Abstract

This paper describes the analysis of solar forcing parameters in the ionosphere and proposes a software solution to analyses and plot different parameters associated with the phenomenon.

List of Figures

List of Tables

Contents

Introduction

A large number of studies are carried out in the ionospheric region of the atmosphere in order to study about solar activities. The solar activity is the physical phenomena that occurs at solar magnetic field which fluctuates on time.

What are solar forcing parameters

Why solar forcing parameters are important

What is correlation

How solar forcing parameters are correlated

Where all we can use correlation between solar forcing parameters

What are the uses of correlation between solar forcing parameters

* Motivation
* Problem Statement
* Goal of the thesis

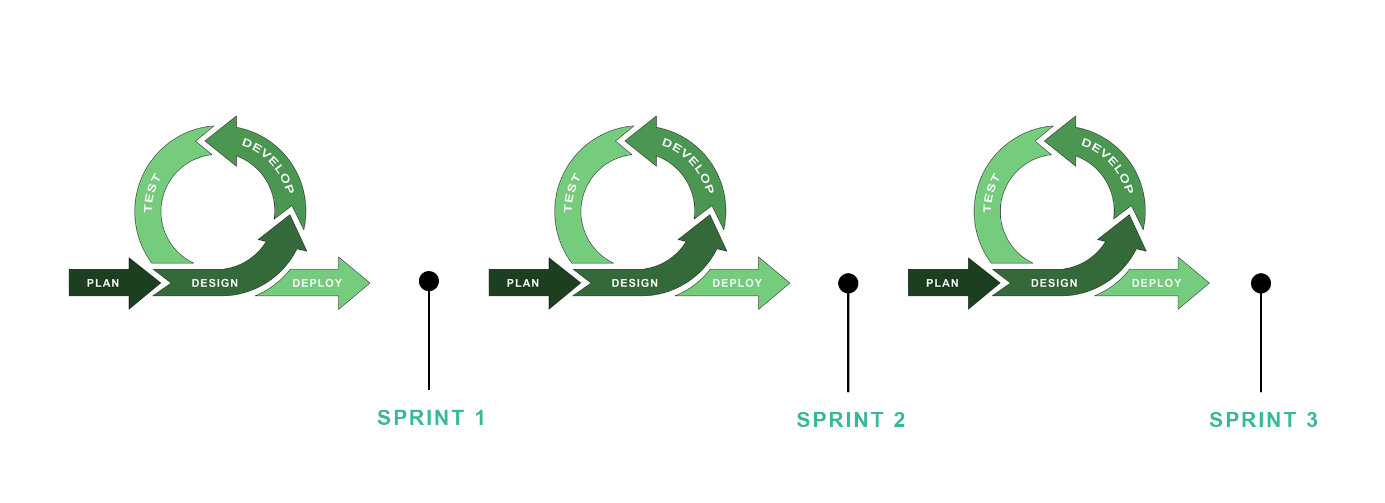
Theory

* Ionosphere
  + Atmosphere
  + Space weather
* Correlation of solar forcing parameters
  + TEC
  + Soalar radio flux
  + Solar wind speed
* Requirements of analysis software
* Existing Tools
* Data
  + Datasets
  + preprocessing

Software Development

Modern software tools are required to analyze large datasets. The data set can be raw data or unstructured data, which has to be formatted according to the requirements of the analysis. Data scientists extract information from these data sets for data driven decision-making. The accuracy of data is critical because inaccurate data leads to faulty conclusions. Frequent use of the same analytics for multiple scientific researches brings the requirement of a software tool to save time and cost of the research.

