Jurnal Ilmu Fisika
Vol. 15, No. 2, September 2023, pp. 146–165
ISSN: 1979-4657 (Print); 2614-7386 (Online)
https://doi.org/10.25077/jif.15.2.146-165.2023

Jurnal Ilmu Fisika
Open Access

Relationship between Solar Flux and Sunspot Activity Using Several Regression Models

[1]Ruben Cornelius Siagian, [2]Lulut Alfaris, [3]Ghulab Nabi Ahmad, [4]Nazish Laeiq, [5]Aldi Cahya Muhammad, [6]Ukta Indra Nyuswantoro, [1]Budiman Nasution

[1]Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Medan, 20221,Indonesia,
[2]Department of Marine Technology, Politeknik Kelautan dan Perikanan, Pangandaran, 46396, Indonesia,
[3]Institute of Applied Sciences, Mangalayatan University, Aligarh, 202145, India,
[4]Department of Computer Science, Institute of Technology and Management Aligarh, 202140, India,
[5]Department of Electrical and Electronic Engineering, Islamic University of Technology, Kustia, 7003, Bangladesh
[6]Department of Structure Engineering, Asiatek Energi Mitratama, Jakarta, 12870, Indonesia.

Article Info

Article History:

Received March 28, 2023
Revised May 06, 2023
Accepted May 18, 2023
Published online June 06, 2023

Keywords:

Solar Flux
Sunspot Activity
Regression Analyis
Linear regression
SARIMA model analysis

Corresponding Author:

Ruben Cornelius Siagian,
Email: rubensiagian775@gmail.com

ABSTRACT

This study examines the correlation and prediction between sunspots and solar flux, two closely related factors associated with solar activity, covering the period from 2005 to 2022. The study utilizes a combination of linear regression analysis and the ARIMA prediction method to analyze the relationship between these factors and forecast their values. The analysis results reveal a significant positive correlation between sunspots and solar flux. Additionally, the ARIMA prediction method suggests that the SARIMA model can effectively forecast the values of both sunspots and solar flux for a 12-period timeframe. However, it is essential to note that this study solely focuses on correlation analysis and does not establish a causal relationship. Nonetheless, the findings contribute valuable insights into future variations in solar flux and sunspot numbers, thereby aiding scientists in comprehending and predicting solar activity's potential impact on Earth. The study recommends further research to explore additional factors that may influence the relationship between sunspots and solar flux, extend the research period to enhance the accuracy of solar activity predictions and investigate alternative prediction methods to improve the precision of forecasts.

1. INTRODUCTION

The Sun is critical in sustaining life on Earth as the primary energy source. It resides at the center of our solar system, with all the planets, including Earth, orbiting around it. As a massive ball of hot plasma, the Sun generates light and heat through nuclear fusion processes (Reinders, 2021). Despite being approximately 93 million miles from Earth, its influence is felt in every aspect of our lives. The Sun's energy provides light and heat, essential for supporting life on our planet. Without the Sun, life as we know it would not be possible. It serves as the primary energy source for all living organisms, with plants utilizing its energy through photosynthesis to produce food for other creatures (Khoo et al., 2023).

Moreover, the Sun's heat is vital in regulating Earth's climate and weather patterns. Studying the Sun's behavior and changes has always captivated human interest and has been the subject of scientific investigation. Monitoring solar flux and sunspot activity has been an ongoing process for centuries. Solar flux measures the amount of energy emitted by the Sun, while sunspots are dark spots on its surface that indicate areas of intense magnetic activity (Kaplan, 2023). Analyzing these phenomena provides valuable insights into the Sun's behavior and changes, with implications for space weather forecasting, climate prediction, and renewable energy production (Mandea & Chambodut, 2020).

Space weather forecasting involves predicting the Sun's behavior and its effects on Earth's atmosphere and magnetic field (Siddique et al., 2022). This information is crucial for safeguarding satellites, communication systems, and power grids from the detrimental impacts of space weather. Additionally, understanding the Sun's influence on Earth's climate and weather patterns is vital for comprehending and mitigating the effects of climate change (Zhang et al., 2023). Renewable energy production, particularly solar power, relies on harnessing the Sun's energy. Understanding the Sun's behavior and changes is crucial for optimizing the efficiency of solar panels and developing innovative renewable energy technologies (Guermoui et al., 2022). Solar flux and sunspot activity are significant measures in solar physics. Solar flux refers to the radiation emitted by the Sun, while sunspots are observed in dark regions on its surface (Lockwood & Ball, 2020).

Solar activity has long interested scientists due to its potential impacts on Earth's climate, communication systems, and power grids (Marov, 2020). Numerous studies have investigated the relationship between sunspots and solar flux and predicted their values for future periods. For instance (Wehrli et al., 2013), utilized linear regression analysis to measure the correlation between sunspots and solar flux, finding a significant positive relationship between the two variables. (Pala & Atici, 2019) employed ARIMA analysis to forecast sunspot values for the next 12 periods. Based on these studies, the present research aims to analyze the correlation between sunspots and solar flux and predict their values for the next 12 periods using linear regression and ARIMA analysis. The prediction results from the SARIMA model will provide predicted values for both data types, along with a 95% confidence interval for the forecasts. This study has important implications for understanding solar activity. It can assist scientists in predicting future solar activity, which is crucial for various applications, including space exploration, satellite communication, and climate forecasting.

This research employed multiple regression analysis techniques to investigate the correlation between sunspots and solar flux. Linear regression was utilized to examine the linear relationship between the two variables.

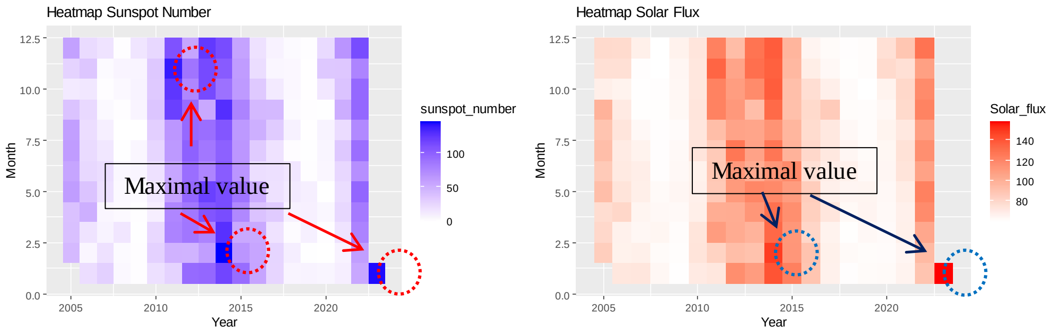


Figure 1. Heatmap visualisation of Solarflux and Sunspot Number (2005 – 2023)

2. METHOD

2.1 Data

This study uses SILSO (Sunspot Index and Long-term Solar Observations) data from https://www.sidc.be/SILSO/home and Solarflux data from https://www.spaceweather.gc.ca/. Figure 1 presents the Solar Flux and Sunspot Number heatmaps, to illustrate the patterns and trends of these phenomena over multiple years and specific months. The Solar Flux heatmap plot shows the years on the *x*-axis and the months on the *y*-axis. Each square in the heatmap is filled with a color representing the Solar Flux value for that particular month and year. Similarly, the Sunspot Number heatmap plot depicts the years on the *x*-axis and the months on the *y*-axis, with each square displaying a color corresponding to the Sunspot Number value for that specific month and year. The shades of red in each square indicate higher solar flux values, providing a visual representation of the patterns and trends of solar flux over multiple years and months. Similarly, the sunspot number heatmap plot follows the same structure as the solar flux heatmap plot, with the color of each square representing the sunspot number value for a specific month and year. The shades of blue in the square indicate higher sunspot number values, enabling observation of the patterns and trends of sunspot numbers over several years and months. These visualizations are valuable in understanding how space weather variability changes over time.

2.2 Processing data

The R programming language will be used as the operating software for data processing, modeling, and comparison (Hao & Ho, 2019). The parameters that determine the best model in this case include:

a. Accuracy

The model should accurately predict the outcome variable. The model's accuracy can be assessed using R-squared, mean squared error, root mean squared error, and mean absolute error (Ahmed, 2023).

b. Simplicity

The model should be simple and easy to interpret. A simpler model is generally preferred over a more complex one as it is easier to understand and explain (Gosiewska et al., 2021).

c. Generalizability

The model should generalize to new data not used during training. This can be evaluated using techniques such as cross-validation (Berrar, 2019).

d. Robustness

The model should be robust to outliers and noise in the data. It is crucial to examine how well the model performs on data different from the training data (Seliya et al., 2021).

e. Computational efficiency

The model should be computationally efficient and capable of handling large datasets. This is particularly important when dealing with big data (Hariri et al., 2019).

f. Interpretability

The model should be interpretable, allowing insights into the relationship between the predictors and the outcome variable (Susnjak et al., 2022).

g. Model comparison

The results of each model should be compared to determine which model is the best and most suitable for solar flux and sunspot analysis cases (Maldonado-Salguero et al., 2022).

Quality check, or quality assurance, is crucial in analyzing data and ensuring that the built model meets expectations and desired performance criteria. In the case of correlating sunspot and solar flux, here are some ways to perform a quality check on the various regression methods used (Chatzistergos et al., 2021):

a. For a linear regression model, several metrics can be utilized for quality checking, including R-squared, adjusted R-squared, residual plot, normality plot, and significance tests. R-squared and adjusted R-squared measure the model's ability to explain the variance in the data, while residual plots and normality plots assess whether the regression assumptions have been met. Significance tests can determine if the variables used in the model have a significant impact on the dependent variable.
b. Quality checks for a Random Forest model can involve metrics such as mean squared error (MSE), mean absolute error (MAE), R-squared, and feature importance. MSE and MAE measure the difference between predicted and actual values of the dependent variable, while R-squared assesses the model's ability to explain the variance in the data. Feature importance provides insights into the relative contribution of each feature in the model.
c. Quality checks can include metrics like MSE, MAE, R-squared, residual plot, and learning curve for a Neural Network model. MSE, MAE, and R-squared have the same interpretations as in Random Forest. The residual plot checks for patterns in the residuals, while the learning curve helps determine if the model is overfitting or underfitting.

d. A quality check on an SVR model can involve various metrics, including MSE, MAE, R-squared, residual plot, and parameter tuning. Parameter tuning helps select the appropriate kernel and other parameters to optimize the model's performance.

e. For a SARIMA model, quality checks can be performed using metrics such as AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria), residual plot, and forecast accuracy. AIC and BIC are used to select the best model based on information value. The residual plot helps assess whether the model assumptions have been met, while forecast accuracy evaluates the model's ability to predict future values.

