЕТЕКТОВАНИХ ПОДАТАКА

Сажетак:

У овом раду је извршено моделовање најзаступљенијих пољопривредних култура на основу временских серија сателитских снимака високе резолуције и вегетационих индекса. Ови подаци су коришћени као показатељи стања усева у моделу мултиваријабилне корелације као зависне променљиве. Утицај климатских фактора (температура, притисак, падавине, инсолација, облачност и влажност ваздуха) на стање усева је утврђен помоћу мултиваријабилне анализе. Овај модел омогућује предвиђање одложеног утицаја климатских фактора на стање усева помоћу вредности NDVI.

Кључне ријечи:даљинска детекција, мултиваријабилна анализа, NDVI, климатски фактори

1. INTRODUCTION

Vegetation indices and climate parameters-based crop phenology analysis have been the subject of many scientific papers and many methodologies have been established in order to extract the most information about current plants state. That information is further used to plan corrective actions and neutralize hazards (diseases, pests, drought) through agro technical measures. The problem in this approach is in the chronological order: the damage happens and, although prompt, is detected later.

Recently, the tendency in the area of remote sensing is to use available datasets to develop models that will take input parameters and predict vegetation state and alarm if there is a harmful influence that needs to be prevented. Similarly, it is possible to develop models that would predict the yield in the early stages of crop development (these models require historical data about crop development and a large number of dependent variables).

Authors van Wart and others [1] concluded that crop simulation models can be used to estimate the influence of current and future climate on crop yield and food safety but demand long-term historical data about daily climate in order to get robust simulation. Many regions that grow crops do not have daily climate data. Alternatively, there are connected databases about climate data with overall coverage of the Earth that usually come from computer models with global flow, interpolated data from meteorological stations and remote sensing data. The aim of this study is to estimate the abilities of computer models with global flow to simulate crop yield potential that can serve as a measuring stick to estimate the influence of climate change on crop productivity. The conclusion is that the study results that rest on computer models with global flow are very uncertain. [1]

Based on the analysis of financial losses, author Nedacelov M. [2] concluded that in the time period from 2007 until 2012 the fluctuation of climate changes intensity induced great yield loss in the past few decades. The results show that emphasized manifestation of climate changes in the past years have added to the decline of the harvest of winter wheat compared to the harvest expected in 2020 calculated by the most drastic Special Report on Emissions Scenarios (SRES) B2A scenario. They concluded that the urgent prevention measures are needed. The climate changes adaptation measures have been taken in Moldova. [2]

Authors Ceglar and others [3] have estimated the influence of inter-season climate variability on inter-annual differences in yield of winter wheat and corn in 92 French administrative areas. Observed monthly time series of temperatures, precipitation and insolation during the vegetation season are analyzed alongside recorded yearly yield with the statistical approach based on partial least squares regression. The results show significant spatial differences in the contribution of main meteorological initiators in the variability of crop yield and time domain of maximal influence. Temperature and strong insolation are identified as the most important variables that influence corn yield in the south, east and north part of France, while precipitation is the most important in central and north-west parts of the country. Positive anomalies of precipitation during summer months lead to corn yield increase, while positive temperature and radiation anomalies have the opposite effect. Extensive irrigation in drought years reduces rain signal. Temperature differences in eastern France mostly influence the yield of winter wheat and precipitation differences influence north, north-west and south-east France. [3]

The aim of this paper was to produce a robust model for crop state prediction based on the current climate parameters. One of the critical steps for model development was to determine the time period of delayed impact of climate factors to the state of crops, quantified through the values of Normalized Difference Vegetation Index (NDVI). Based on the built model it is possible to predict the values of the NDVI for different crops if the current climate parameters are known. In this way, it is possible to roughly determine if there is a need for corrective actions in order to neutralize harmful influence of the weather on the plant health.

2. METHODOLOGY

The area of statistics deals with the acquisition, representation, analysis and application of data used in decision making, problem solving and product and process designing [4]. Statistical methods are used to improve variability description and understanding. Variability is presented by successive observations or phenomena that do not produce the same results. Two methods have been used in this paper: Z-score method and multivariable regression**.**

2.1. Z-score method

Outlier is a term in statistics that represents an observation significantly different from other observations [4]. This definition suggests that the outlier is something separated or different from the rest. Every data science project begins with the data acquisition and in that phase, there is no knowledge about the outliers. Outliers can be the result of data acquisition errors or the sign of the differences in data. There are several methods in statistics that can be used to detect the outliers [5]. Z-score method was used in this paper. Z-score method is a statistical test used to determine if the two mean values within a population differ when their variances are known and the sample sizes are large. This is performed under the assumption that the data is normally distributed and the standard deviation is known. The point of Z-score method is to detect and eliminate the outliers from the dataset [6].

2.2. Multivariable analysis

In the process of scientific explanation of the nature of some phenomenon, starting points consist of the data about one or more objects [7]. These objects can be: individuals, communities, different physical objects but also natural phenomena or a result of human activities. Sometimes it is not possible to look at the nature of the object as a whole. However, it is possible to observe one multidimensional phenomena that consists of several characteristics. Those characteristics are the subjects of the observation, and are usually called variables. Multivariable analysis represents an aggregate of statistical methods that simultaneously analyse multidimensional observations acquired for each unit of observation of examined objects. Assuming that the data about *j* characteristics about *i* objects was acquired and that this data was presented as a matrix (rows represent objects, columns represent variables), the table of data has the following structure:

$$\begin{array}{cccccc}
 & Variable\ 1 & Variable\ 1 & \dots & Variable\ j & \dots & Variable\ p \\
Feature\ 1: & X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\
Feature\ 2: & X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2p} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
Feature\ i: & X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
Feature\ n: & X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np}
\end{array} \tag{1}$$

where (i, j) element of the matrix represents the value of *j* variable measured on *i* object. In matrix notation, this matrix is noted as X, or X_{ij}, i=1,2,...,n; j=1,2,...,p.

