UNIVERSITÀ DEGLI STUDI DI ROMA TOR VERGATA
MACROAREA/FACOLTÀ DI SCIENZE MATEMATICHE,
FISICHE E NATURALI

CORSO DI STUDIO IN
FISICA - Curriculum Erasmus Mundus MASS

TESI DI LAUREA IN
SPACE WEATHER

TITOLO

The Sun as a star: Investigating the Influence of Solar Activity Features on Radial Velocity Signals

Relatore:
Physics Dept., University of Rome (Tor Vergata)
Prof. Francesco Berrilli

Laureando:
matricola: 0325661
Brandon Rajkumar

Correlatore:
UMR Lagrange, Observatoire de la Côte d'Azur
Dr. Lionel Bigot

Anno Accademico 2023/2024

Erasmus Mundus Master
in Astrophysics and Space Science

Master Thesis

The Sun as a star: Investigating the Influence of Solar Activity Features on Radial Velocity Signals

Supervisors:

Prof. Francesco Berrilli
Physics Dept., University of Rome (Tor Vergata)

Dr. Lionel Bigot
UMR Lagrange, Observatoire de la Côte d'Azur

Author:

Brandon Rajkumar

Academic Year 2023/2024

This Master's thesis is submitted in partial fulfillment of the requirements for the degree FISICA - Curriculum ErasmusMundus as part of a multiple degree awarded in the framework of the Erasmus Mundus Joint Master in Astrophysics and Space Science −− MASS jointly delivered by a Consortium of four Universities: Tor Vergata University of Rome, University of Belgrade, University of Bremen, and Université Cote d'Azur, regulated by the MASS Consortium Agreement and funded by the EU under the call ERASMUS-EDU-2021-PEX-EMJM-MOB.

Table of Contents

The Sun as a star: Investigating the Influence of Solar Activity Features on Radial Velocity Signals

Brandon Rajkumar

Abstract

The search for small, Earth-like planets in the habitable zone of sun-like stars using the radial velocity (RV) method is limited by stellar activity noise which can hide the exoplanet signal. This thesis investigates the influence of stellar activity features on RV signals using the Sun as a star approach, with the Sun serving as a proxy for understanding stellar activity in sun-like stars. Comparing various solar activity indicators with solar RV data revealed strong correlations between the RV data and several solar activity indicators including the $\log(R'_{HK})$ index, S_{MW} index, daily sunspot numbers, 10.7 cm radio flux, Mg II index, and total solar irradiance (TSI), along with cross-correlation metrics such as the bisector span, FWHM, and contrast suggesting that these indicators strongly reflect variations in RV measurements. Results showed that including multi-proxy data like the 10.7 cm radio flux, Mg II index, and TSI improves the accuracy of stellar models, highlighting the importance of broad wavelength observations for exoplanet searches. This thesis suggests that a multi-proxy approach is necessary to account for stellar activity and improve the reliability of exoplanet detections, thus, applying the techniques highlighted in this study to other sun-like stars can improve the likelihood of finding Earth-like exoplanets in habitable zones.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Search for Exoplanets and the Detection Methods

Humanity has long been interested in searching for life beyond the Earth, with much effort spent on exploring the potential habitability of other solar system bodies. However, the discovery of exoplanets, planets orbiting stars outside our solar system, has opened a new search beyond the solar system allowing for a better understanding of the distribution of life in the universe. Studying the distribution of planet sizes, masses, and orbits as well as the properties of the stars they orbit can also improve our understanding of the habitability of exoplanets. (Catala and Consortium, 2009).