A good software product is developed through certain steps and guidelines, which are called software development life cycle (SDLC). SDLC is a used by software industries to design, develop, test and maintain quality software products. (Yadav, 2015) The software to analyze scientific data has been developed by the popular SDLC method called agile software development methodology. It is a modern software development methodology having flexibility and incorporates levels of practicality into the final product. Agile methodology focuses on keeping the code simple and delivering functional parts of the application as it finishes. Agile development is a kind of incremental software development with rapid development cycles. (Yadav, 2015)



SDLC – software development life cycle steps

* Developing a tool to analyze scientific data
  + Design
    - Technologies Used
      * Python
      * Dependencies
        + Numpy
        + Pandas

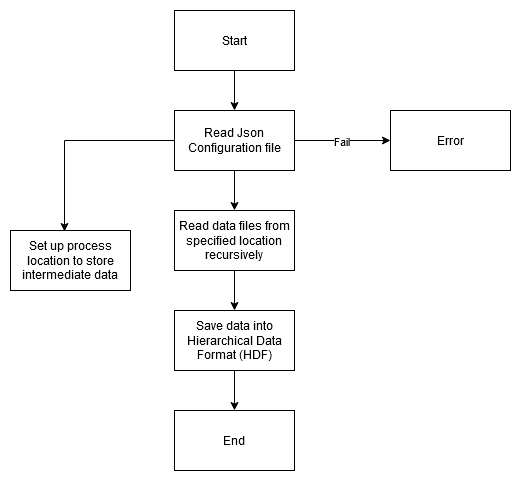
dataframe

* + - * + Pycdf
        + matplotlib

The analyzing process is divided into 4 independent steps, which are setting up the project and loading the data, data cleaning and formatting, computations, visualization. Each step in the process flow has its own data configurations to handle different data sets and different operations. The configurations files are written in JSON format and saved as json file, which makes it easier to handle in the software. There will be independent folder locations for each process step and these folders contains the configuration files.

* + - Workflow
      * Setting up the project and loading the data

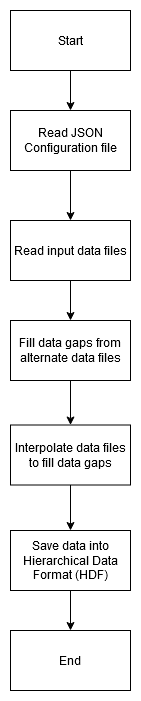
Workflow start with setting up the project. Here in this step a project folder is created in project location with name of the project. All the intermediate files and final results are saved into this folder during the process workflow. Project name and project location is saved into the configuration file for easy reference during the program execution. The raw data for the analysis is available in different file formats like comma-separated values (CSV), text files and other formats that may make the read process slower. To make the data handling and computations faster, data files are read and converted to a binary file format called hierarchical data format (HDF) in this process step.



* + - * Data cleaning and formatting

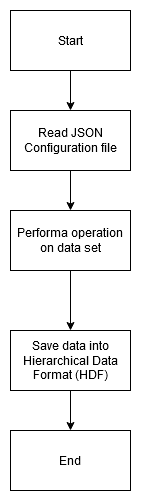
Accuracy of the analysis performed is highly depended on the quality of data. Raw data consist of data from different sources, duplicates, corrupted data or incorrect data. In this step raw data files are read from the specified location and making it ready for operations. The step ensures there is sufficient number of data points in the time series data, merging data columns from different data sources and interpolating data points if needed. Since the program handles time series data, it has to ensure the time series has data for all the data points according the temporal resolution. However, leaving out those missing data may result in a temporal mismatch. The dataset can be in different files for different temporal resolution, this leads to gaps in the time series. This scenario can be managed by merging data columns from different files based on the time frequency index. This mechanism helps to maintain the consistent of data for analysis. After applying these data filling mechanisms, still there can be missing data points because no availability of data from the source. When dealing with temporal data, the removal of data point is not a solution. One approach is to fill in the missing data with the mean value of adjacent series. Interpolation is a mathematical approach for extrapolating missing data by adjusting a function to the data. The basic form of interpolation is linear interpolation, which takes the mean of the values before and after the missing data.

Data cleaning and formatting is a time-consuming process as the data scientists spend 60% of their time on this step. Due of the varying nature of data curation, most activities require human supervision. So, the user must provide integrity criteria, specify statistical parameters, and annotate data values as part of the data cleaning process.