Methods for data matrix analysis are categorized into two categories: methods of dependence and methods of interdependence. If the goal is to examine dependence between two groups of variables, where one group represents dependent and the other group are independent variables, a former group of methods is used. Later group is used when there's no apriori, theoretical base to divide all the variables into two groups. With the methods of dependence, one or more dependent variables have to be predicted based on the group of independent variables. In the other case this is not mandatory.

2.2.1. Multivariable regression

The main objective of a regression is to discover as many factors (independent variables) that influence a dependent variable. First assumption is that if there are more variables included in the model, latent variables (standard errors) will have less impact [8]. It is very important to decide which variables will be included in the model. Basic multivariable regression model is defined with the following:

$$\widehat{x_\iota} = a_1 + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_m x_{im} + \varepsilon_i, \ \ i = 1, 2, \dots n \tag{2}$$

where:

a_1 – free term

$\widehat{x_\iota}$, *i = 1,2,...,n* – single value of regression

x_{ij}, *i = 1,2,...,n, j = 1,2,...,m* – values of independent variables

b_j, *j= 1,2,...,m* – regression coefficient

ε_i, *i + 1,2,...,n* – latent variable

m – number of independent variables

n – sample size

This model offers best possible prediction of dependent variable values based on the values of independent variables, if all of the assumptions are met. It is possible to conclude the relative impact or importance of each of the independent variables based on the regression coefficients if those coefficients are converted into beta coefficients. These coefficients are a result of variable values standardization. One of the assumptions is that in order to use regression analysis, there has to be linear dependence between the variables. It is a mandatory assumption because the analysis begins with the calculation of simple correlation coefficients for all tuples of variables, and all these calculations require linear relationship between the tuple members.

2.2.2. Multicollinearity

Multicollinearity shows the interdependence between independent variables. The bigger the multicollinearity, the more it reflects on beta coefficients and they can't further be used to show relative impact of each independent variable. The reason behind this is that regression coefficients, b and beta are always calculated so that they give the best possible prediction of dependent variable Y, and not to show relative importance of each independent variable X. When the multicollinearity is low or doesn't exist, then regression coefficients are proportional to simple correlation coefficients and both of those give similar ideas about relative importance of independent variables. If there is a significant multicollinearity, then the most important independent variable gets the real value of beta coefficient, while the rest get lower values so that interdependence and mutual impact of the independent variables is avoided [9].

2.2.3. Variance inflation factor (VIF)

Variance inflation factor quantifies the level of correlation between one independent variable (predictor) and other independent variables in the model. It is used to determine collinearity and multicollinearity. Larger values show that it is difficult or impossible to determine the contribution to the model. Multicollinearity produces the problem in multivariable regression because the input factors have mutual impact, they are not independent which makes it hard to test how many combinations of independent variables influence dependent variables. Multicollinearity reduces the power and legitimacy of the model. The extent to which the predictor is correlated to other predictors can be quantified as R^2 statistics of the regression, where the predictor is predicted through all the other predictors' values [10]. VIF is then calculated as:

$$VIF = \frac{1}{1 - R^2} \tag{3}$$

VIF can be calculated for each predictor in the model. Value of 1 means that the specified predictor is not correlated to any other variables. If the value is larger, it means that there is more correlation between the observer variable and other variables in the model. Values larger than 5 are classified as medium and high values, and values larger than 10 are classified as very high values. This classification should be taken with a caution, because sometimes the value of 2 can cause issues. If one variable has a high VIF, it means that there is at least one more variable with a high VIF (two highly correlated values).

2.2.4. R-Squared

Multiple regression also shows how strong the interdependence is between dependent variable and all independent variables through the R index. R^2 index shows what the percentage of variability of dependent variables is explained through variability of independent variables. Since the correlation index and determination index are calculated based off the acquired data, there is no methods to improve them. Still, it is advised that a pilot research should be conducted in order to identify all the variable with the most impact, and only after that the whole research should be done. Multicollinearity can be determined through a specific indicators such as level of tolerance. Level of tolerance is a proportion of variable variance that is not connected to other variables in the regression model. High level of tolerance, above 0.8 means that those variables are relatively not correlated with other variables. Low level of tolerance, under 0.2 means that there is a high multicollinearity and that that variable doesn't contribute a lot to the explanation of dependent variable in the model. Their statistical significance should be tested when the results are explained. If R, b and beta are not statistically significant, the conclusion is that no independent variable has any real connection to the dependent variable. If all of the regression coefficients b are significant, then the index of correlation R will be significant.

2.2.5. Heteroscedasticity

Heteroscedascity refers to the situation where variability of dependent variable is not equal in the range of values of independent variables that predicts it. Scatterplot of these variables is often conic shaped, where dependent variable expands or narrows when independent variable grows [11]. The opposite case is homoscedasticity, which shows that the variability of dependent variable is the same within the scope of values of independent variable.