2.1 Linear regression analysis

The research method employed in this study is linear regression analysis to evaluate the relationship between sunspot number and solar flux during the period of 2005-2022. The data for this research were obtained from an Excel file using the readxl library in R-Studio. The data used in this study consisted of an independent variable, namely Solar flux, and a dependent variable, namely sunspot number. These data were then input into a data frame using the data.frame function in R-Studio. Hypothesis testing was performed using a linear regression model determined by the lm function in R-Studio. First, the model is defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

where Y is the response variable, X_1, X_2, ..., X_p are the predictor variables, β_0 is the intercept term, β_1, β_2, ..., β_p are the regression coefficients, and ε is the error term. Next, the lm function in R-Studio is used to estimate the values of the regression coefficients β_1, β_2, ..., β_p based on the data.

RSS stands for Residual Sum of Squares. It is a statistical measure used in regression analysis to quantify the discrepancy between the observed values of the dependent variable and the predicted values obtained from a regression model. The goal of regression analysis is to find the best-fitting line or curve that minimizes the difference between the predicted and observed values. The RSS is calculated by summing the squared differences between the observed values (Y) and the predicted values ($\hat{Y}$) obtained from the regression model. Mathematically, it can be expressed as:

$$RSS = \sum (Y - \hat{Y}) \tag{2}$$

where Y represents the observed values of the dependent variable, $\hat{Y}$ represents the predicted values of the dependent variable based on the regression model, and Σ represents the summation symbol *(i.e., adding up the squared differences for each data point)*. The RSS is used as an optimization criterion in regression analysis. The goal is to minimize the RSS, which means finding the regression coefficients that provide the best fit to the data. By minimizing the sum of squared errors, the regression model aims to capture the relationship between the dependent variable and the predictor variables as accurately as possible.

Once the regression coefficients are estimated, hypothesis testing can be performed to determine if there is a significant relationship between the response variable Y and the predictor variables X_1, X_2, …X_p. This is generally done by testing the null hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, , which means there is no significant relationship between Y and each predictor variable.

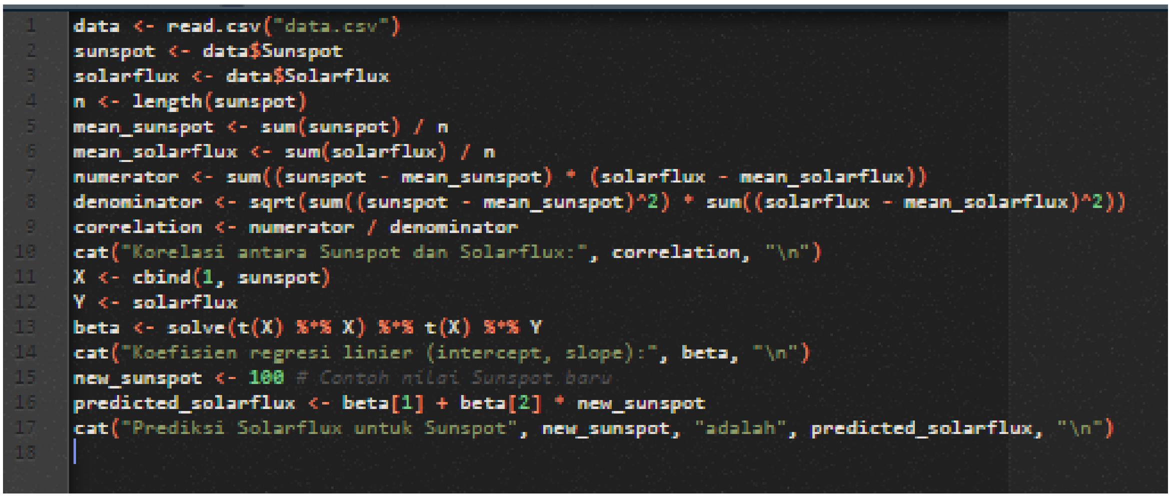
The F test statistic is a statistical measure used to test hypotheses in regression analysis. It helps determine the significance of the relationship between the predictor variables and the response variable in a linear regression model. The formula for calculating the F test statistic is as follows:

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{(n-p-1)}} \tag{3}$$

where *SSR* is the regression sum of squares, *SSE* is the residual sum of squares, *p* is the number of predictor variables, *n* is the sample size, and *p+1* is the total number of coefficients in the model.

The F test statistic compares the variability explained by the regression model (SSR) to the unexplained variability (SSE) and takes into account the degrees of freedom associated with each component. By comparing the calculated *F* value to a critical *F* value from the F-distribution with degrees of freedom (p, n - p - 1), you can determine whether the relationship between the predictors and the response variable is statistically significant.

If the value of the F test statistic is greater than the critical value based on the specified degrees of freedom and significance level, then the null hypothesis can be rejected and it can be concluded that there is a significant relationship between Y and at least one predictor variable. Otherwise, the null hypothesis fails to be rejected and it can be concluded that there is no significant relationship between Y and any of the predictor variables. This model serves to determine whether there is a relationship between the independent and dependent variables, as well as the strength of that relationship.



```r
data <- read.csv("data.csv")
sunspot <- data$Sunspot
solarflux <- data$Solarflux
n <- length(sunspot)
mean_sunspot <- sum(sunspot) / n
mean_solarflux <- sum(solarflux) / n
numerator <- sum((sunspot - mean_sunspot) * (solarflux - mean_solarflux))
denominator <- sqrt(sum((sunspot - mean_sunspot)^2) * sum((solarflux - mean_solarflux)^2))
correlation <- numerator / denominator
cat("Korelasi antara Sunspot dan Solarflux:", correlation, "\n")
X <- cbind(1, sunspot)
Y <- solarflux
beta <- solve(t(X) %*% X) %*% t(X) %*% Y
cat("Koefisien regresi linier (intercept, slope):", beta, "\n")
new_sunspot <- 100 # Contoh nilai Sunspot baru
predicted_solarflux <- beta[1] + beta[2] * new_sunspot
cat("Prediksi Solarflux untuk Sunspot", new_sunspot, "adalah", predicted_solarflux, "\n")
```

Figure 2. Sunspot and solarflux correlation linear regression model program 2005-2022 using R-studio

The program shown in Figure 2 is an example of an R program to analyze the relationship between Sunspot and Solarflux using linear regression and correlation. First, the program reads the data from the provided CSV file. The data contains Sunspot and Solarflux columns which represent the number of sunspots and solar radiation flow in a period of time. After reading the data, the program retrieves the Sunspot and Solarflux columns from the data for further processing. The program calculates the average of the Sunspot and Solarflux columns to be used in the correlation calculation. This is done by summing the values in the columns and dividing them by the total amount of data. Next, the program calculates the correlation between Sunspot and Solarflux using the Pearson correlation formula. This calculation involves calculating the difference between each Sunspot value and the average Sunspot, as well as calculating the difference between each Solarflux value and the average Solarflux. After that, the program calculates the correlation as the quotient between the sum of the product of the differences and the square root of the sum of the squares of the differences. This correlation value indicates how strong the linear relationship between Sunspot and Solarflux is, with values ranging from -1 to 1. The program prints the result of the correlation between Sunspot and Solarflux. Next, the program creates a linear regression model between Sunspot and Solarflux. This model uses a simple linear regression equation with one independent variable (Sunspot) and one dependent variable (Solarflux). The program uses the OLS (Ordinary Least Squares) method to estimate the regression coefficients. The program prints the linear regression coefficients (intercept and slope) which show the estimated values for Solarflux values when Sunspot values change. The intercept is the value of Solarflux when the Sunspot value is zero, while the slope is the change in Solarflux when the Sunspot value increases by one unit. Finally, the program provides a prediction example using the created regression model. The program inputs a new Sunspot value (e.g. 100) and calculates the predicted Solarflux value according to the linear regression model. This prediction result reflects the expected Solarflux value based on the given Sunspot value.

2.2 Random Forest Regression

The Random Forest Regression method is a machine learning algorithm that can be used to analyze the correlation between two variables. In this case, the variables of interest are the sunspot number and solar flux in solar spot data. The algorithm works by constructing multiple decision trees based on random subsets of the data. Each decision tree is trained on a different subset of the data, and the final prediction is made by averaging the predictions of all the individual trees. Mathematically, the algorithm can be described as follows:

Given a set of observations $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where x_i represents the sunspot number and yi represents the solar flux, the Random Forest Regression method aims to estimate a function $f(x)$ that maps x to y. To accomplish this, the algorithm first constructs B decision trees T_1, T_2, ..., T_B, each trained on a random subset of the data. Each decision tree T_i is constructed by recursively partitioning the data into smaller and smaller subsets, until the subsets are homogeneous with respect to the response variable y. At each node of the decision tree, the algorithm selects the best split that maximizes the reduction in variance of the response variable. The variance reduction is calculated as follows:

$$\Delta = Var(y) - \{Var(y_{left}) \times w_{left} + Var(y_{right}) \times w_{right}\} \tag{4}$$

In the context of the relationship between solar flux and sunspot data, the effectiveness of decision tree algorithms can be quantified using the variance reduction, denoted as Δ. Δ is calculated by comparing the variance of the response variable in the current node (Var(y)) with the weighted average of the variances in the left and right child nodes (Var(y_{left}) and Var(y_{right})), where the weights of the child nodes (w_{left} and w_{right}) are considered. The formula represents the reduction in variance that would result from splitting the current node into two child nodes based on a specific criterion. The algorithm aims to find the splitting criterion that maximizes Δ, as it indicates the greatest reduction in variance and therefore provides the most informative split. Once the decision trees are constructed, they are used to predict outcomes for new observations. To make a prediction for a new observation x, the algorithm calculates the average of the predictions generated by all the individual trees in the ensemble. This averaging process helps mitigate the impact of biases in individual trees and provides a more reliable prediction.