* + - * + Read data files
        + File formats(hdf)
        + Save formatted files
      * Computations

The objective is to evaluate raw data from different data sources. The applications rely on extracting relevant patterns and information from massive data volumes. This is because many of these systems' complexity or governing equations are unknown or hard to extract. Thus, spectral approaches and linear algebra rely on data, its statistical features, and analysis. Statistics may be a potent instrument in the practice of data analysis. In a broad sense, statistics is the application of mathematics to the technical study of data. A simple visualization such as a bar chart may provide some high-level information, but statistics allows us to work on the data in a far more information-driven and focused manner. The mathematics involved enables us to draw solid inferences from our data, as opposed to mere speculation. The third step of data analysis is the actual data analysis. Finding meaning in the data can be challenging without the appropriate tools. This section examines some of the operations with Python.



* + - * + Mean

The (arithmetic) sample mean is the most frequent measure of the center of a quantitative variable. Typically, when individuals refer to taking an average, they are referring to the mean. The sum of the variable's observed values in a data set is divided by the total number of observations to arrive at the sample mean for that variable.

To effectively communicate the concepts and accompanying computations, it is practical to represent variables and observed values of variables with symbols. So to signify the variable in issue, and to denote the variable's th observation in the data set.

Considering sample size as then the mean of the variable is

To further simplify the sum of the sample is denoted as

The function receives data set as input and returns the data set after performing the operation. Whether to calculate mean based on column wise or row wise decided by the specified input fields in the configuration file.

* + - * + Interpolation

Python's interpolation function estimates unknown data points between two known data points. Interpolation is typically employed during data analysis to fill in missing values in the dataframe or series. Interpolation may be used to discover a missing value using its neighbors. When averaging missing data does not provide the greatest fit, here comes the significance of interpolation. Interpolation is commonly employed when working with time-series data since missing values are typically recovered with the preceding values. Linear Interpolation is the estimation of a data point by connecting dots in ascending or descending order along a straight line. The unknown value is estimated in the same progressive sequence as previous values. The default technique employed by Interpolation is Linear. Therefore, linear interpolation operates in the same sequence. Note that it interprets data by connecting points in a straight line rather than considering the index.

* + - * + Moving average

Smoothing is performed by averaging measurements over many time intervals to decrease noise. The moving average is a basic smoother that generates a series of averages by averaging variables over a window of specific periods of time. A moving average with window size entails averaging over each set of w consecutive data, where is a user-specified variable. There are two kinds of moving averages in general: the centered rolling average and the trailing moving average. Because the averaging procedure can reduce seasonality and noise, centered moving averages are effective for highlighting patterns. For predicting, trailing moving averages are beneficial. The positioning of the average window over the time - series data differentiates the two.

Moving average at time is calculated by centering the moving window frame at time and averaging the data point values within the window.

The only decision the user needs make when using moving average modeling or visualization is the width of the window (). (Shmueli & Lichtendahl, 2016)

* + - * + Relative difference

When numbers change, people can report the magnitude of the change in relative or absolute form. It determines what percentage (greater or lesser) the number changed from the initial data value. Small relative changes might appear to be more important than they are. This is due to the fact that even a modest absolute change in a number can result in a significant percentage change.

* + - * + Correlation

Numerous researches perform correlation analysis to determine the degree of relationship between variables. Analysis of linear correlation is a method for depicting the proximity between two variables. Almost all studies employ correlation analysis for the same objective, which is to investigate the association between independent and dependent variables.

The purpose of correlation is to examine the degree of association between two variables under study. The correlation coefficient quantifies the degree of link between variables. Pearson's Product Moment Correlation Coefficient and Spearman's Rank Correlation Coefficient are the two most common correlation coefficients often used in applications. This study focuses on the applications of Pearson's Simple Linear Correlation to investigate the relationship between variables.