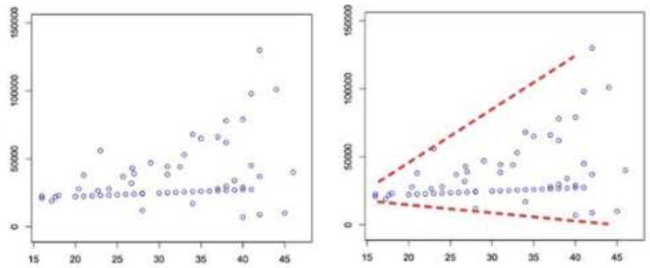

Figure 1. *Heteroscedasticity in the data (notable differences in the values of standard deviation alongside horizontal axis)*

Heteroscedascity influences the analysis results in the following ways:

- Although it does not provoke bias in the coefficients estimation, it makes them less precise. Lower precision means that it's probable that estimated coefficients are further away from the correct value
- Heteroscedasticity tends to produce p-values lower than they should be. This effect happens because heteroscedasticity increases coefficient estimates, and OLS is not able to detect this increase. This problem can cause the conclusion that the model is statistically significant when in fact it is not.

2.2.6. Autocorrelation

Autocorrelation refers to the extent of correlation between the values of one variable in different moments of observation [12]. The concept is usually considered in time series where observations happen in a different time domains (temperature during one month). If two values that are close to each other in time domain were more similar to each other in comparison to other values that were observed later during the month, it could mean that the data is autocorrelated. This can cause problems in conventional analyses that assume independence of observations.

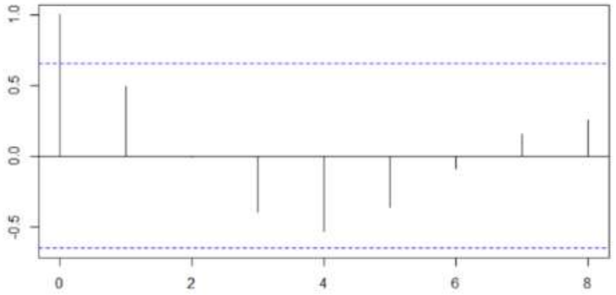

Figure 2. *Correlogram for demo data. Lag is connected to the order of correlation. Correlation has the value of 1 when there is no lag. With the lag present, the value of the correlation changes and it can be within or outside of the limits*

Autocorrelation can be seen in any dataset if it is observed based on the error sampling. That is the reason why statistical tests should be performed in order to avoid the situation of sampling causing autocorrelation. Standard test for this is Durbin-Watson. This test explicitly tests for correlation of first order. Problems that arise due to autocorrelation are OLS coefficient estimates that are not good enough as well as the estimates performed off those estimates, data overfitting, low values of standard errors, high t-statistics values.

2.2.7. Error measurements

Mean absolute error measures the average value of error in the prediction set, without the consideration of their directions. It is the average value for the sample of absolute differences between predicted and true values of observation where all the differences have the same weights.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{4}$$

If the absolute value is omitted, mean absolute error becomes mean bias error and it's usually meant for measuring model bias. This error can provide a lot of information, but has to be carefully analyzed because there is a chance of positive and negative errors undoing each other.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{5}$$

Given the fact that RMSE errors are first squared before the average value is calculated, RMSE gives relatively higher weights to larger errors. This means that RMSE is more useful when the goal is to eliminate larger errors [13].

2.3. NDVI

Green plants absorb sunlight and use it as an energy source in the process of photosynthesis. Chlorophyll, the plant leaves pigment, absorbs visible part of the spectrum (0.4-0.7 nm) and reflects near infrared spectrum (0.7-1.1nm). The absorption at these wavelengths would make plant overheat and the tissue would be damaged. That is why green plants look relatively dark in the visible spectrum (RGB) and relatively bright in the NIR band [14].

It's well known that the combination of near infrared and red bands can be used as an indicator of plant's health and give more useful information than separate bands can discover. NDVI is a quantitative measure of a plant's health based on the way that the plants reflect lights at different frequencies. The equation for NDV was developed several decades ago in order to use satellite images in agriculture. The structure of the equation makes it insensitive to the total brightness of the scene. Basically, the relation between NIR and red band was described through the equation and that relation doesn't change with the total scene brightness.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{6}$$

Today, NDVI is often used in many different analyses. Agricultural producers use NDVI to measure biomass, while in forestry it is used to quantify forest farms. Its most important application is that NDVI allows detection of changes in plants up to two weeks before human eye could see. That way it is possible to discover diseases, pests, fungi on time and react properly [15].

3. RESULTS

Proposed methodology has been tested on the study area shown in Figure 3. Data that was used is consisted of 10 parcels located in the vicinity of Bečej, Republic of Serbia. 20 sample points were generated for each parcel (200 in total) and NDVI was calculated for each sample point based on the satellite images available for this area for the year of 2017. Over 15 Sentinel-2 satellite images were used in the sampling process.