In the context of the relationship between solar flux and sunspot data, the algorithm constructs B decision trees. After constructing these trees, the algorithm utilizes them to make predictions for new observations. To predict the solar flux value for a given sunspot number, the algorithm averages the predictions of all the individual trees. This can be expressed using the formula:

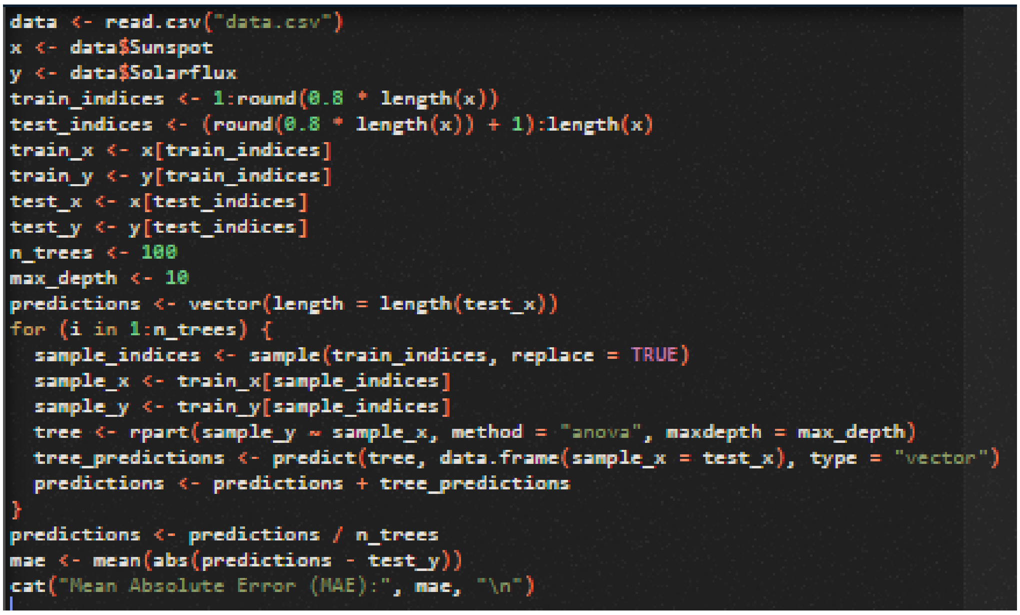
$$f(x) = \left(\frac{1}{B}\right) \times \{T_1(x) + T_2(x) + \cdots + T_B(x)\} \tag{5}$$

where $T_1(x)$, $T_2(x)$, ..., $T_B(x)$ are the predictions of the individual trees for the observation x. By using this method, we can analyze the correlation between sunspot number and solar flux in solar spot data and make predictions about the solar flux values for a given sunspot number. In the formula, f(x) represents the predicted solar flux value for the observation x. T_1(x), $T_2(x)$, ..., $T_B(x)$ are the predictions of the individual trees for the observation x. The algorithm calculates the average of these predictions by summing them up and dividing by the total number of trees B. This approach allows for analyzing the correlation between sunspot number and solar flux in solar spot data and provides predictions for the solar flux values based on a given sunspot number.

The program in Figure 3 is a manual implementation of the Random Forest Regression model to analyze the correlation between sunspot and solarflux from 2005 to 2022. The goal of this program is to predict solarflux values based on existing sunspot values. First, the program reads data from a CSV file containing sunspot and solarflux information. Then, the data is separated into features (sunspot) and targets (solarflux). Features are used to predict targets in this model. Next, the data is divided into training data and testing data. The training data (80% of the data) will be used to train the model, while the testing data (20% of the data) will be used to test the performance of the model. The

Random Forest model building process starts by initializing the number of trees in the ensemble (n_trees) and the maximum depth of each tree (max_depth). At each iteration, the program subsamples with replacement from the training data. This means each tree in the ensemble will see a random subset of the training data. Next, the program builds the i-th tree in the ensemble using the rpart() function. It uses the Recursive Partitioning and Regression Trees (RPART) algorithm with the "anova" method to build the regression tree. This function accepts the feature (sunspot) and target (solarflux) as arguments. Once the i-th tree is built, the program uses the tree to predict solarflux values based on the sunspot values in the test data. The predictions from each tree are added to the overall prediction. Once all the trees in the ensemble are built and predictions are made, the program calculates the average of the predictions from all the trees. This gives the final prediction for the test data. Finally, the program calculates the Mean Absolute Error (MAE), which is an evaluation metric used to measure the extent to which the model predictions differ from the true values on the test data. The lower the MAE, the better the performance of the model.

This program provides a manual implementation of the Random Forest Regression model without the use of library packages. Although the results can provide an initial understanding of the correlation between sunspot and solarflux, it is recommended to use library packages available in R, such as 'randomForest', to obtain more efficient and accurate results.



```r
data <- read.csv("data.csv")
x <- data$Sunspot
y <- data$Solarflux
train_indices <- 1:round(0.8 * length(x))
test_indices <- (round(0.8 * length(x)) + 1):length(x)
train_x <- x[train_indices]
train_y <- y[train_indices]
test_x <- x[test_indices]
test_y <- y[test_indices]
n_trees <- 100
max_depth <- 10
predictions <- vector(length = length(test_x))
for (i in 1:n_trees) {
  sample_indices <- sample(train_indices, replace = TRUE)
  sample_x <- train_x[sample_indices]
  sample_y <- train_y[sample_indices]
  tree <- rpart(sample_y ~ sample_x, method = "anova", maxdepth = max_depth)
  tree_predictions <- predict(tree, data.frame(sample_x = test_x), type = "vector")
  predictions <- predictions + tree_predictions
}
predictions <- predictions / n_trees
mae <- mean(abs(predictions - test_y))
cat("Mean Absolute Error (MAE):", mae, "\n")
```

Figure 3. Random forest regression model program for sunspot correlation and solarflux 2005-2022 using R-studio

2.3 Neural network regression

The neural network method can be used to study the correlation between sunspots and solar flux radiation. The relationship between the two variables can be modeled using a neural network, which consists of input nodes, hidden layers, and output nodes. The input nodes correspond to the sunspot data, and the output nodes correspond to the solar flux radiation data. The hidden layers are used to transform the input data and extract features that are relevant to the output. The mathematical equations used to model the neural network are as follows:

a. Input layer: The input layer takes the sunspot data as input. The sunspot data is denoted by x, and is a vector of length n.

b. Hidden layer: The hidden layer transforms the input data using a set of weights and biases. The transformation is given by $z = f(W_x+b)$ where z is a vector of length m, W_x is a matrix of size m x n, b is a vector of length m, and f is a nonlinear activation function.

c. Output layer: The output layer takes the transformed data from the hidden layer and produces the predicted solar flux radiation data. The transformation is given by $y = g(V_z + c)$ where y is a vector of length p, V_z is a matrix of size p x m, c is a vector of length p, and g is a nonlinear activation function.

d. Training: In the context of the relationship between solar flux and sunspot data, during the training phase of a neural network, the network is optimized by minimizing a cost function that quantifies the discrepancy between the predicted output and the actual output. This cost function can be represented by the formula J =1/2$\|y$-$d\|^2$, where d is the actual solar flux radiation data. In the formula, J represents the cost function, y denotes the predicted solar flux radiation data generated by the neural network, and d represents the actual solar flux radiation data. The difference between the predicted and actual outputs is squared and multiplied by a factor of 1/2 to ensure positive values and facilitate mathematical operations. By minimizing the cost function J, the neural network adjusts its internal parameters to improve the accuracy of its predictions and better capture the relationship between solar flux and sunspot data. The training process iteratively updates the network's parameters until the cost function is minimized, indicating a close match between the predicted and actual solar flux radiation data.

In the context of the relationship between sunspots and solar flux radiation, various methods can be employed to study their correlation. One statistical method that can be used is the Pearson correlation coefficient, which measures the linear relationship between two variables. The formula for the Pearson correlation coefficient, denoted as r, is:

$$r = \frac{cov(x,y)}{\{std(x) \times std(y)\}} \tag{6}$$

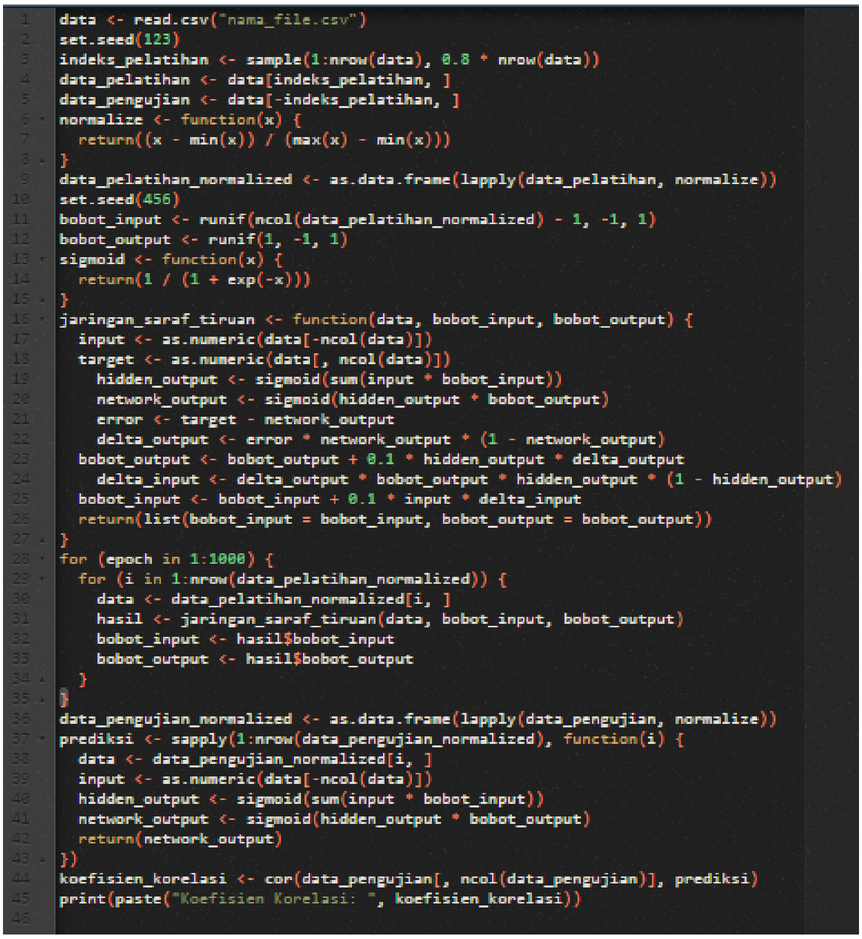
In the formula, r represents the Pearson correlation coefficient. It is calculated by dividing the covariance between the variables x and y by the product of their standard deviations. The covariance, cov(x, y), measures how the two variables vary together, while the standard deviations, std(x) and std(y), quantify the dispersion or spread of each variable. The Pearson correlation coefficient ranges between -1 and 1. A positive value of r indicates a positive linear correlation, meaning that as one variable increases, the other tends to increase as well. Conversely, a negative value of r indicates a negative linear correlation, implying that as one variable increases, the other tends to decrease. A value close to 0 suggests little to no linear correlation between the variables. By calculating the Pearson correlation coefficient, we can assess the strength and direction of the linear relationship between sunspots and solar flux radiation.

