USING SATELLITE IMAGERY AND SIMPLE LINEAR REGRESSION TO PREDICT MAIZE YIELD FOR FOOD SECURITY

Shamim Rashid	I63/4255/2019
Victorine Imbuhila	I63/4262/2019
Olweny Lynn Nicole Adhiambo	I63/137207/2019
Justus Oriya	I63/136922/2019
Gabriel Hamgera Obuwa	I63/4234/2019

SUPERVISOR: Dr. Musiga

UNIVERSITY OF NAIROBI

DEPARTMENT OF MATHEMATICS

June 4, 2023

Overview I

Overview II

Introduction and Statement of the Problem

Background of the Study

- Drought and other severe climatic conditions contribute to food insecurity in Africa.
- The region has experienced reduced agricultural productivity due to environmental destruction, unsustainable land use, and population growth.
- Employment of technology in analyzing land use and how to maximize productivity is key to realization of agricultural produce in the region.

Statement of the Problem

- Problem Statement is determination of the relationship between Normalized Difference Vegetation Index and maize yield in Kenya. NDVI is used to determine density of vegetation health.

Objectives of the Study and Methodology

General Objective

- Prediction of maize yield using satellite imagery and simple linear regression method.

Specific Objective

- To perform exploratory data analysis to identify patterns, trends, and relationships between maize yield and NDVI obtained from satellite images.
- Fit and evaluate the fit of the models.
- Use the models for maize yield prediction.

Methodology

- Simple linear regression model was used
- R software was used for analysis.
- Data was obtained from Ministry of Agriculture, Kilimo House and Kenya Space Agency.

Assumptions and Justification of the Study

Assumptions

- Error term ε has zero mean.
- Error term ε has constant variance σ^2 .
- Errors are uncorrelated.
- Errors are normally distributed.

Justification of the Study

- The study will enable the government and non-governmental organizations(NGOs) to predict maize yield for food security.

Literature Review

Ines et al. (2013)

- Developed data assimilation crop modeling framework that improved the prediction of crop yields in the USA from 2003 to 2009
- Used the Ensemble Kalman filter(EnKF), a Monte Carlo application.
- Results: Assimilation of soil moisture and LAI would be more suitable under normal conditions.

Kasampalis et al. (2018)

- Reviewed contribution of remote sensing data to crop growth models.
- Used regression models to estimate crop yield
- Results: Remote sensing provided accurate timely estimation of land resources.

Sayago & Bocco (2018)

- Conducted study on use of satellite images for crop yield estimation in Argentina.
- Used Multiple Linear regression & Neural Networks for estimating crop yields using satellite data
- Results: NN outperformed MLR. NN provided accurate & reliable estimates of crop yield.

Literature Review

Torre et al.(2021)

- Reviewed use of various models to estimate yield of rice.
- Used both crop models and statistical methods such as the multiple linear regression and logistic regression.
- Results: Remote sensing-based estimation of rice yields has advanced in recent years, but limitations such as low quality of satellite data, inadequate sample sizes, and the need for more robust models still exist.

Ali et al. (2022)

- Reviewed crop yield prediction using Multi sensors remote sensing.
- They identified different types of sensors used to estimate crop yield e.g satellite imagery, RADAR & LIDAR.
- Results: remote sensing could be employed in all aspects of agricultural process from land preparation to harvesting. Satellite data is the most efficient.

Roznik et al. (2022)

- Investigated how crop yield estimation could be improved by applying higher resolution satellite imagery in the USA from 2008 to 2018.
- Used Cubic Spline Regression Model to estimate crop yield using NDVI
- Results: Accuracy of crop yield estimation models could be moderately improved using higher satellite resolution NDVI.