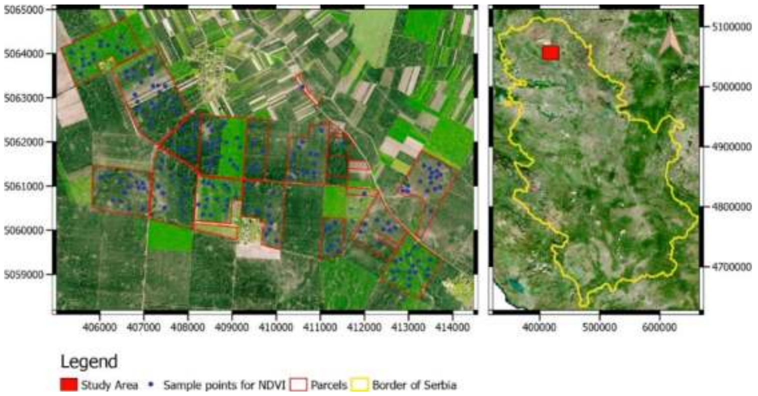

Figure 3. *Parcels, located in the vicinity of Bečej, Republic of Serbia, which were used for development and testing of regression model*

Data visualization plays a very important role in statistical analysis. Certain rules and exceptions in the data can be seen this way that can serve as a starting point for future decisions in the analysis. One of the visualization methods is BoxPlot. This method displays 5 parameters that describe the data: minimal value, first quartile, median, third quartile and maximum value. The data that fall under the minimal value and above the maximum value are called outliers. Based on the box plot (Figure 5) it is visible that each culture has observations that are significantly different than others within the same culture and the conclusion is that they are the outliers that need to be excluded from the future analysis in order not to skew the results.

Box plot of wheat (Figure 4) gives insight into a large range of values throughout the year which leads to a deeper analysis of this culture. The result of this research shows that there are two types of wheat in the Republic of Serbia: one that is sowed in the spring and one that is sowed in the autumn.

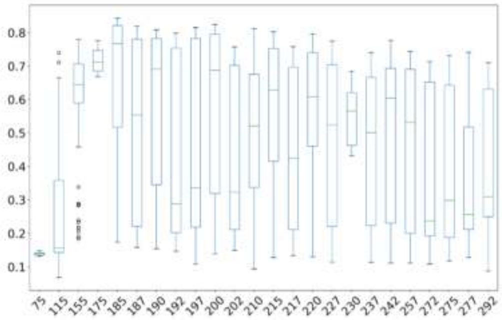

Figure 4. *BoxPlot of wheat data with a notable wide range of values*

The data has been filtered based on the calendar of agricultural actions, two types of wheat were separated and analysed separately.

Even though the negative buffer eliminated the possibility of sample points falling out of the specified culture, the additional test for outliers was performed. The possibility of outliers existence can be explained by the following situations: sample points fell on weed, cloud, passage in the field, excess water etc. The value of the vegetation index can be significantly different and this is why the Z-score test was performed. All the values that were declared an outlier were removed from the dataset, which can be seen on the example of winter wheat (Figure 5).

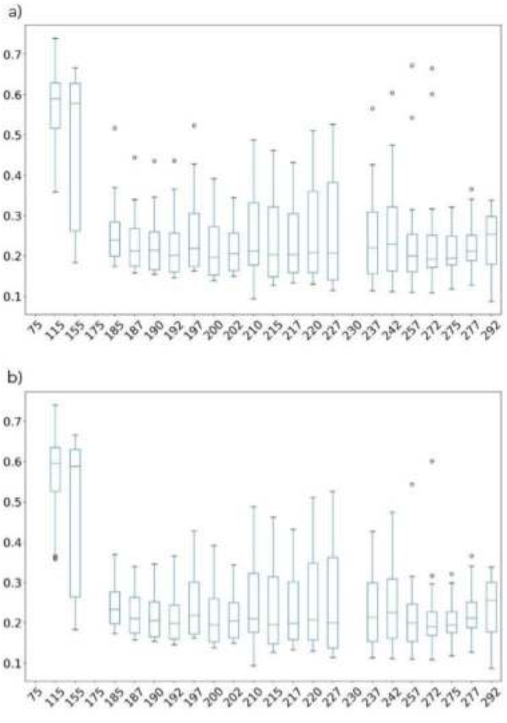

Figure 5. *BoxPlot for winter wheat before and after Z-score test for outliers*

When all the data is within the limits of Z-score and all the outliers are eliminated, the mean value of the index is calculated for each culture and each date. This results in a value that corresponds to each date (Figure 6).

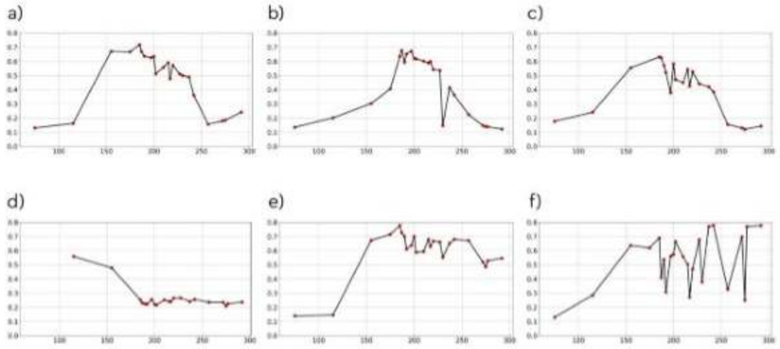

Figure 6. *Mean values of the vegetation index (NDVI) for: corn (a), sugar beet (b), barley (c), winter wheat (d), spring wheat (e) and clover (f) during the year of 2017*

In order to get a mathematical model of crop phenology during the whole period of observation, approximation with a polynomial function of 4th degree was done through nonlinear regression.

