DATASCI 507 – Final Project Proposal

Jackson MacTaggart

Overview

In this project, we will use machine learning models to attempt to predict the strength of the solar cycle. First, let us establish some background. The sun goes through an 11-year, sinusoidal cycle where it goes from a period of "baseline", calm, and quiet activity (called solar minimum) to a period of higher level, volatile activity 5 ½ years later (called solar maximum). One of the open questions in solar physics is why the sun goes through this cycle and why this cycle is 11 years long (on average). Even more pressing is the question of whether the behavior of the solar cycle can be predicted.

In this project, we will use a variety of parameters to attempt to create predictions for the activity of the solar cycle. We will focus on using a quantity known as total open magnetic flux. It will be helpful for us to break down this term. In physics, flux is the amount of "stuff" traveling through some area. Total open magnetic flux is the amount of all the magnetic field lines flowing through some area at some distance from the sun. As we will discuss in the next section, this has been shown to be a potentially incredibly useful quantity for predicting the behavior of the solar cycle.

I personally want to focus on this project because it is the focus of my research as a PhD student. My research into this topic focuses primarily on utilizing Python to analyze in situ data taken by satellites studying the sun. I utilize this data and apply it to developing theory to validate said theory and drive it further forward. One of the primary reasons I took this class was to learn better skills to help me in my use of Python in my research. While taking this class, it has become apparent to me that utilizing machine learning models with scikit-learn could be very useful and has the potential to produce quality results.

The data I will work on is in situ data from solar satellites. In situ data is collected by a given satellite at the location of said satellite. I will use data from the Advanced Composition Explorer (ACE) and Ulysses satellites, as these satellites produce the most reliable data used to study this research topic. This data is multifaceted, meaning that there is a plethora of different parameters available for use. The fact that there are so many parameters in the data is particularly useful because of the types of insights I am expecting from this project. This is a supervised learning problem. I expect to find that some set of solar physics parameters will be able to predict the activity of the solar cycle to a reliable degree. This specific machine learning algorithm/model that performs the best is hard to predict, but I expect that this will be an additional insight.

Prior Work

Prior to 2011, the standing theory in solar physics concerning the open magnetic flux at solar minimum was that the solar minimum should represent the "baseline" level of activity for the

sun and the open magnetic flux should remain constant across the minima. However, this was found to vary. In fact, from the cycle 23 minimum in 1998 to the cycle 24 minimum in 2009, the open magnetic flux was found to drop by about 40%. For a quantity that is supposed to remain constant (or relatively so) from solar minimum to solar minimum, this was a large change and a contradiction to the standing theory. In 2011, Zhao and Fisk modified how open magnetic flux is defined and found that this new definition open magnetic flux to be conserved from solar minimum to solar minimum. Their new theory was validated for the cycle 22, 23, and 24 solar minima and the results were published in the *Solar Physics* journal. In 2024, MacTaggart, Zhao, Lepri, and Fisk (unpublished) validated this theory for the most recent solar cycle 25 minimum.

In addition to validating the most recent solar minimum, the 2024 work showed in brief data analysis that a quantity known as the "solid angle of the open flux region" could act as a predictor for the strength of the following solar cycle. This was not truly followed up on until now.

The potential methods that can be used to achieve the stated project goal are listed as follows.

- Visualization through Python packages such as matplotlib or seaborn will be crucial for establishing initial patterns to follow up on.
- Pandas will be another crucial package to use, as it will be the best way for us to set up and perform data analysis
- Scikit-learn will allow us to easily access and implement a variety of machine learning algorithms/models