Karl Pearson formed the correlation coefficient and looked into it for the first time in 1896 (Hauke and Kossowski, 2011). Francis Galton and Auguste Bravais came up with the ideas of correlation and relative contribution (Denis, 2001). Hauke and Kossowski (2011) agree that the Pearson's Product Moment Correlation Coefficient (R or r) is a scale for figuring out how strongly two variables are linked linearly. As it measures how linearly two variables are related, interval or ratio variables should be taken into account, but only if the variables in study have a normal distribution. (Senthilnathan, 2019)

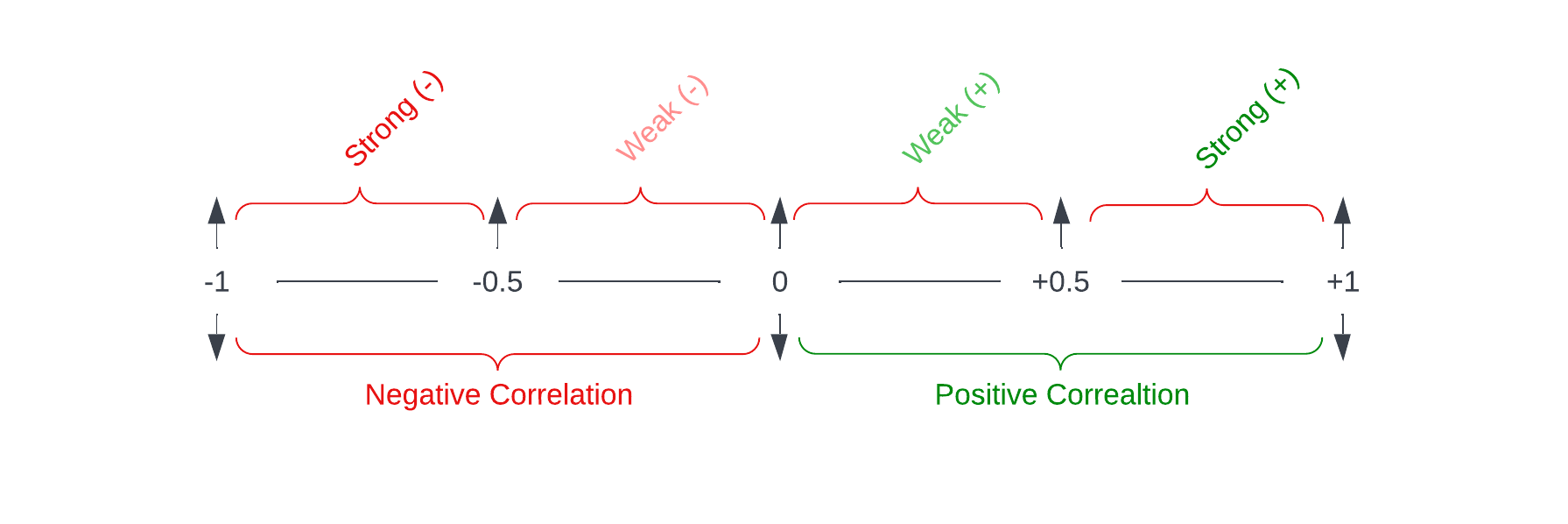
Pearson's formula for quantifying the level of correlation (R) between two variables, specifically X and Y, is as follows:

Where is the number of observations, is the measure of variable 1 and is the measure variable 2.

Depending on the direction, the degree of correlation can be classified as Positive, Zero, or Negative. In practice, perfect zero correlation coefficients between variables are uncommon; consequently, positive and negative correlations can be classified identically in studies.

If the pattern of a variable is positive and nearly identical to that of another variable, there is a possibility of positive association between the variables, and such an association can result in a positive correlation coefficient; conversely, if the trend of a variable is positive and nearly opposite to that of another variable, there is a possibility of negative association between the variables, and such an association can result in a negative correlation coefficient.