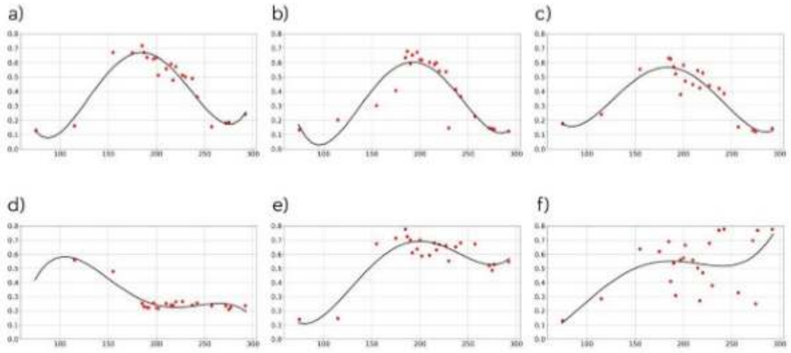

Figure 7. *Mean values and regression curve for: corn (a), sugar beet (b), barley (c), winter wheat (d), spring wheat (e) and clover (f) during the year of 2017*

Meteorological data was taken from the archive of the Republic Hydrometeorological Service of Serbia for the year of 2017 [16]. The data included in this research are air pressure, temperature, precipitation, insolation, humidity and cloudiness and refer to the geographical area in the vicinity of Novi Bečej. After the data has been loaded and checked for completeness, the dates of interest were picked. Meteorological data were taken for the time period from 31. of March until 2nd of October in 2017. This time domain was determined in the accordance of crop development.

The first step in the meteorological data analysis was the test for multicollinearity. From a mathematical standpoint, the problem with multicollinearity is that it produces unreliable coefficient estimates. In addition to that, standard errors of the coefficients become artificially enlarged. Since standard error is used to calculate p-values, this leads to a bigger probability that a variable will be declared statistically insignificant when in fact the truth is the opposite. A correlation matrix was generated and the values are presented in Table 1.

Table 1. *Correlation coefficient values for the meteorological data*

	air pressure	temperature	humidity	insolation	cloudiness	precipitation
air pressure	1	-0.297766	-0.015235	0.084048	-0.197213	-0.126089
temperature	-0.297766	1	-0.701491	0.61549	-0.493282	-0.219265
humidity	-0.015235	-0.701491	1	-0.780406	0.731885	0.381271
insolation	0.084048	0.61549	-0.780406	1	-0.838166	-0.248826
cloudiness	-0.197213	-0.493282	0.731885	-0.838166	1	0.270916
precipitation	-0.126089	-0.219265	0.381271	-0.248826	0.270916	1

Correlation heatmap (Figure 8) is often used when there are several variables. This offers a great insight into variables that have a high correlation coefficient (darker colors). By looking at this heatmap, it is possible to determine several highly correlated variables. For example, insolation and cloudiness are highly correlated. This is expected, since the amount of sunlight is directly tied to the cloudiness.

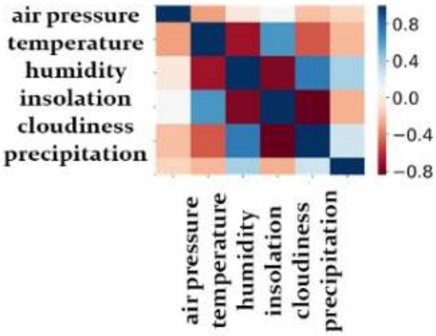

Figure 8. *Graphical representation of meteorological parameter correlation. Darker colours represent strong positive (blue) and negative (red) correlation*

Still, systematic search and removal of the variables with a high degree of correlation is a must. One of the methods used in this process is Variance Inflation Filter – VIF which measures how much the variable contributes to the amount of standard error in the regression model. When there is a significant multicollinearity, VIF will have a high value for the variable used in the model. General advice is that if any variable has a VIF factor 5 or higher, it should be excluded from the model. In this case, since no factor has a value higher than 5, all the variables will be kept in the model (Table 2).

Table 2. *VIF values for meteorological data. Value of VIF for insolation is close to the limit value of 5, but this factor was kept in the analysis.*

air pressure	1.359661
temperature	2.565814
humidity	3.790195
insolation	4.484016
cloudiness	3.820751
precipitation	1.202791

For the better insight into data, all standard statistical parameters (mean, standard deviation, minimum and maximum etc) were calculated and are shown in Table 3.

Table 3. *Standard statistical parameters for meteorological data*

	air pressure	temperature	humidity	insolation	cloudiness	precipitation
count	186	186	186	186	186	186
mean	1005.640323	19.625806	63.892473	9.35	4.098925	1.701613
std	4.325417	6.095375	13.057844	4.32494	2.883729	5.630424
min	994.7	2.7	37	0	0	0
25%	1002.725	15.475	54	6.8	1.7	0
50%	1005.65	20.1	62.5	10.5	4	0
75%	1008.1	24.075	72	12.675	6.6	0.275
max	1018.5	31.8	97	14.9	10	48.6
+3_std	1018.616573	37.911933	103.066006	22.324821	12.750111	18.592885
-3_std	992.664072	1.33968	24.718941	-3.624821	-4.552261	-15.189659

After the data has been cleared of errors, the next step is the model development. The first step is the definition of dependent and independent variables, and then the data needs to be divided into training and test sets. Independent variables are meteorological data, and dependent variable is the NDVI value. Good ratio of test and train data division is 20% to test and 80% to train the model.