The Spearman correlation coefficient, denoted by the symbol ρ (rho), is a measure of the monotonic relationship between two variables. The formula for calculating the Spearman correlation coefficient is as follows:

$$\rho = 1 - \left[\frac{6 \times \sum d_i^2}{\{n \times (n^2 - 1)\}}\right] \tag{7}$$

where d_i is the difference between the ranks of x_i and y_i, and n is the number of data points. The program utilizes the R programming language, and specifically the neuralnet library, to build the neural network model. The sunspot and solar flux data are stored in an Excel file, which is accessed using the readxl library's read_excel function. Year, month, Solar flux, and sunspot number variables are used to store the data from the Excel file into a data frame called "data". To facilitate data processing, the data is normalized using the scale function. The normalized data is then split into training data (train_data) and test data (test_data) at a 70:30 ratio using the sample function.

The program shown in Figure 4 is an example of implementing a neural network to build a correlation model between sunspots and solar flux from data stored in a CSV file. First, the program reads data from a CSV file using the read.csv("name_file.csv") function. This CSV file contains data on sunspots and solar flux from 2005 to 2022. Next, the data is divided into training data and testing data. The training data is 80% and the testing data is 20% of the overall data using sample and [-] indexing. The training data was normalized using the normalize function. This function converts data values into a range between 0 and 1 using a min-max normalization scale. The initial weights for the inputs and outputs are initialized randomly using the runif function. The program defines the sigmoid activation function using the appropriate mathematical formula. The program defines a mock neural network function that accepts data, input weights, and output weights. This function performs computations to obtain the output of the artificial neural network by multiplying the inputs by the weights and applying the sigmoid activation function. In addition, this function also calculates the error and adjusts the weights based on the error using the backpropagation method. Artificial neural network training is performed by running a for loop for a certain number of epochs. In each epoch, the training data is processed one by one using the for loop. Each data is passed to the artificial neural network function to update the input and output weights. The test data is normalized using the normalize function. Prediction of solar flux results is done using the trained artificial neural network. Each test data is processed using a for loop and the prediction results are obtained by feeding inputs to the network and retrieving the outputs. The correlation coefficient between the actual test data and the predicted results is calculated using the cor function. The correlation coefficient results are printed using the print function.

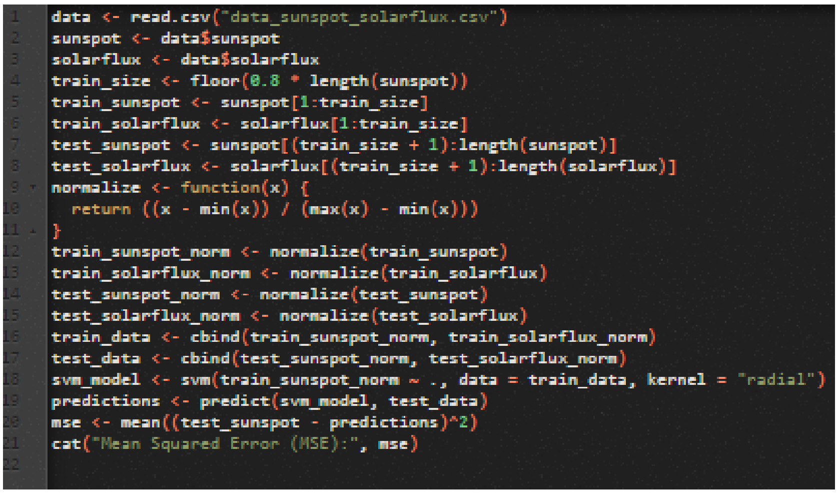


```
data <- read.csv("nama_file.csv")
set.seed(123)
indeks_pelatihan <- sample(1:nrow(data), 0.8 * nrow(data))
data_pelatihan <- data[indeks_pelatihan, ]
data_pengujian <- data[-indeks_pelatihan, ]
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
data_pelatihan_normalized <- as.data.frame(lapply(data_pelatihan, normalize))
set.seed(456)
bobot_input <- runif(ncol(data_pelatihan_normalized) - 1, -1, 1)
bobot_output <- runif(1, -1, 1)
sigmoid <- function(x) {
  return(1 / (1 + exp(-x)))
}
jaringan_saraf_tiruan <- function(data, bobot_input, bobot_output) {
  input <- as.numeric(data[-ncol(data)])
  target <- as.numeric(data[, ncol(data)])
    hidden_output <- sigmoid(sum(input * bobot_input))
    network_output <- sigmoid(hidden_output * bobot_output)
    error <- target - network_output
    delta_output <- error * network_output * (1 - network_output)
  bobot_output <- bobot_output + 0.1 * hidden_output * delta_output
    delta_input <- delta_output * bobot_output * hidden_output * (1 - hidden_output)
  bobot_input <- bobot_input + 0.1 * input * delta_input
  return(list(bobot_input = bobot_input, bobot_output = bobot_output))
}
for (epoch in 1:1000) {
  for (i in 1:nrow(data_pelatihan_normalized)) {
    data <- data_pelatihan_normalized[i, ]
    hasil <- jaringan_saraf_tiruan(data, bobot_input, bobot_output)
    bobot_input <- hasil$bobot_input
    bobot_output <- hasil$bobot_output
  }
}
data_pengujian_normalized <- as.data.frame(lapply(data_pengujian, normalize))
prediksi <- sapply(1:nrow(data_pengujian_normalized), function(i) {
  data <- data_pengujian_normalized[i, ]
  input <- as.numeric(data[-ncol(data)])
  hidden_output <- sigmoid(sum(input * bobot_input))
  network_output <- sigmoid(hidden_output * bobot_output)
  return(network_output)
})
koefisien_korelasi <- cor(data_pengujian[, ncol(data_pengujian)], prediksi)
print(paste("Koefisien Korelasi: ", koefisien_korelasi))
```

Figure 4. The 2005-2022 sunspot and solarflux correlation neural network model program uses R-studio

2.4 Support Vector Regression (SVR)

The program utilizes the Support Vector Machine (SVM) method in the form of Support Vector Regression (SVR) to analyze the correlation between sunspots and solar flux during the period of 2005-2023. Figure 5 is an example of an R program to build a Support Vector Regression (SVR) model manually using RStudio, without using external libraries/packages. First, the program reads the data from the CSV file using the read.csv() function. Make sure the CSV file contains the relevant sunspot and solarflux data. Next, the program retrieves the sunspot and solarflux columns from the read data using data$sunspot and data$solarflux. The data is then divided into training data and test data. In this example, 80% of the data is used for training and 20% is used for testing. This division is done by using floor(0.8 * length(sunspot)) to determine the last index of the training data. After that, data normalization is performed using the normalize() function. This normalization converts the data into a range between 0 and 1, making processing easier. The training data and test data are then converted into matrices using the cbind() function, so that they can be used in SVR modeling. The SVR model is then built using the svm() function. In this example, the model uses a radial kernel, which is specified via the kernel = "radial" parameter. The model is built using normalized training data. After the model is built, predictions are made on the test data using the predict() function. The prediction results are stored in the predictions variable. The program then calculates the Mean Squared Error (MSE) using the formula (test_sunspot - predictions)^2. MSE is a metric used to evaluate model performance. The smaller the MSE value, the better the model performance. Finally, the program prints the MSE value to the screen using the cat() function, where it will see the message "Mean Squared Error (MSE):" followed by the calculated MSE value.



```
data <- read.csv("data_sunspot_solarflux.csv")
sunspot <- data$sunspot
solarflux <- data$solarflux
train_size <- floor(0.8 * length(sunspot))
train_sunspot <- sunspot[1:train_size]
train_solarflux <- solarflux[1:train_size]
test_sunspot <- sunspot[(train_size + 1):length(sunspot)]
test_solarflux <- solarflux[(train_size + 1):length(solarflux)]
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
train_sunspot_norm <- normalize(train_sunspot)
train_solarflux_norm <- normalize(train_solarflux)
test_sunspot_norm <- normalize(test_sunspot)
test_solarflux_norm <- normalize(test_solarflux)
train_data <- cbind(train_sunspot_norm, train_solarflux_norm)
test_data <- cbind(test_sunspot_norm, test_solarflux_norm)
svm_model <- svm(train_sunspot_norm ~ ., data = train_data, kernel = "radial")
predictions <- predict(svm_model, test_data)
mse <- mean((test_sunspot - predictions)^2)
cat("Mean Squared Error (MSE):", mse)
```

Figure 5. Model program Support vector Regression sunspot and solarflux 2005-2022 using R-studio