Simple Linear Regression Model

- Simple linear regression is a method for modelling the relationship between two variables, where one variable (the independent variable), is used to predict the value of the other variable (the dependent variable).
- We denote it using the equation:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

- With the mean and variance:

$$E(y|x) = \beta_0 + \beta_1 x \tag{2}$$

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{3}$$

Estimation of Model Parameters

- The method of least squares is used to estimate β_0 and β_1.
- The least-squares estimators of β_0 and β_1 are obtained by minimizing the total squared prediction errors, thus

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{4}$$

and

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{5}$$

- The solutions to the normal equations are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{[\sum_{i=1}^{n} y_i][\sum_{i=1}^{n} x_i]}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{[\sum_{i=1}^{n} x_i]^2}{n}} \tag{7}$$

Estimation of Model Parameters

- The fitted simple linear regression model that gives a point estimate of the mean of y for a particular x. is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{8}$$

- Mathematically the i^{th} residual is

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), i = 1, 2, ..., n \tag{9}$$

Hypothesis Testing and Coefficient of Determination

- The hypotheses are:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

- The t-test statistic for the hypothesis test is:

$$t_0 = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} \tag{10}$$

- where,
  - $\hat{\beta}_1$ is the estimate of the slope obtained from fitting the regression model
  - $s.e(\hat{\beta}_1)$ is the estimated standard error or the standard error of $\hat{\beta}_1$.
- Coefficient of determination, R^2 is a summary of the strength of the relationship between the x_i and the y_i in the data.

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_{Res}}{SS_T} \tag{11}$$

- $0 \leq R^2 \leq 1$

Data for the Study

- The table below contains maize yield for the five counties which are considered the bread-basket of the country.

Table 1: Table showing maize yield for the five counties from 2018-2022

County	Yield/Ha(2018)	Yield/Ha(2019)	Yield/Ha(2020)	Yield/Ha(2021)	Yield/Ha(2022)
Bungoma	3.16	3	3.66	2.89	2.98
Nakuru	1.14	3	3.19	2.77	3.22
Nandi	1.01	3	2.80	2.39	1.96
Trans-Nzoia	5.09	4	4.60	3.90	3.57
Uasin Gishu	4.26	3	3.07	3.68	3.50

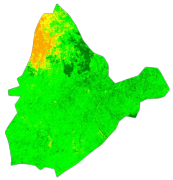
Data for the Study

- The table below contains mean NDVI for the five counties obtained from satellite images.

Table 2: NDVI for the five counties from 2018-2022

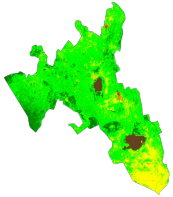
County	NDVI 2018	NDVI 2019	NDVI 2020	NDVI 2021	NDVI 2022
Bungoma	0.56842	0.356925	0.480692	0.353525	0.297571
Nakuru	0.3033913	0.34985	0.3917917	0.39055	0.357579
Nandi	0.2955	0.331462	0.285267	0.206318	0.2223
Trans Nzoia	0.71	0.633	0.67	0.6136	0.4815
Uasin Gishu	0.657882353	0.4365	0.452947	0.384	0.364

Application of Model, Results and Discussions- Satellite Images



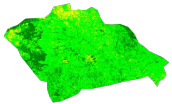
(a) NDVI Map for Bungoma County

(b) NDVI Map for Uasin Gishu County



(c) NDVI Map for Nakuru County

(d) NDVI Map for Nandi County



(e) NDVI Map for TransNzoia County

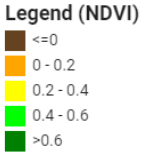


Figure 1: NDVI Maps for the 5 Counties

Application of Model, Results and Discussions

Exploratory Data Analysis

Simple Linear Regression Models quantify relationships between response and predictor variables using a straight line. The following plots were obtained using the average NDVI and Yield/Ha per year.