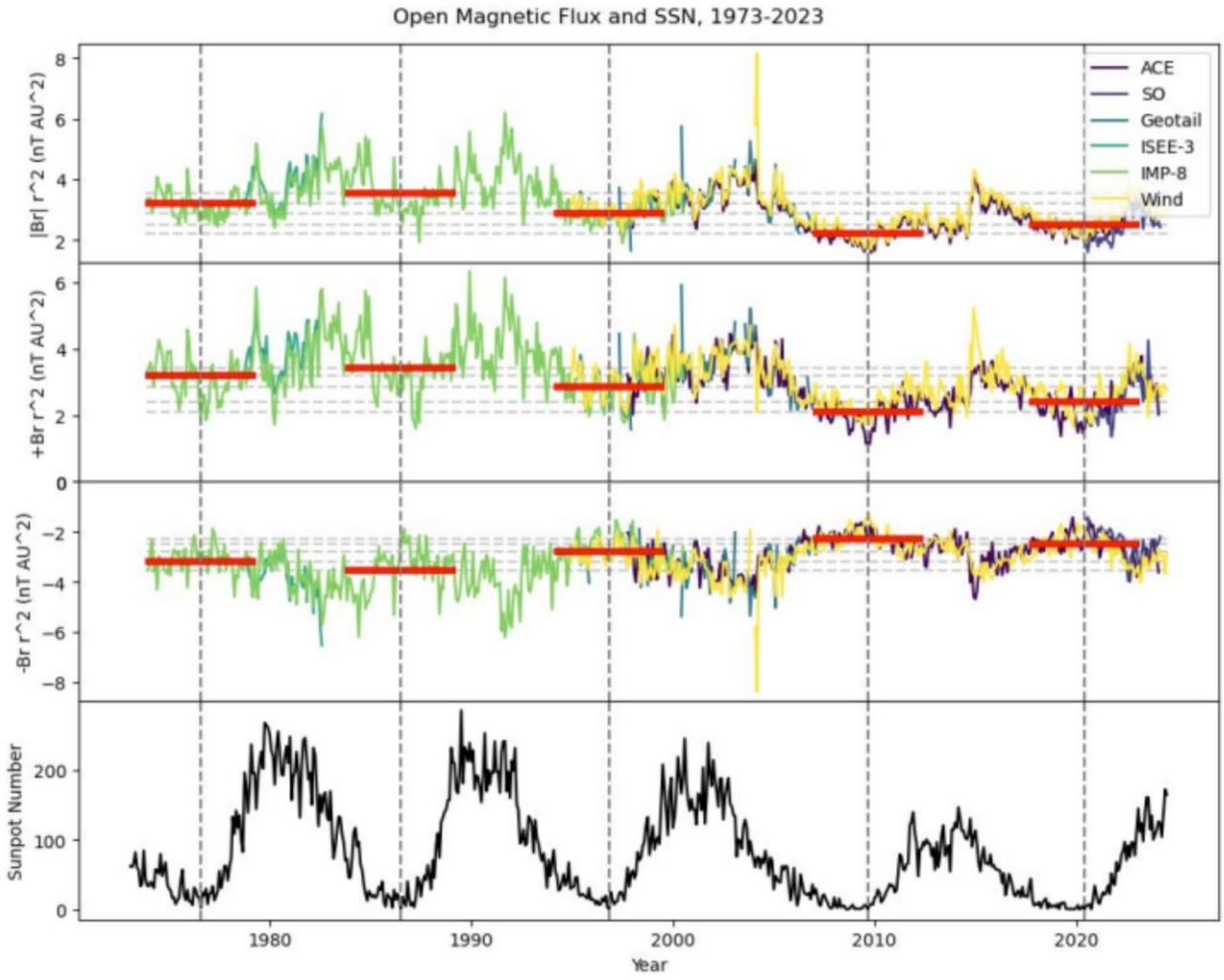
Preliminary Results

Pandas was extremely useful in helping to produce preliminary quantitative results. Our preliminary results pull from current analysis and past published results (such as Zhao and Fisk 2011). This can be seen in the following table.

	Streamer half-width (degree)	Non-streamer-stalk region solid angle	Relative Ratio of $\|Br\|r^2$	Relative Ratio of Total Open Magnetic Flux
22-23 minimum	25	1	1	1
23-24 minimum	10	1.43	0.74	≈1
24-25 minimum	20	1.14	0.87	0.99≈1

This table shows the relative ratios of total open magnetic flux and associated parameters for the last three solar minima. The first two rows are the results of Zhao and Fisk 2011. The third row shows the new results of this work from 2024 that further validated the 2011 conservation theory. What is new to this project as a preliminary result is the use of Pandas to study the non-streamer stock region solid angle measurements in conjunction with the measurements of open magnetic flux in the third column. From this preliminary Pandas data analysis, we can see that, for a given

minimum, if there is a bigger solid angle of the open flux/non-streamer stock region, physics tells us that the open flux is more spread out. This makes it harder to form what are known as polar coronal holes, which in turn causes a weaker solar cycle. This was easier to visualize through the use of matplotlib, which allowed us to create the following figure.



First, let us focus on the last panel labeled “Sunspot Number.” This is a good measure of the sinusoidal behavior of the solar cycle. For each of the solid angles produced by Pandas and shown in the above table, we can see the aforementioned relationship appear. A bigger solid angle at a corresponding minimum tends to be followed by a smaller peak in the following sunspot number maximum. The opposite is also true. We can also see similar patterns produced by in situ measurements of the open magnetic flux.

The fact that these two parameters alone indicate some sort of accurate predictive model is very promising. If we include other in situ parameters and explore further with as many different machine learning algorithms/models as time allows, we are confident that we can produce a successful project.

Project Deliverables

A successful project will produce an answer for what combination of machine learning model and solar physics parameters produce accurate predictions for the behavior of the solar cycle. Ideally, we will have a list of parameters that we will be able to feed into a given machine learning model to produce accurate predictions. We hope that we can show that these predictions will have high accuracies, low cross validation errors, etc.

Our primary sub-goal is to be able to extend our predictive analysis to a wider range of time periods. Solar physics is a relatively new field and the idea of using satellites to study the sun is even newer. Therefore, the amount of reliable data is somewhat limited. As of right now, we only have three solar cycles worth of usable data. As we use data going back into the 70s and 80s, we simultaneously reach back towards the beginning of satellite science where data collection methods were scarcer and less reliable than today's methods. For example, parameters such as solid angle of certain regions were not widely available 50 years ago. If we can produce confident results for more modern solar minima, we would like to see if we can extend at least a similar analysis even further to acquire more data points to further support our (hopefully) successful results.

Timeline

- Week 1-2: Process data using Pandas and matplotlib/seaborn to choose what parameters will be most useful
- Week 3-4: Focus on feeding data into machine learning models/algorithms to see which models produce the best predictive results
- Week 5: Refine results and produce final report