Data Analytics for Environmental Science

Wiredu, Charles | Shepherd, Emily | McClure, Kaitlyn

2024-04-30

# Introduction

The survival of humanity has high dependency on the existence of natural resources.There has however resulted in the need to improve the management of these resources as well as strengthen the knowledge base on the interactions between humans and the earth through resources available to them.The understanding of these interactions and variations of vegetation (NDVI) and land cover therefore becomes very essential source of information because it tend to highlight more spatial activities in the environment. The aim of this project is to apply Key data analysis techniques on the cleaned data set which have the same coordination systems as the land cover/land use (MCD12Q1) data. The project is structured as follows; Data loading, Exploratory Data Analysis (EDA), Data Modeling, results discussion and conclusion.

# Data Loading

The required dataset for this project labelled “Landcover\_NDVI\_Rainfall\_filter” and “NDVI\_timeseries\_Month” will be loaded into R to initiate all the necessary steps mentioned above. The MCD12Q1 dataset is part of the MODIS Land Cover Type product, which provides global land cover classification at a spatial resolution of 500 meters. Each pixel in the dataset represents a land cover class based on a classification scheme. MCD12Q1 has been Stage 2 Validated based on cross-validation of the training dataset used to create the maps.

# Exploratory Data Analysis (EDA)

To better understand the data set, we performed various spatial mappings, which helped us to better understand the data. In the spatial visualization of rainfall, we could see that the amount of rainfall in southern Indiana is higher than in the northern part. When we looked at the map of the landcover, we observed that the southern part of the state was covered with forest, and the northern part of the state had more grasslands. These results matched what we understood about the geography of Indiana. The Hoosier National Forest is in the southern part of the state, and northern Indiana is farmland. Lastly, we created mappings of rainfall by season. The amount of rainfall is less throughout Indiana in Fall 2019 and Winter 2019. In Spring 2019, rainfall is heavier in all parts of the state, but southern Indiana received more rain than the northern part of the state. The largest amount of rainfall fell in the northwest corner of Indiana during Summer of 2019. This can be observed in appendix Fig1 and Fig2.

# Data Modeling/ Results Discussion

For the purpose this project the following modeling techniques will be employed on both dataset to gain more insight into the landcover classifications, Vegetation index and Precipitation trends.

* Anova Analysis A two Anova analysis is conducted on NDVI with month and land cover types using data in 2019.The model lists the independent variables being tested in the model which included month and landcover. Again, the study conducted normality check on the residuals of the Anova model(two-way). From the histogram and QQ-plot above, we can observe that the normality assumption seems to be met. The model from this analysis suggests that month and landcover are statistically significant with p-values less than 0.001. Again, there is a suggestion from the normality check that shows the histogram roughly forming a bell curve, indicating that the residuals follow a normal distribution. Furthermore, points in the QQ-plots roughly follow the straight line and most of them are within the confidence bands, also indicating that residuals follow approximately a normal distribution as shown in appendix Fig3.
* *Model: two\_way <- aov(MeanNDVI ~ Month + Landcover, data= Landcover\_NDVI\_Rainfall\_filter)*
* Regression Analysis

Regression analysis (ordinary linear regression) is further conducted on the landcover\_NDVI\_Rainfall\_filter data set on NDVI with precipitation. The linear regression model allows us to determine the value of one response variable based on the value of the given predictor variable. The model report on residuals and coefficients. The model created to predict NDVI from rainfall suggests that the p-value for the model is less than 0.05, therefore suggesting that there is a statistically significant relationship between rain and NDVI.

* *model: NDVI= 0.0004597(rain) + 0.4611341*
* Time series Analysis

Time series as a statistical technique allows for analysis of a time series data or trend with the goal of extracting meaningful statistics. For this project time series analysis was conducted with NDVI time series from 2013-2022 over croplands. After the first exploration of the time series data, we concluded it was stationary. We compared the AIC and BIC for four different models: Seasonal ARIMA, Auto ARIMA, MSTL (), and STLM(). The model created using MSTL had the lowest absolute value of AIC and BIC, and therefore, fit the data the best. To verify this graphically, we plotted the residuals, which were close to the line passing through the observed and predicted NDVI’s. This can be observed in appendix Fig 4- Fig7.

# Conclusion

The findings in this project suggest that vegetation index changes throughout the year, with a decline in NVDI in the last two quarters of the year. Visual plots as projected in this study suggest the estimated trend showing a small decrease in 2018-19, which is followed by a steady increase from 2021 and rapid fall into 2022.This project further utilizes findings from OLR, ANOVA and Time series which enhanced our understanding on working on environmental science data set.

Appendix

Fig 1. Landcover Classification

A green and blue map

Description automatically generated with medium confidence

Fig 2. mappings of rainfall by season

A graph of a fall in indiana

Description automatically generatedA graph of a fall

Description automatically generated

A graph of a number of colored squares

Description automatically generated with medium confidenceA graph of a fall in indiana

Description automatically generated

Fig3. Normality check on Anova(two-way Model)

A graph and diagram of a graph

Description automatically generated

Fig4 . Time series of the original data

A graph of a graph showing the number of years

Description automatically generatedA graph of a graph

Description automatically generated

Fig5. Decomposition of time series

A graph of different types of waves

Description automatically generated with medium confidenceA graph of a graph of a graph

Description automatically generated with medium confidence

Fig6. Accuracy Comparison of models

A screenshot of a graph

Description automatically generated

Fig7. AIC and BIC values comparison of four models

A white background with black lines

Description automatically generated