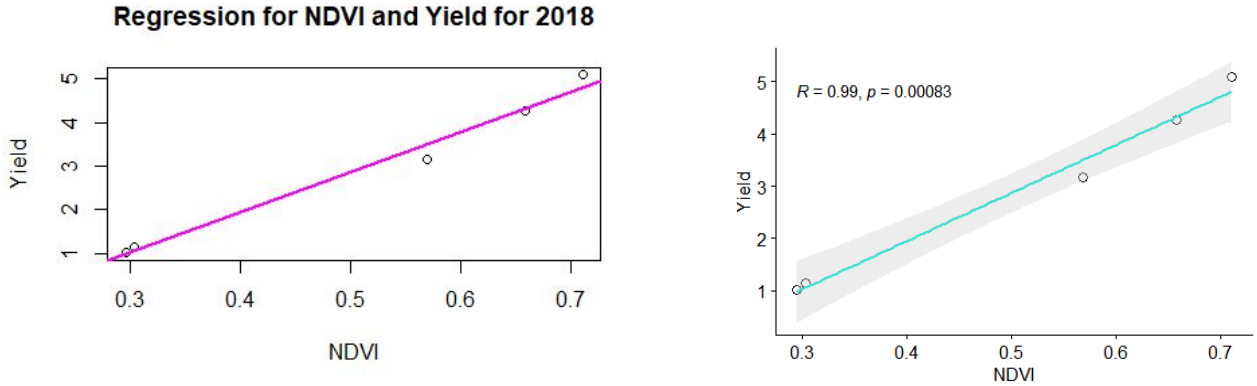


Figure 2: Regression and Correlation plots for NDVI and Yield/Ha for 2018

Application of Model, Results and Discussions

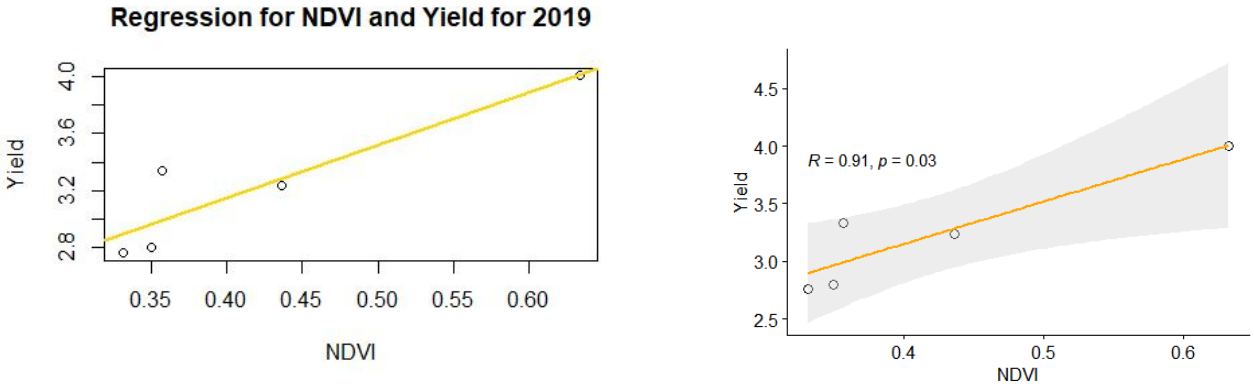


Figure 3: Regression and Correlation plots for NDVI and Yield/Ha for 2019

Application of Model, Results and Discussions

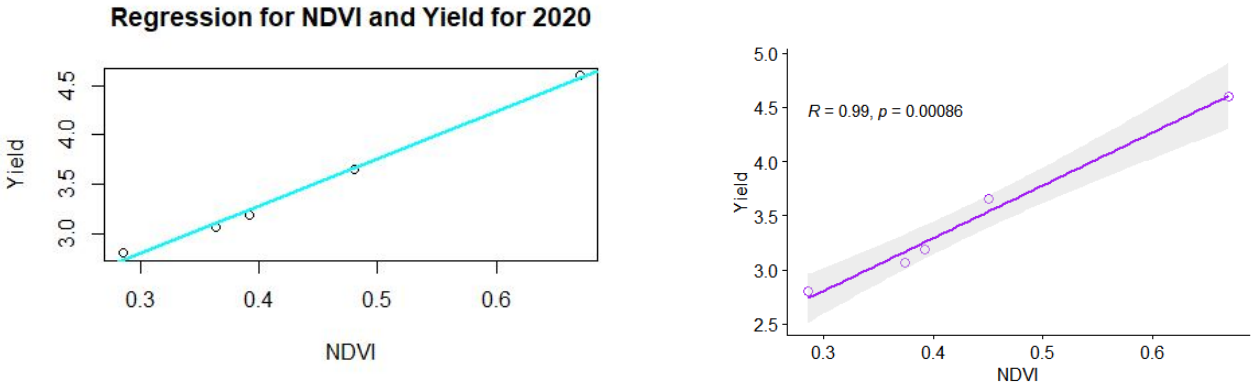


Figure 4: Regression and Correlation plots for NDVI and Yield/Ha for 2020

Application of Model, Results and Discussions

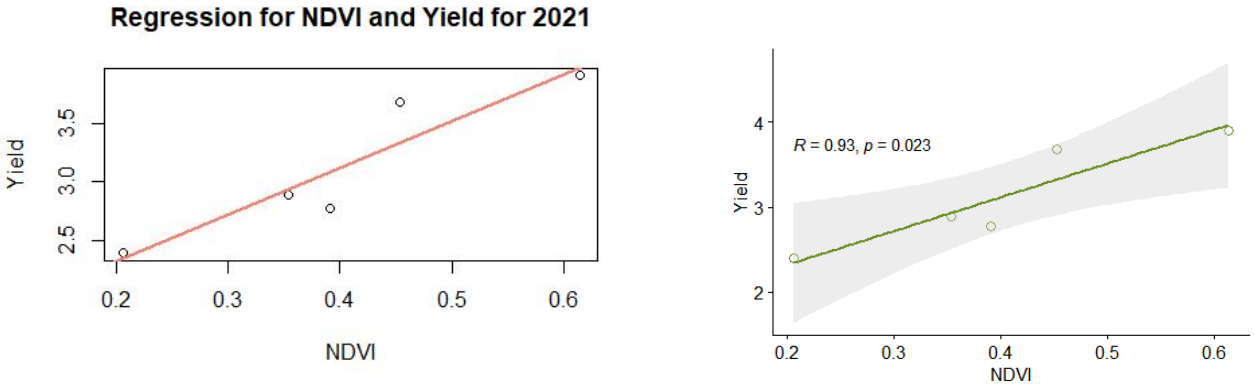


Figure 5: Regression and Correlation plots for NDVI and Yield/Ha for 2021

Application of Model, Results and Discussions