When the data is fitted in the model, results are analysed. First part consists of the intercept model analysis. Intercept value is the value of dependent variable when all of the independent variables are zero. For each slope coefficient, it is the estimated difference of dependent variable for single unit change of specific independent variable. Intercept model variables are shown in Table 4.

Table 4. *Intercept values of the model. If all of the variables except temperature are zero, single unit change in temperature would contribute to a change of 2.9% to NDVI*

air pressure	0.0055
temperature	0.029
humidity	0.0046
insolation	0.023
cloudiness	0.027
precipitation	0.00064

The second part of the analysis consisted of tests for normal distribution of regression residuals, residual homoscedasticity and lack of multicollinearity (this has already been tested). Using various statistical tests all of the assumptions were confirmed: the residuals are homoscedastic with normal distribution. To get a better feel of model fitting, several indicators were calculated: mean average error, mean square error and root mean square error. The values of these errors are shown in Table 5.

Table 5. *Values of mean average error, mean square error and root mean square error*

MSE	0.0166
MAE	0.108
RMSE	0.129

R^2 provides a way to measure goodness of fit. Larger R^2 means better fitting. One of the limitations is that this value gets larger if there are more variables, which leads to larger R^2 value when new variables are added even though they are not necessarily good choice. R^2 value in the case of corn is 0.56 which is a low value, but it was expected in a way. It is normal that the climate changes affect the crops only after several days, not instantly. Keeping this in mind, NDVI will not show results in real time. For example, if day 100 is observed and the value of NDVI is 0.45 and it rains, an increase of NDVI value is expected in 7 to 10 days. In order to identify the time domain when the climate change impact is the strongest, the same above-explained process was repeated for time ranges from 1 to 14 days, and R^2 indicates the impact that has happened.

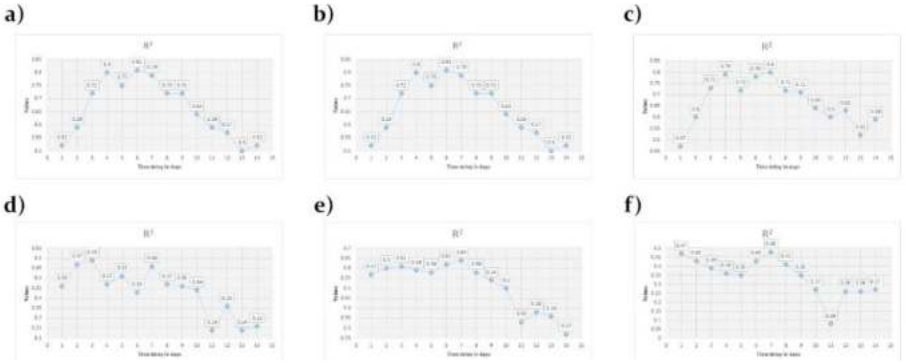

Figure 9. *Values of R2 with time delay taken into consideration for: corn (a), sugar beet (b), barley (c), winter wheat (d), spring wheat (e) and clover (f)*

The conclusion of the step above is that the climate parameters will have the strongest impact after 6 days. The analysis parameters have changed and the values for the new model are given below:

Table 6. *New values for model intercept*

air pressure	0.0035
temperature	0.029
humidity	0.0037
insolation	0.016
cloudiness	0.014
precipitation	0.0025

Table 7. *New model values for MSE, MAE and RMSE*

MSE	0.00736
MAE	0.0714
RMSE	0.0858

Table 8 shows the values of the 95% confidence interval for each meteorological parameter. Hypothesis tests were performed after this with the goal of determining statistical significance of coefficients estimates. Null hypothesis was that there is no connection between independent and dependent variables, while the alternative hypothesis claimed the opposite.

Table 8. *95% Confidence intervals for each of the meteorological parameters*

	0	1
air pressure	-0.001155	0.007161
temperature	0.024506	0.032801
humidity	0.001054	0.005723
air pressure	0.006921	0.022144
temperature	0.004110	0.025483
humidity	-0.001301	0.004720

Table 9 presents the p-values for each parameter, and Table 10 gives the summary of p-values for meteorological parameters before and after the removal of excess variables.

Table 9. *Estimated p-values for meteorological parameters*

air pressure	1.56e-01
temperature	1.74e-29
humidity	4.68e-03
air pressure	2.23e-04
temperature	6.92e-03
humidity	2.64e-01

Table 10. *P-values report for meteorological parameters before (above) and after (below) removal of insignificant independent variables*