Fundamentally, the correlation coefficient R will vary between and , or . The correlation coefficient cannot be interpreted in a particular way. Figure 1 illustrates, according to Gogtay and Thatte (2017), how the correlation coefficient might be interpreted depending on its value.



Correlation with box window

Time series data can be characterized as continuous, equal-interval sequences. In practice, recent series values is compared using the sliding window technique. Let be a stream time series, consisting of , where is the most recent data and is the user-defined window length. The data in the sliding window will be continuously updated as sliding windows goes from the start of the analysis period to end. Sliding window size can be defined by user in the configuration file.

Correlation with gaussian window

In order to obtain the cross correlation, a gaussian window function is applied on each data sets before calculating the cross correlation.

* + - * + Confidence interval

A confidence interval for a correlation coefficient is a set of values that, with a certain degree of certainty, contains the population correlation coefficient. This procedure computes the sample size required to get a particular confidence interval width for the Pearson product-moment correlation coefficient at a specified confidence level.

* + - * Visualization

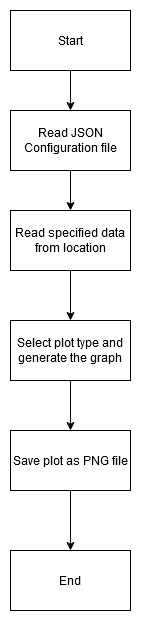
Data visualization is an essential component of the scientific method. A scientist will be able to comprehend their own data and communicate their ideas to others through the use of effective visualizations. These objectives can be advanced by graph-specification tools that strike a fair trade - off between performance and adaptability. The matplotlib project is relatively well within scientific Python community, having been continuously developed for over a decade. It offers fine-grained flexibility over the arrangement and visual appearance of things in a plot and is very adaptable. (Waskom, 2021)

Data visualization is among the most essential data finding and exploration strategies. Despite the fact that visualization is not termed a data science technique, visual mining and pattern identification based on visuals are becoming increasingly popular in the corporate sector. Data visualization is the practice of conveying data in an abstract graphic form. The visual depiction of complicated data with various properties and their underlying interconnections facilitates comprehension. Motives for employing data visualization include:

*Comprehension of complex material:* A basic visual chart may include massive amounts of data without difficulty. By utilizing visuals, the viewer is able to grasp the overall picture as well as longer term trends that are exceedingly difficult to comprehend when facts are expressed just in numerical form.

*Relationships*: Visualizing data in Cartesian coordinate system facilitates investigation of the characteristics' relationships. Whereas able to represent more than three components on the x, y, and z-axes is not possible in Cartesian coordinates, there are a few innovative solutions available, such as altering the size, color, and shape of data markers or employing flow maps, which use more than two attributes in a two-dimensional medium. (Kotu & Deshpande, 2018)

Visualisation in the project can be customized within the configuration file saved in the project folder. It supports different kinds of plots and can be extended in future development. Currently there are line graphs and color maps for visualizing data. On executing the visualization command with corresponding configurations will generate the graphs specified and save to project folder.



* + - * + line graphs

A line chart illustrates the progression of one or more numerical variables. It is among the most used chart types, therefore it can be constructed using any python graphic library, such as matplotlib, seaborn, or plotly.

* + - * + colour maps

The use of color in the presentation of scientific findings is increasingly commonplace. Color permits the plotting of more (complicated) data in the same plot without resorting to complicated 3D plots or subplots or interactive applications. However, a sometimes unappreciated feature of data visualization is the effect of color on the interpretation of the shown data, making the selection of the appropriate colormap critical (van der Velden, 2020).

* + - * + interactive plots
    - UI
      * Design
      * Libraries
        + streamlit
  + Build
    - File Formats
      * Cdf
      * Hdf
      * Csv
      * Txt
      * json
    - Metadata configurations
  + Test
  + Deployment
    - Deployment in cloud machine
  + Maintenance
    - repository

Results

Conclusion

Reference

Appendix