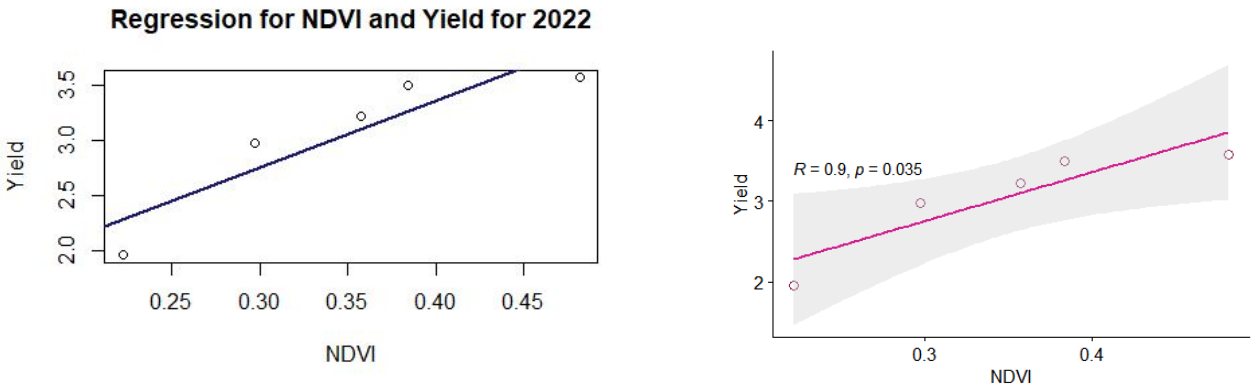


Figure 6: Regression and Correlation plots for NDVI and Yield/Ha for 2022

Estimation of Model Parameters and Test of Fit

- A simple linear regression model is fit when there is a positive or negative linear relationship between the response and predictor variables.
- There is a positive linear relationship between NDVI and yield for the years 2018-2022
- To check the significance of the model we compare the test statistics to the critical values.
  - Critical value for t-test $t_{\alpha/2,n-2}$ is 3.182
  - Critical value for F-test $F_{\alpha,1,n-2}$ is 10.128