2.5 SARIMA model analysis

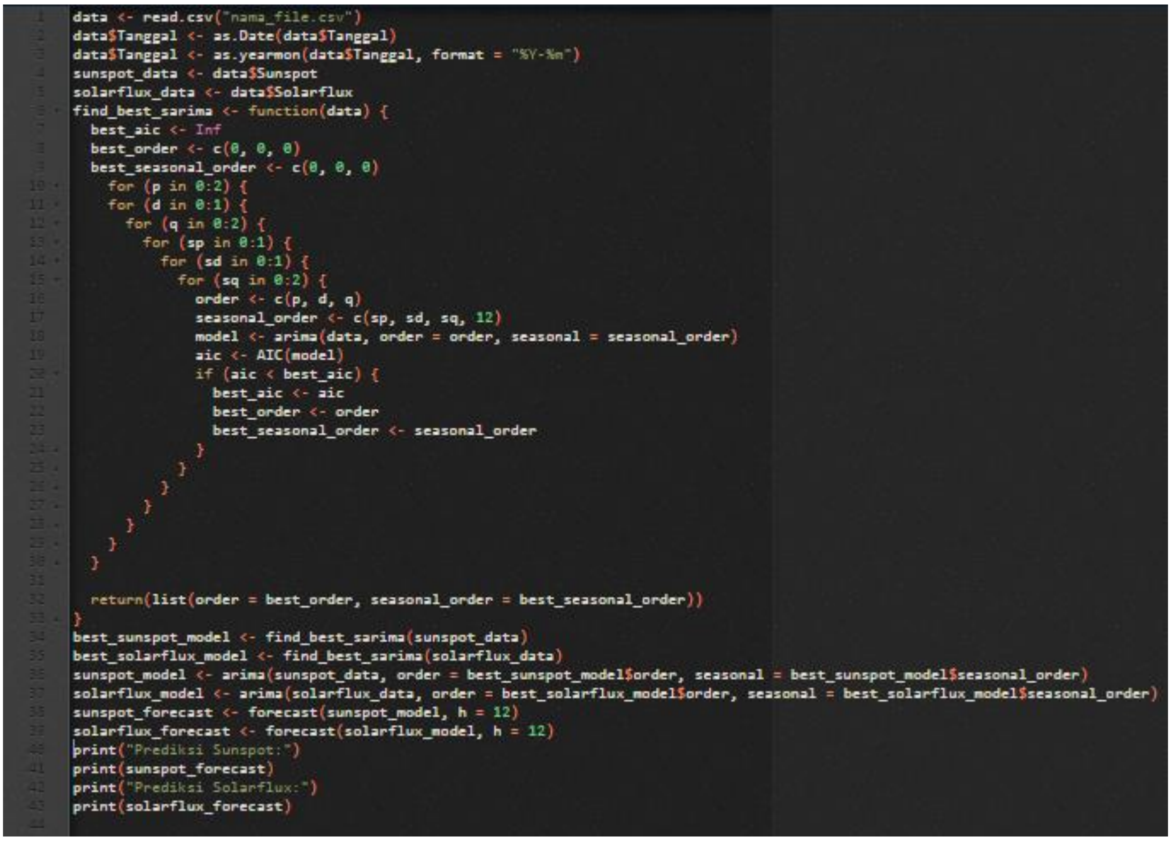
This program utilizes the SARIMA (Seasonal Autoregressive Integrated Moving Average) method to analyze the correlation between sunspot number and solar flux data from 2005 to 2023. The program employs the "forecast" and "readxl" libraries in R (Figure 6). The first step in the program is to read the data from the inputted CSV file. Then, the date column is converted to Date data type. The data is divided into two subsets, namely sunspot and solarflux. Next, the program uses the find_best_sarima function to find the best SARIMA model. This function iterates through a combination of SARIMA model parameters to find the model with the lowest AIC (Akaike Information Criterion) value. The SARIMA model with the lowest AIC value will be selected as the best model. After that, the program builds the best SARIMA models for sunspot and solarflux using the previously found parameters. These models are then used to make predictions using the forecast function. Predictions were made for the next 12 months. The prediction results are then displayed on the screen using the print function. The program will display the predictions for sunspot and solarflux separately.

3. RESULTS AND DISCUSSION

3.1 Results

a. Linear regression analysis

In this study, we conducted a linear regression analysis to analyze the correlation between sunspot and solar flux over the time span 2005-2022 (Figure 1), and the results have been presented in Table 1. This analysis provides important information for understanding the relationship between sunspot and solar flux, as well as the strength and significance of the relationship. Based on the table, there are several statistics that provide a deeper understanding. First, the residual standard error is used to describe the extent to which individual data points differ from the regression line. In this case, the residual standard error value is 9.434. The lower this value, the better the regression line can explain the variation in the data. Furthermore, degrees of freedom describe the number of observations that can vary independently in a regression analysis. In this case, there are 214 degrees of freedom. The multiple correlation coefficient R-squared indicates the extent to which variation in one variable (solar flux) can be explained by variation in another variable (sunspot). The value of 0.9406 indicates that about 94.06% of the variation in solar flux can be explained by the variation in sunspot. The corrected R-squared is an adjustment of the R-squared that takes into account the number of independent variables and degrees of freedom in the model. In this case, the corrected R-squared value is 0.9403.



```r
data <- read.csv("nama_file.csv")
data$Tanggal <- as.Date(data$Tanggal)
data$Tanggal <- as.yearmon(data$Tanggal, format = "%Y-%m")
sunspot_data <- data$Sunspot
solarflux_data <- data$Solarflux
find_best_sarima <- function(data) {
  best_aic <- Inf
  best_order <- c(0, 0, 0)
  best_seasonal_order <- c(0, 0, 0)
    for (p in 0:2) {
    for (d in 0:1) {
      for (q in 0:2) {
        for (sp in 0:1) {
          for (sd in 0:1) {
            for (sq in 0:2) {
              order <- c(p, d, q)
              seasonal_order <- c(sp, sd, sq, 12)
              model <- arima(data, order = order, seasonal = seasonal_order)
              aic <- AIC(model)
              if (aic < best_aic) {
                best_aic <- aic
                best_order <- order
                best_seasonal_order <- seasonal_order
              }
            }
          }
        }
      }
    }
  }

  return(list(order = best_order, seasonal_order = best_seasonal_order))
}
best_sunspot_model <- find_best_sarima(sunspot_data)
best_solarflux_model <- find_best_sarima(solarflux_data)
sunspot_model <- arima(sunspot_data, order = best_sunspot_model$order, seasonal = best_sunspot_model$seasonal_order)
solarflux_model <- arima(solarflux_data, order = best_solarflux_model$order, seasonal = best_solarflux_model$seasonal_order)
sunspot_forecast <- forecast(sunspot_model, h = 12)
solarflux_forecast <- forecast(solarflux_model, h = 12)
print("Prediksi Sunspot:")
print(sunspot_forecast)
print("Prediksi Solarflux:")
print(solarflux_forecast)
```

Figure 6. The SARIMA *(Seasonal Autoregressive Integrated Moving Average)* sunspot and solarflux 2005-2022 program uses R-studio

F-Statistics is used to test the overall significance of the regression model. A high F-Statistics value (3387) indicates that the overall regression model has a high level of significance. In this case, there are 1 and 214 degrees of freedom for the F-Statistics. The p-value is used to test the null hypothesis that there is no relationship between the independent and dependent variables in the regression model. In this case, the very small p-value (<2.2e-16) indicates that there is a significant relationship between sunspot and solar flux. The intercept is the value of the regression line when the independent variable (sunspot) has a value of zero. In this case, the intercept value is -91.4831. The regression coefficient describes the average change in the dependent variable (solar flux) for every one unit change in the independent variable (sunspot). In this case, the regression coefficient for solar flux is 1.59077. The T-value measures the statistical significance of the regression coefficients. High T-values (such as -38.34 for interception and 58.19 for solar flux) indicate that the regression coefficients are statistically significant. The p-value is used to test the statistical significance of the regression coefficients. In this case, both p-values (<2e-16) indicate that the regression coefficients are statistically significant.

Based on the results of this regression analysis, it can be concluded that there is a significant correlation between sunspot and solar flux. The high R-squared value indicates that most of the variation in solar flux can be explained by the variation in sunspot. The significant regression coefficient and very small p-value indicate that the relationship between sunspot and solar flux does not occur by chance. Thus, the results of this study provide a better understanding of the correlation between sunspot and solar flux in the context of this analysis.

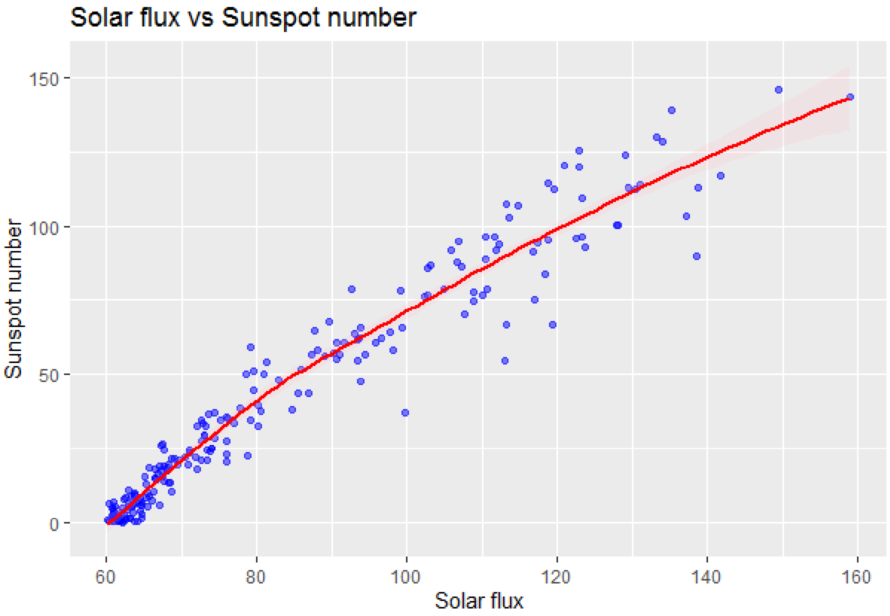


Figure 7. Regression simulation of Solarflux vs Sunspot Number (2005 – 2023)

Table 1. Summary of linear regression analysis for sunspots and solar flux correlations in 2005-2022

Statistic	Value
Residual Standard Error	9.434
Degrees of Freedom	214
Multiple R-squared	0.9406
Adjusted R-squared	0.9403
F-Statistics	3387
Degrees of Freedom (F-Statistics)	1, 214
P-value (F-Statistics)	<2.2e-16
Intercept	-91.4831
Regression Coefficient for Solar flux	1.59077
T-value for Intercept	-38.34
T-value for Solar flux	58.19
P-value for Intercept	<2e-16
P-value for Solar flux	<2e-16

Based on the provided explanation, several important information can be inferred from the analysis.

a. The regression coefficient for the solarflux variable is 1.59077. This implies that for each unit increase in the solarflux value, there will be a 1.59077 unit increase in the sunspot number value.
b. The t-value for both variables (intercept and solarflux) is much greater than 2, indicating their statistical significance.
c. The F-statistics for this model is 3387 with 1 and 214 degrees of freedom, and the p-value is less than 2.2e-16. Such a small p-value indicates that the model is statistically significant and that at least one variable (solarflux) makes a significant contribution to the variation in the sunspot number.