OLS Regression Results						
Dep. Variable:	NDVI_corn		R-squared:			0.698
Model:	OLS		Adj. R-squared:			0.688
Method:	Least Squares		F-statistic:			68.97
Date:	Fri, 27 Sep 2019		Prob (f-statistic):			5.77E-44
Time:	17:18:12		Log-Likelihood:			155.83
No. Observations:	186		AIC:			-297.7
Df Residuals:	179		BIC:			-275.1
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std. err	t	P>|t|	[0.025	0.975]
const	-3.5804	2.147	-1.667	0.097	-7.818	0.657
air pressure	0.0038	0.002	1.425	0.156	-0.001	0.007
temperature	0.0287	0.002	13.633	0.000	0.025	0.033
humidity	0.0034	0.001	2.864	0.005	0.001	0.006
insolation	0.0145	0.004	3.768	0.000	0.007	0.22
cloudiness	0.0148	0.005	2.732	0.007	0.004	0.25
percipitation	0.0017	0.002	1.12	0.264	-0.001	0.005
Omnibus:	7.84		Durbin-Watson:		0.602	
Prob(Omnibus):	0.02		Jarque-Bera(JB):		8.213	
Skew:	0.582		Prob(JB)		0.0165	
Kurtosis:	2.769		Cond. No.		2.77E+05	
OLS Regression Results						
Dep. Variable:	NDVI_corn		R-squared:			0.698
Model:	OLS		Adj. R-squared:			0.688
Method:	Least Squares		F-statistic:			68.97
Date:	Fri, 27 Sep 2019		Prob (f-statistic):			5.77E-44
Time:	17:18:12		Log-Likelihood:			155.83
No. Observations:	186		AIC:			-297.7
Df Residuals:	179		BIC:			-275.1
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std. err	t	P>|t|	[0.025	0.975]

const	-0.5474	0.11	-4.974	0.000	-0.765	-0.33
temperature	0.0275	0.002	14.492	0.000	0.024	0.031
humidity	0.0036	0.001	3.193	0.002	0.001	0.006
insolation	0.015	0.004	3.902	0.000	0.007	0.023
cloudiness	0.0133	0.005	2.51	0.013	0.003	0.024
Omnibus:	6.424		Durbin-Watson:		0.578	
Prob(Omnibus):	0.04		Jarque-Bera(JB):		6.5	
Skew:	0.427		Prob(JB)		0.0388	
Kurtosis:	2.672		Cond. No.		9.65E+02	

Table 11. *True and estimated values of NDVI for test dataset*

Day	True	Estimated	Residual	143	0.4739	0.4705	0.0035	198	0.6486	0.6183	0.0302
94	0.0866	0.2648	-0.1782	144	0.4824	0.5122	-0.0299	207	0.6113	0.6429	-0.0316
101	0.1172	0.0453	0.0719	146	0.4989	0.5567	-0.0579	208	0.6062	0.6585	-0.0523
104	0.1351	0.0079	0.1272	149	0.5226	0.5637	-0.0411	212	0.5836	0.6759	-0.0922
106	0.1484	0.0837	0.0647	156	0.5725	0.5139	0.0586	217	0.5516	0.6629	-0.1112
109	0.1701	0.2887	-0.1186	168	0.6361	0.6088	0.0272	220	0.5306	0.5487	-0.0181
119	0.2540	0.2825	-0.0285	171	0.6470	0.5804	0.0667	252	0.2744	0.2632	0.0111
121	0.2721	0.3041	-0.0320	172	0.6502	0.6217	0.0285	255	0.2532	0.3500	-0.0968
125	0.3092	0.4110	-0.1018	183	0.6695	0.6227	0.0468	257	0.2400	0.2914	-0.0514
130	0.3562	0.4582	-0.1020	187	0.6692	0.5988	0.0704	260	0.2218	0.3302	-0.1084
132	0.3749	0.4149	-0.0400	192	0.6634	0.4786	0.1848	268	0.1854	0.1578	0.0275
137	0.4211	0.2805	0.1405	195	0.6570	0.6144	0.0426	270	0.1795	0.3542	-0.1747
138	0.4301	0.5177	-0.0876	196	0.6544	0.6262	0.0282	271	0.1772	0.3055	-0.1283

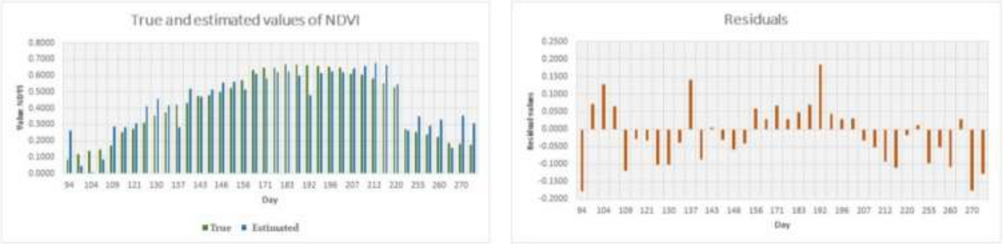

Figure 10. *True and estimated values of NDVI (left) and residual values (right)*

4. DISCUSSION

After initial acquisition of NDVI values, based on BoxPlots it was concluded that the data contain large amount of outliers that come from different sources and all of them were eliminated with Z-score method. Based on the ScatterPlot of mean values by culture, large oscillations were noticed throughout the year. This can be justified by the fact that climate conditions change during the year and they have an impact on plant development: drought, precipitation and agro-technical measures such as watering and fertilization after which the values of NDVI can improve or deteriorate (Figure 26). This fact was also confirmed by looking at the ScatterPlot of residuals. After the elimination of the outliers, a mathematical model was defined in order to get the values of the index for each date in the time period of plant development. Due to the nonlinearity of the data, a polynomial function of 4th order was chosen as it was the one that fit the data in the most adequate way. Resulting data was further used as dependent variables in the multivariable regression.