2018 model

```
                    Residuals:
     Min       Q1       Median       Q3       Max
-0.33931  0.08978   0.03646    0.28173   -0.06866

                    Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept)  -1.7558   0.3606    -4.869   0.016551 *
2018NDVI      9.2465   0.6721    13.758   0.000831 ***
                    ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Residual standard error: 0.2637 on 3 degrees of freedom
   Multiple R-squared:  0.9844,Adjusted R-squared:  0.9792
    F-statistic: 189.3 on 1 and 3 DF,  p-value: 0.0008311
```

The y intercept is $\beta_0 = -1.7558$ and the NDVI coefficient is $\beta_1 = 9.2465$, therefore the model for the year 2018 is **Yield=−1.7558+9.2465NDVI**

2019 model

```
Residuals:
     Min       1Q   Median       3Q      Max
0.345235 -0.164705 -0.133575 -0.004225 -0.042729

Coefficients:
            Estimate   Std. Error t value  Pr(>|t|)
(Intercept) 1.6770     0.4108     4.083    0.0265 *
2019NDVI    3.6809     0.9419     3.908    0.0298 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.2352 on 3 degrees of freedom
Multiple R-squared:  0.8358,Adjusted R-squared:  0.7811
F-statistic: 15.27 on 1 and 3 DF,  p-value: 0.02976
```

The y intercept is $\beta_0 = 1.6770$ and the NDVI coefficient is $\beta_1 = 3.6809$, therefore the model for the year 2019 is **Yield=1.6770+3.6809NDVI**

Application of Model, Results and Discussions

2020 model

```
                    Residuals:
      Min       1Q       Median     3Q        Max
-0.008572 -0.049288  0.071017  0.025969 -0.039127

                    Coefficients:
             Estimate   Std.Error t value Pr(>|t|)
(Intercept)   1.36007    0.08864   15.34 0.000601 ***
2020NDVI      4.79668    0.19369   24.77 0.000144 ***
                   ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.05701 on 3 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9935
F-statistic: 613.3 on 1 and 3 DF,  p-value: 0.0001444
```

The y intercept is $\beta_0 = 1.36007$ and the NDVI coefficient is $\beta_1 = 4.79668$, therefore the model for the year 2020 is **Yield=1.36007+4.79668NDVI**

Application of Model, Results and Discussions

2021 model

```
                    Residuals:
     Min      1Q      Median      3Q      Max
-0.03809 -0.30599  0.05080 -0.06443  0.35771

                    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.5187     0.3959   3.836   0.0312 *
2021NDVI       3.9884     0.9322   4.278   0.0235 *
                   ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.2767 on 3 degrees of freedom
Multiple R-squared:  0.8592,Adjusted R-squared:  0.8122
F-statistic:  18.3 on 1 and 3 DF,  p-value: 0.02345
```

The y intercept is $\beta_0 = 1.5187$ and the NDVI coefficient is $\beta_1 = 3.9884$, therefore the model for the year 2021 is **Yield=1.5187+3.9884NDVI**

Application of Model, Results and Discussions

2022 model

```
                    Residuals:
     Min      1Q  Median      3Q     Max
 0.2424  0.1222 -0.3198 -0.2824  0.2376

                    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.9250     0.5978   1.547   0.2196
2022NDVI       6.0872     1.6644   3.657   0.0353 *
                   ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Residual standard error: 0.3225 on 3 degrees of freedom
   Multiple R-squared:  0.8168,Adjusted R-squared:  0.7557
     F-statistic: 13.38 on 1 and 3 DF,  p-value: 0.03531
```