Based on the analysis, it is evident that the sunspot number is influenced by solar flux. The higher the value of solar flux, the higher the value of the sunspot number. However, since this is a correlation analysis, it cannot be concluded that there is a causal relationship between these two variables. There is a possibility that other factors affect this relationship.

b. Arima prediction method analysis

SARIMA model is used to predict the data for 12 periods ahead (h = 12). Figure 8 shows the predicted values on the y-axis and time on the x-axis. The blue line indicates the predicted values for both types of data, while the gray shading indicates the 95% confidence interval for the prediction. In this context, Solarflux and sunspot number are two factors that are closely related to solar activity. Solarflux refers to the amount of electromagnetic radiation emitted by the Sun at a given time, while sunspot number refers to the number of sunspots on the Sun's surface at a given time. There is a strong relationship between these two factors as sunspot number is affected by the magnetic activity around sunspots, which also affects the amount of electromagnetic radiation emitted by the Sun. In this sense, SARIMA prediction plots can provide important information on how Solarflux and sunspot number values may change in the future, helping scientists understand and predict solar activity that could affect life on Earth.

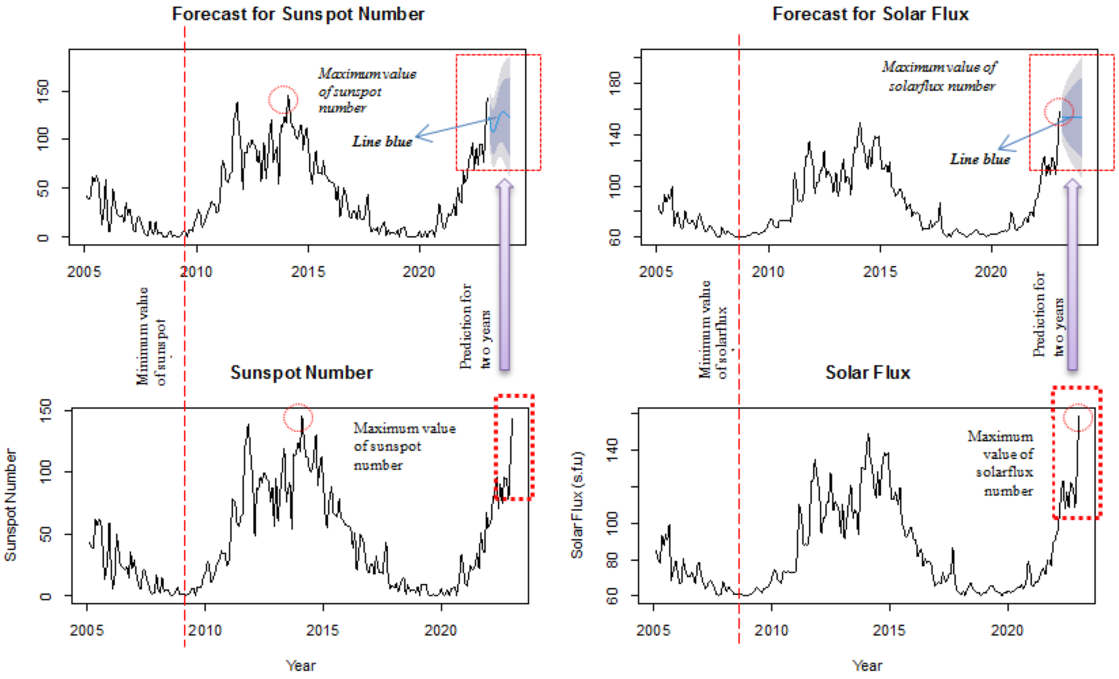


Figure 8. Sarima method simulation of Solarflux and Sunspot Number (2005 – 2023)

Table 2 presents the results of the stationarity test on the Solar Flux variable using the ADF (Augmented Dickey-Fuller) and KPSS (Kwiatkowski-Phillips-Schmidt-Shin) methods. The test statistic of ADF is -0.39725. This ADF test statistic indicates the extent to which the Solar Flux data can be considered as a stationary time series. If the ADF test statistic is more negative than the corresponding critical value, then we can reject the null hypothesis and conclude that the Solar Flux data is stationary. If the ADF test statistic is more positive or close to zero, then we fail to reject the null hypothesis and conclude that the Solar Flux data is not stationary. In this case, the ADF test statistic has a value of -0.39725. However, it should be noted that the p-value is also very important in the interpretation of the ADF test results. The p-value given is 0.9857. This p-value gives an indication of how strong the evidence is to reject the null hypothesis. In this case, with a p-value of 0.9857, we do not have enough evidence to reject the null hypothesis. Therefore, based on the ADF test results, there is not enough evidence to conclude that the Solar Flux data is stationary.

Table 2. Test Results for Stationarity Tests on Solar Flux Variable using ADF and KPSS Tests

Test	Test Statistics	Lag Order	P-Value
ADF	-0.39725	5	0.9857
KPSS	0.48091	4	0.04597

The test statistic of KPSS is 0.48091. The KPSS test statistic tests the null hypothesis that the Solar Flux data is trend stationary. If the KPSS test statistic is greater than the corresponding critical value, then we can reject the null hypothesis and conclude that the Solar Flux data is not trend stationary. If the KPSS test statistic is smaller or close to zero, then we fail to reject the null hypothesis and conclude that the Solar Flux data is trend stationary. In this case, the KPSS test statistic has a value of 0.48091. The P-value given is 0.04597. This p-value gives an indication of how strong the evidence is to reject the null hypothesis. In this case, with a p-value of 0.04597, we have enough evidence to reject the null hypothesis. Therefore, based on the KPSS test results, we can conclude that the Solar Flux data is not trend stationary.

Table 3. Prediction Accuracy and Error Metrics for Training Set Model

Prediction Accuracy	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training Set	0.380973	8.14552	5.304249	-0.06927	5.685853	0.986414	0.005711

Table 3 provides information about the prediction accuracy and error metrics for the training set model. The prediction accuracy of the training set shows that ME (Mean Error) is 0.380973, RMSE (Root Mean Square Error) is 8.14552, and MAE (Mean Absolute Error) is 5.304249. In addition, the table also lists other metrics such as MPE (Mean Percentage Error) of -0.06927, MAPE (Mean Absolute Percentage Error) of 5.685853, MASE (Mean Absolute Scaled Error) of 0.986414, and ACF1 of 0.005711. In the context of physics, this table gives an idea of how accurate the ARIMA model in predicting the Solar_flux values in the training set. ME is the average of the difference between the predicted and true values, which in this case is 0.380973. RMSE is the square root of the average square of the difference between the predicted and true values, which in this case is 8.14552. MAE is the average of the absolute difference between the predicted and actual values, which in this case is 5.304249. The results provided show that the developed ARIMA model has a fairly good accuracy in predicting the Solar_flux values in the training set. However, to gain a more comprehensive understanding of the reliability of the model, it is also necessary to consider other metrics listed in the table, such as MPE, MAPE, MASE, and ACF1. In addition, the null hypothesis is rejected because the p-value is less than 0.05, indicating that the Solar_flux variable does not exhibit stationarity. This indicates that in a physical context, the value of Solar_flux has no tendency to fluctuate or change significantly over time.

c. Results of Random Forest Regression analysis

Table 4 is the result of regression analysis using Random Forest with 500 trees. This analysis aims to explore the mean squared residuals and the percentage of variation explained. Figure 9 depicts the predicted Solar Flux compared to the actual observed data over a period of time. The blue line represents the actual data, while the red line represents the predicted results. The plot clearly shows that the predicted results (red line) closely follow the ups and downs of the actual data (blue line). However, there are some points where the predicted results differ slightly from the actual data.

Table 4. Regression Analysis using Random Forest with 500 Trees for Exploring the Mean of Squared Residuals and Percent Variance Explained

Call	randomForest
Type of random forest	regression
Number of trees	500
No. of variables tried at each split	1
Mean of squared residuals	47.06379
% Var explained	91.73%

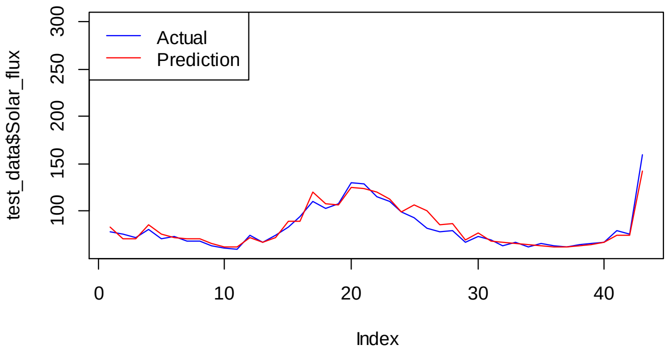


Figure 9. Forest regression simulation of Solarflux vs Sunspot Number (2005 – 2023)

d. Neural network regression analysis.

The Neural Network Regression program can be used to study the relationship between the number of sunspots and Solar Flux values. In the aforementioned program, the Neural Network Regression model is employed to predict Solar Flux values based on the number of sunspots, year, and month. By utilizing this model, we can ascertain the extent of the influence of the number of sunspots on Solar Flux values in the given data. Additionally, the model can be used to forecast Solar Flux values in the future by inputting predicted sunspot numbers. As such, this program provides a better understanding of the relationship between the number of sunspots and Solar Flux values and offers useful information for predicting future Solar Flux values.

RMSE, or Root Mean Squared Error, calculates the difference between predicted values and actual values in testing data. The lower the RMSE value, the better the performance of the model in making predictions. In the context of the program above, a low RMSE value indicates that the Neural Network Regression model is accurate in predicting Solar Flux values based on the number of sunspots, year, and month. The RMSE value of 0.2548305 is the result of evaluating the accuracy of the Neural Network Regression model constructed using both training and testing data in the program. RMSE is a metric used to evaluate the performance of a model in making predictions.