Meteorological data (temperature, air pressure, humidity, insolation, cloudiness and precipitation) were taken from the website of Republic Hydrometeorological Service of Serbia for the area of interest. A test for multicollinearity was performed before the model development in order to check if any factors have the same impact. Result of this test showed that insolation impacts the model in the same way as cloudiness, which was expected since the two phenomena are directly connected, although VIF factor was less than the limit which resulted in keeping it in the analysis.

The data was split in 80% for training and 20% for verification of the multivariable regression model. Model analysis showed that the impact of coefficients of single climate parameters are very low, which is expected since the values of index are between 0.2 and 0.8. Heteroscedascity and autocorrelation tests confirmed that the data is homoscedastic with high autocorrelation. Again, this result was expected since there is a well-known trend in the meteorological data, where it is possible to predict future values based on past ones. Mean value of the residuals was zero, and residuals are normally distributed. Initial value of R^2 test gave a very low value (0.56 for corn) since the observed climate parameters have no impact on the index in real time. By observing the R^2 values for different time domains, it was determined that the changes in climate parameters will be noticed after 6 days in the case of corn (it varies from 3 to 7 days for other crops). Analysis of residuals also confirms that the model predicts the values of NDVI on a very satisfactory level. If the more detailed data about agro technical measures was available (crop yield for past years), this model could be used to predict yield in the early stages of planting the crops.

5. CONCLUSION

Cloud platforms for data processing allow fast processing of large amounts of data which makes it easy to perform certain analysis and discover trends in data. Time series of satellite images would be almost impossible to process with classic methods. Statistical methods used in this paper could be used as steps of preprocessing this data. One of the biggest sources of errors in the process of image classifications is the input data (known locations of classes that are identified and contain outliers). Input data (true data) is usually blindly taken as correct when in fact this paper shows that sometimes outliers can be found within.

Multivariable regression offers methods to determine the impact of independent variables on a dependent variable as well as to generate a model that can estimate the values of dependent variable based on the independent variable values. This paper tests the impact of climate parameters (temperature, air pressure, humidity, cloudiness, insolation and precipitation) on several crops during its development in 2017. Based on the model, there are several conclusions: the impact of climate parameters, examined in this paper, will be best noticed after 3 to 6 days; climate factors precipitation and air pressure do not have a significant impact on changes of crops plant state; if model gets real time data, it will estimate NDVI values for the culture in 3 to 7 days from the day of observation, depending on the crop type.

Future work requires the expansion of independent parameters that are determined by the application of agricultural measures (fertilization, watering, usage of herbicide and pesticide etc.) as well as the values of yield for several years back in order to predict yield for the current year in the early stages of plant development.

LITERATURE

[1] J. Van Wart, P. Grassini, K. G. Cassman, “Impact of derived global weather data on simulated crop yields”, Global change biology, vol. 19, pp. 3822–3834, Sep. 2013.

[2] M. Nedealcov, “The Impact of Weather and Climate Risks on Cereal Crops Productivity”, Present Environment and Sustainable Development, vol. 8, pp. 195-208, Aug. 2014.

[3] A. Ceglar, A. Toreti, R. Lecerf, M. Van der Velde, F. Dentener, “Impact of meteorological drivers on regional inter-annual crop yield variability in France”, Agricultural and Forest Meteorology, vol. 216, pp. 58–67, Jan. 2016.

[4] Douglas C. Montgomery and George C. Runger, “Applied Statistics and Probability for Engineers, Third edition, United States of America, Publisher: John Wiley & Sons, Inc., ISBN 0-471-20454-4, 2003.

[5] Towards Data Science, Ways to Detect and Remove the Outliers, Available: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba, [25.08.2019].

[6] Investopedia, Z-test, Available: https://www.investopedia.com/terms/z/z-test.asp, [25.08.2019].

[7] Das Panchanan, “Econometrics in Theory and Practice”. 10.1007/978-981-32-9019-8_8, pp.207-243, 2019.

[8] W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. Peter He, JW Jr. Lillard, “A Study of Effects of MultiCollinearity in the Multivariable Analysis”, Int J Appl Sci Technol, vol. 4. pp. 9–19, Oct. 2014.

[9] Investopedia, Multicollinearity, Available: https://www.investopedia.com/terms/m/multicollinearity.asp [25.08.2019.].

[10] Cecil Robinson, Randall E. Schumacker, “Interaction Effects: Centering, Variance Inflation Factor, and Interpretation Issues”, Multiple Linear Regression Viewpoints, Vol. 35, Jan. 2009.

[11] E. Peksova-Szolgayova, Z. Lukac, P. Roncak, J. Szolgay, “Considering heteroscedastic in the modelling and forecasting of time series of mean daily discharges of the Hron river at station Brehy in Slovakia”, 18th International Multidisciplinary Scientific GeoConference SGEM 2018, vol.18, pp. 151-158, July. 2018.

[12] G. E. P. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, USA: Holden-Day, Inc. 1976.

[13] T. Chai, R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? –

[14] Arguments against avoiding RMSE in the literature”., Geoscientific Model Development, vol. 7, pp. 1247–1250 , Jun. 2014.

[15] NDVI – normalized difference vegetation index, Available: https://gisgeography.com/ndvi-normalized-difference-vegetation-index/ [26.08.2019.].

[16] NDVI Available: https://eos.com/ndvi/ [26.08.2019.].

[17] Republic Hydrometeorological Service of Serbia, Available: http://www.hidmet.gov.rs/index.php [26.08.2019.].