The y intercept is $\beta_0 = 0.9250$ and the NDVI coefficient is $\beta_1 = 6.0872$, therefore the model for the year 2022 is **Yield=0.9250+6.0872NDVI**

Predictions of New Observation

- The table below are maize yield values predicted for the years 2023-2027:

County	2023	2024	2025	2026	2027
Bungoma	3.5000757	2.990854	3.665791	2.928679	2.736324
Nakuru	1.0494881	2.964811	3.239365	3.076349	3.101603
Nandi	0.9765212	2.897126	2.728401	2.341557	2.278136
TransNzoia	4.8091950	4.007068	4.573840	3.965963	3.855933
UasinGishu	4.3272893	3.283764	3.104230	3.325215	3.262433

- In order to check the accuracy of the prediction of yield, the R^2 value is calculated:

```
[1] 0.7102979
```

- The above output indicates that observed and predicted values are highly correlated.

Application of Model, Results and Discussions

Discussions

- Exploratory data analysis revealed a positive linear relationship between the Normalized Difference Vegetation Index (NDVI) and Yield/Ha for the years 2018-2022
- Correlation plots indicated a strong correlation between NDVI and yield for each year, with Pearson correlation coefficients ranging from 0.90 to 0.99.
- The Simple Linear Regression equations indicated that a 0.1 increase in NDVI resulted in a corresponding increase in yield.
- The t-test statistics and associated p-values indicated that the NDVI coefficients were statistically significant for all years.
- The multiple R-squared values,were high for all years suggesting that NDVI can explain a significant portion of the variability in yield.
- Overall, the analysis consistently demonstrated a positive linear relationship between NDVI and yield for the years 2018-2022

Summary, Conclusions and Recommendation

Summary

- The theory of Simple Linear Regression Model was used for analysis of maize yield and NDVI obtained from satellite images.
- R software was used to fit the models.
- Predictions were made for the next 5 years

Conclusions

- For each year that was analyzed, a linear relationship was obtained between NDVI and yield.
- Low NDVI resulted in low yield and high NDVI resulted in high yield

Recommendations

- Use satellite imagery techniques to analyze how land topology affects vegetation health and growth in Kenya.
- Use satellite imagery techniques in precision agriculture and pest control.

References

[1] Ali A.M., Abouelghar M., Belal A.A., Saleh N., Yones M., Selim A.I, Amin M.E.S., Elwesemy A., Kucher D.E., Maginan S. and Savin I., Crop Yield Prediction Using Multi Sensors Remote Sensing, *The Egyptian Journal of Remote Sensing and Space Sciences*, 25, 711-716, 2022.

[2] Ines A.V.M., Das N.N., Hansen J.W. and Njoku E.G., Assimilation of Remotely Sensed Soil Moisture and Vegetation with a Crop Simulation Model for Maize Yield Prediction, *Remote Sensing of Environment*, 138, 149-164, 2013.

[3] Kasampalis D.A., Alexandridid T.K., Deva C., Challinor A., Moshou D. and Zalidis G., Contribution of Remote Sensing on Crop Models, *Journal of Imaging*, 52, 1-19, 2018.

[4] Montgomery D.C., Peck E.A. and Vining G.G., Introduction to Linear Regression Analysis, John Wiley & Sons, 2012.

[5] Roznik M., Boyd M. and Porth L., Improving Crop Yield Estimation by Applying Higher Resolution Satellite NDVI Imagery and High-resolution Cropland Masks, *Remote Sensing Applications: Society and Environment*,25, 100-693, 2022.

[6] Sayago S. and Bocco M., Crop Yield Estimation Using Satellite Images: Comparison of Linear and Non-Linear Models, *Agriscientia*, 35, 1-9, 2018.

[7] Torre D.M.G.D., Gao J. and Macinnis-Ng C., Remote Sensing-based Estimation of Rice Yields using Various Models, *Geo-Spatial Information Science*, 24:4, 580-603, 2021.

[8] Weisberg S., Applied linear regression, John Wiley & Sons, 2014.

THANK YOU!