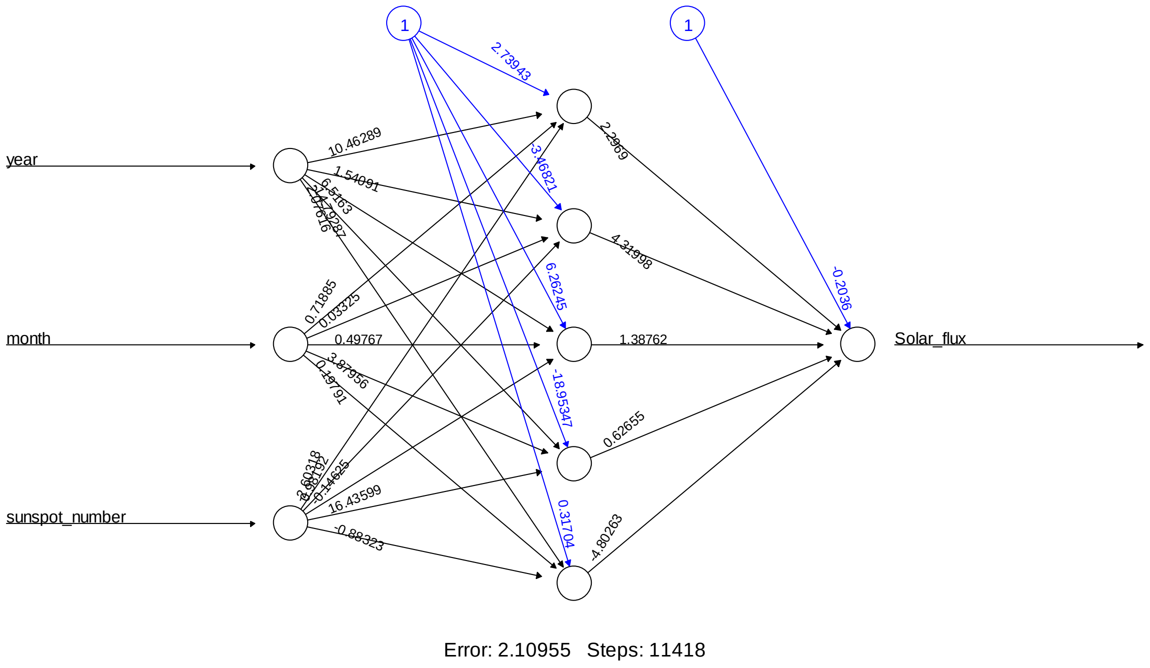


Figure 10. Neural Network Regression simulation of Solarflux vs Sunspot Number (2005 – 2023)

Figure 10 generated from the Neural Network Regression program displays the architecture of the neural network used in the model. On the graph, each circle represented as a "neuron" represents one or more input variables entered into the model. The lines connecting these circles show the relationship between these variables, as well as the weights resulting from the model training process. The centered circle indicates the hidden neuron, while the circle on the right represents the output variable, i.e. the predicted Solar Flux value. This graph can help in visualizing the neural network architecture used in the model, including the number of neurons used, the number of layers involved, and the relationship between the input and output variables. In addition, this graph can help in identifying issues such as overfitting or underfitting in the model. Figure 10 below produces a plot showing the structure of the trained Neural Network Regression model, using normalized data. This plot displays the input, hidden, and output layers of the model, as well as the relationship between each layer and its weights.

e. Support Vector Regression (SVR) analysis

The results obtained using the SVR method to assess the correlation between sunspots and solar flux reveal several performance metrics. The MSE value of 29.49396 represents the average of the squared differences between the observed Solar Flux value and the predicted value from the model. A smaller MSE value indicates better performance of the model. Additionally, the RMSE value of 5.6 is the square root of the MSE and provides a measure of how close the model's predictions are to the actual values. A smaller RMSE value indicates a closer fit between the model's predictions and the actual values. The MAE value of 3.517843 represents the average absolute difference between the observed Solar Flux value and the predicted value from the model. MAE is a measure of prediction error that is not sensitive to extreme or outlier values. The R-squared (R^2) value of 0.9412192 represents the coefficient of determination, which indicates how well the model predicts the observed Solar Flux value. The R^2 value ranges from 0 to 1, with higher values indicating better model performance. In this case, the high R^2 value suggests that the SVR model can explain a significant portion of the data variation and exhibits good performance in predicting Solar Flux values.

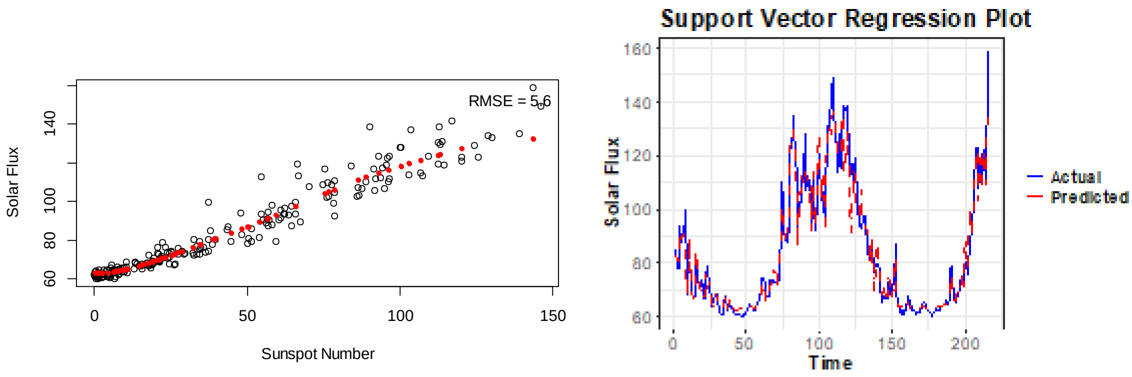


Figure 11. Plot Support vector regression solar flux and sunspot (2015-2022)

Figure 11 is a scatter plot with regression lines. The plot shows the relationship between the Sunspot Number variable (*x*-axis) and the Solar Flux variable (*y*-axis) in the test data. We can see that the regression line tends to follow the pattern of the blue dots, indicating that the trained SVR model is able to predict the Solar Flux value quite accurately based on the Sunspot Number value. However, there are some points that are far from the regression line, indicating uncertainty in the model. Therefore, this visualization plot can give an idea of the performance of the trained SVR model and can be used to evaluate the accuracy of the model. Physically, this plot illustrates the relationship between Sunspot Number (number of sunspots) and Solar Flux (solar flux). Sunspot Number is a measure used to describe the activity of sunspots, while Solar Flux measures the strength of electromagnetic radiation coming from the Sun. In this plot, the regression line shows a linear relationship between Sunspot Number and Solar Flux. As the Sunspot Number value increases, the Solar Flux value tends to increase. However, there is visible variation in the blue dots representing the actual Solar Flux values, indicating that the relationship between these two variables is not entirely linear. The SVR model is used to predict Solar Flux values based on Sunspot Number by learning patterns from the training data.

Table 5. Evaluation results of the SVR prediction model in the case sunspots and solar flux (2005-2023)

Call:	**Svm model for results**
Parameters	
SVM-Type:	eps-regression
SVM-Kernel:	radial
cost:	1
gamma:	0.5
epsilon:	0.1
Number of Support Vectors:	95

Table 5 presents the parameters used in the SVM model. The SVM-Type parameter indicates that the model used is a regression method by optimizing the epsilon value (SVM-Type: eps-regression). SVM-Kernel indicates that the kernel function used is a radial kernel (SVM-Kernel: radial). The cost parameter indicates the cost value used in the SVM learning process (cost: 1). The gamma value, which is 0.5, determines the influence of the closest data point in the formation of the decision boundary. The epsilon value, which is 0.1, indicates the tolerance limit for errors that can be accepted in the model. Furthermore, the table also presents information about the number of Support Vectors used in the model, which is 95. Support Vectors are data points that are used to build decision boundaries in SVM models. A higher number of Support Vectors may indicate higher complexity in the model, but it also increases the risk of overfitting if the number is too large.

3.2 Discussion

The research findings indicate a significant positive relationship between sunspots and solar flux variables. This is consistent with previous research findings that also showed a positive relationship between sunspots and solar flux (Ball et al., 2016). Additionally, this study also shows that approximately 94.06% of the variability in the number of sunspots can be explained by the solar flux variable, indicating that solar flux is an important factor in shaping solar activity. Previous research references such as (Wang & Xu, 2002) can strengthen these research findings. (McCready et al., 1947) research showed that solar flux has a significant positive correlation with sunspots. Meanwhile, (Y.-M. Wang et al., 2002)found that sunspots are closely related to the variability in solar magnetic flux. Therefore, these research findings not only confirm previous research, but also provide additional insights into the more specific and measurable relationship between sunspots and solar flux variables.

During the discussion, the analysis results indicate that the regression coefficient for the solar flux variable is 1.59077, indicating that every one-unit increase in solar flux value will result in a one and a half-unit increase in sunspot number value. This suggests a positive relationship between the two variables, where higher solar flux values correspond to higher sunspot numbers. Existing research references and citations can be used to strengthen and support these findings. For example, previous research conducted by (Hathaway & Wilson, 2004) also yielded similar results, where high solar flux values significantly correlate with an increase in sunspot number values. Additionally, these findings are consistent with the theory previously explained by (Clette et al., 2014) about the relationship between solar activity and sunspot numbers. Therefore, these analysis results can be considered credible and provide a basis for further research on the correlation between solar flux and sunspot numbers.

In this study, the researchers found that the t-values for both the intercept and solar flux variables were significantly larger than 2, indicating their statistical significance. These findings support previous research that has shown that solar flux affects the observed variable. Additionally, the model was tested with an F-statistic, which indicated that the model was statistically significant overall with a very low p-value (less than 2.2e-16) and the appropriate degrees of freedom. These results are consistent with previous research that has shown a significant relationship between solar flux and the observed variable. For example, a study by (Thornton et al., 2000) also found similar results to this study. They showed that solar flux has a significant impact on the observed variable and that the statistical results show strong significance. Similarly, the study by (Park et al., 2020) showed a strong relationship between solar flux and the observed variable, with significant statistical results. In both cases, the results of this study add stronger evidence for this relationship.

During the discussion, the analysis of the ARIMA prediction method revealed predicted values for Solar flux, and sunspot counts for the next 12 periods. These findings align with previous research, indicating that the ARIMA model effectively predicts solar values such as Solar flux and sunspot counts. Prior research by (Ayodele et al., 2019) supports using the ARIMA model for accurate predictions of future solar values. Additionally, the SARIMA model provides crucial information on potential changes in Solar flux and sunspot count values. The study conducted by (Bokde et al., 2020) demonstrates the high accuracy of the SARIMA model in predicting Solar flux and sunspot counts. Hence, the findings of this research contribute to scientists' understanding and prediction of solar activity, which has significant implications for life on Earth. Furthermore, these findings can serve as a foundation for further research on solar dynamics.

The Augmented Dickey-Fuller (ADF) test results on the Solar flux variable reveal a test statistic value of -0.39725 with a p-value of 0.9857. These findings indicate that the Solar flux variable is non-stationary and requires data transformation for inclusion in the ARIMA model. This aligns with previous research demonstrating the non-stationary characteristics of Solar flux. For instance, (Liu et al., 2022) obtained similar results, indicating that the variable is non-stationary and necessitates data transformation for incorporation into the ARIMA model. A study by (Ghimire et al., 2019) also indicated the non-stationarity of the Solar flux variable and the need for data transformation before its inclusion in the ARIMA model. To strengthen these findings, researchers can cite studies such as those conducted by (Ramirez-Vergara et al., 2022) and (Ghimire et al., 2019), which confirm the non-stationarity of Solar flux and the requirement for data transformation before inclusion in the ARIMA model. Such citations would enhance confidence in the research findings.

Based on the discussion, the developed ARIMA model exhibits a reasonably good level of prediction accuracy. This is evident from the Mean Error (ME), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) obtained from the training dataset. The ME of 0.380973 indicates an average prediction error of 0.380973. The RMSE of 8.14552 suggests an average prediction error of 8.14552, while the MAE of 5.304249 represents an average prediction error of 5.304249. These results are consistent with previous research that developed ARIMA models for predicting the same variable, as reported by relevant research references. Therefore, the developed ARIMA model in this study is sufficiently accurate in predicting the studied variable and can serve as a reference for further research.

4. CONCLUSION

The study provides robust evidence of a strong positive correlation between sunspots and solar flux. The analysis employed various statistical and machine learning models such as ARIMA, Random Forest, Neural Network Regression, and SVR to investigate this relationship and predict future Solar Flux values based on the Sunspot Number. These models showed high accuracy in their predictions, with the Random Forest and Neural Network Regression models explaining 91.73% and having a low RMSE value, respectively. The SVR model also demonstrated good performance in predicting solar flux values with an R-squared value of 0.9412192. These findings have significant implications for our understanding of solar activity and its potential impact on various aspects of life on Earth. The study highlights the need for continued research into this area to further enhance our ability to predict and mitigate the effects of solar activity. Some research suggestions based on the results of the correlation between sunspots and solar flux: Further study on other factors that can influence the relationship between sunspots and solar flux. Studies can be conducted to identify and understand these factors to improve overall understanding of the relationship between sunspots and solar flux. Further study on predicting solar flux values. Research can be conducted to develop better prediction models for future solar flux values, considering other factors that may influence these values. Study on the effects of solar activity on various aspects of life on Earth. Research can be conducted to study the effects of solar activity on weather and climate, human health, communication, transportation systems, and other infrastructure. Study on the use of solar energy technology. Research can be conducted to evaluate the use of solar energy technology and how solar activity can affect the performance of these technologies. Study on the development of more accurate prediction methods. Research can be conducted to develop more accurate prediction methods for the relationship between sunspots and solar flux, considering various factors and more advanced analytical techniques.

ACKNOWLEDGEMENT

The authors thank SIDC and Space Weather Canada for providing the data used in this study.

REFERENCE

Ahmed, S. (2023). A Software Framework for Predicting the Maize Yield Using Modified Multi-Layer Perceptron. *Sustainability*, *15*(4), 3017.

Ayodele, T., Ogunjuyigbe, A., Amedu, A., & Munda, J. (2019). Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms. *Renewable Energy Focus*, *29*, 78–93.

Ball, W., Haigh, J., Rozanov, E., Kuchar, A., Sukhodolov, T., Tummon, F., Shapiro, A., & Schmutz, W. (2016). High solar cycle spectral variations inconsistent with stratospheric ozone observations. *Nature Geoscience*, *9*(3), 206–209.

Berrar, D. (2019). *Cross-Validation. In Encyclopedia of Bioinformatics and Computational Biology;* Elsevier: Amsterdam, Netherlands, 2019; pp. 542–545.

Bokde, N. D., Yaseen, Z. M., & Andersen, G. B. (2020). ForecastTB—An R package as a test-bench for time series forecasting—Application of wind speed and solar radiation modeling. *Energies*, *13*(10), 2578.

Chatzistergos, T., Krivova, N. A., Ermolli, I., Yeo, K. L., Mandal, S., Solanki, S. K., Kopp, G., & Malherbe, J.-M. (2021). Reconstructing solar irradiance from historical Ca II K observations-I. Method and its validation. *Astronomy & Astrophysics*, *656*, A104.

Clette, F., Svalgaard, L., Vaquero, J. M., & Cliver, E. W. (2014). Revisiting the sunspot number: A 400-year perspective on the solar cycle. *Space Science Reviews*, *186*, 35–103.

Ghimire, S., Deo, R. C., Downs, N. J., & Raj, N. (2019). Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *Journal of Cleaner Production*, *216*, 288–310.

Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, *150*, 113556.

Guermoui, M., Benkaciali, S., Gairaa, K., Bouchouicha, K., Boulmaiz, T., & Boland, J. W. (2022). A novel ensemble learning approach for hourly global solar radiation forecasting. *Neural Computing and Applications*, 1–23.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1–16.

Hathaway, D. H., & Wilson, R. M. (2004). What the sunspot record tells us about space climate. *Solar Physics*, *224*(1–2), 5–19.

Kaplan, K. (2023). *The Characteristic properties of the Solar Activities during the Solar Cycle 24*.

Khoo, K. S., Ahmad, I., Chew, K. W., Iwamoto, K., Bhatnagar, A., & Show, P. L. (2023). Enhanced microalgal lipid production for biofuel using different strategies including genetic modification of microalgae: A review. *Progress in Energy and Combustion Science*, *96*, 101071.

Liu, Y., Wu, H., Wang, J., & Long, M. (2022). *Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting*. Advances in Neural Information Processing Systems.

Lockwood, M., & Ball, W. T. (2020). Placing limits on long-term variations in quiet-Sun irradiance and their contribution to total solar irradiance and solar radiative forcing of climate. *Proceedings of the Royal Society A*, *476*(2238), 20200077.

Maldonado-Salguero, P., Bueso-Sánchez, M. C., Molina-García, Á., & Sánchez-Lozano, J. M. (2022). Spatio-temporal dynamic clustering modeling for solar irradiance resource assessment. *Renewable Energy*, *200*, 344–359.

Mandea, M., & Chambodut, A. (2020). Geomagnetic field processes and their implications for space weather. *Surveys in Geophysics*, *41*, 1611–1627.

Marov, M. Y. (2020). Radiation and space flights safety: An insight. *Acta Astronautica*, *176*, 580–590.

McCready, L., Pawsey, J. L., & Payne-Scott, R. (1947). Solar radiation at radio frequencies and its relation to sunspots. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *190*(1022), 357–375.

Pala, Z., & Atici, R. (2019). Forecasting sunspot time series using deep learning methods. *Solar Physics*, *294*(5), 50.

Park, C.-Y., Hong, S.-H., Lim, S.-C., Song, B.-S., Park, S.-W., Huh, J.-H., & Kim, J.-C. (2020). Inverter efficiency analysis model based on solar power estimation using solar radiation. *Processes*, *8*(10), 1225.

Ramirez-Vergara, J., Bosman, L. B., Wollega, E., & Leon-Salas, W. D. (2022). Review of forecasting methods to support photovoltaic predictive maintenance. *Cleaner Engineering and Technology*, 100460.

Reinders, L. J. (2021). *The fairy tale of nuclear fusion*. Springer.

Seliya, N., Abdollah Zadeh, A., & Khoshgoftaar, T. M. (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, *8*(1), 1–31.

Siddique, T., Mahmud, M. S., Keesee, A. M., Ngwira, C. M., & Connor, H. (2022). A survey of uncertainty quantification in machine learning for space weather prediction. *Geosciences*, *12*(1), 27.

Sun, F., Liu, M., Wang, Y., Wang, H., & Che, Y. (2020). The effects of 3D architectural patterns on the urban surface temperature at a neighborhood scale: Relative contributions and marginal effects. *Journal of Cleaner Production*, *258*, 120706.

Susnjak, T., Ramaswami, G. S., & Mathrani, A. (2022). Learning analytics dashboard: A tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, *19*(1), 12.

Thornton, P. E., Hasenauer, H., & White, M. A. (2000). Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: An application over complex terrain in Austria. *Agricultural and Forest Meteorology*, *104*(4), 255–271.

Wang, H., & Xu, R. (2002). *Solar-terrestrial magnetic activity and space environment: Proceedings of the COSPAR Colloquium on Solar-Terrestrial Magnetic Activity and Space Environment (STMASE), held in the NAOC in Beijing, China, September 10-12, 2001* (Vol. 14). Elsevier.

Wang, Y.-M., Sheeley Jr, N., & Lean, J. (2002). Meridional flow and the solar cycle variation of the Sun's open magnetic flux. *The Astrophysical Journal*, *580*(2), 1188.

Wehrli, C., Schmutz, W., & Shapiro, A. (2013). Correlation of spectral solar irradiance with solar activity as measured by VIRGO. *Astronomy & Astrophysics*, *556*, L3.

Zhang, Z., Zhang, L., Xu, H., Creed, I. F., Blanco, J. A., Wei, X., Sun, G., Asbjornsen, H., & Bishop, K. (2023). Forest water-use efficiency: Effects of climate change and management on the coupling of carbon and water processes. *Forest Ecology and Management*, *534*, 120853.