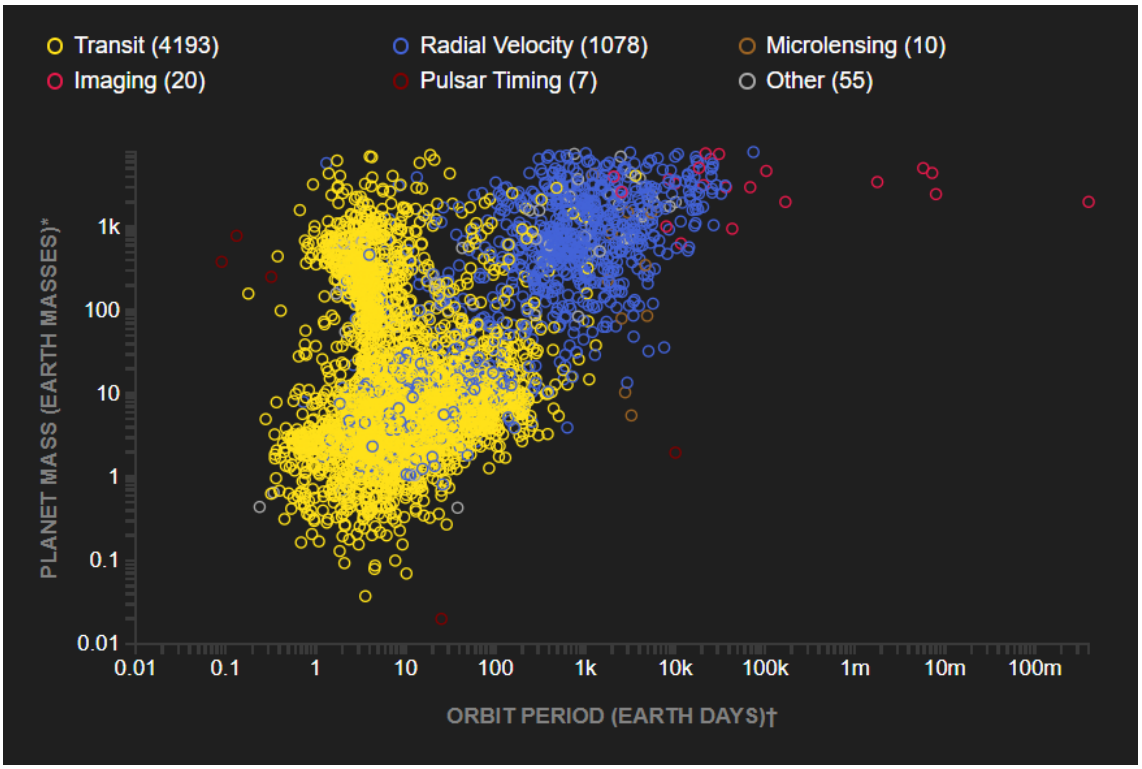


Figure 1.1: Distribution of exoplanet masses and orbital periods discovered as of July 2024 (Brenner et al., 2024).

Since the discovery of the first exoplanet, approximately 5678 confirmed exoplanets have been identified as of July 2024 (Brenner et al., 2024). Figure 1.1 shows the distribution of exoplanet masses and orbital periods discovered using various techniques with the most effective methods being the transit method, the radial velocity (RV) method, direct imaging, and the microlensing method in that order.

The transit method has been used to detect 74.3 % (4216) of the confirmed exoplanets by measuring the dimming of the host star's brightness as the planet passes directly between the star and the observer (Exoplanet Archive, 2023). The orbital period, semi-major axis of the exoplanet's orbit, and the planet-to-star radius ratio can be determined using high-precision photometric monitoring. This method can also be used to detect the presence of satellites and rings of the observed exoplanets (Catala and Consortium, 2009) and allows for the study of multiple planets in a single planetary system as they orbit the host star Winn (2010). However, the transit method requires the planetary system to be aligned, which lowers the probability of detection. The orbital inclination also affects the accurate classification of orbits while binary stars and stellar activity can mimic transiting signals leading to false positive detections (Seager and Mallen-Ornelas, 2003).

In contrast, the RV method has been used to detect 19.2 % (1089) of confirmed exoplanets (Exoplanet Archive, 2023) by measuring the shift of the star's spectral lines due to the gravitational effect of the orbiting planet Mayor and Queloz (1995). Figure 1.2 shows the motion of an exoplanet and its host star as they orbit around a common center of mass. A periodic Doppler shift in the star's spectral lines is observed as the orbiting planet induces a wobble in the host star. The Doppler shift can then be analyzed to derive the RV of the star.

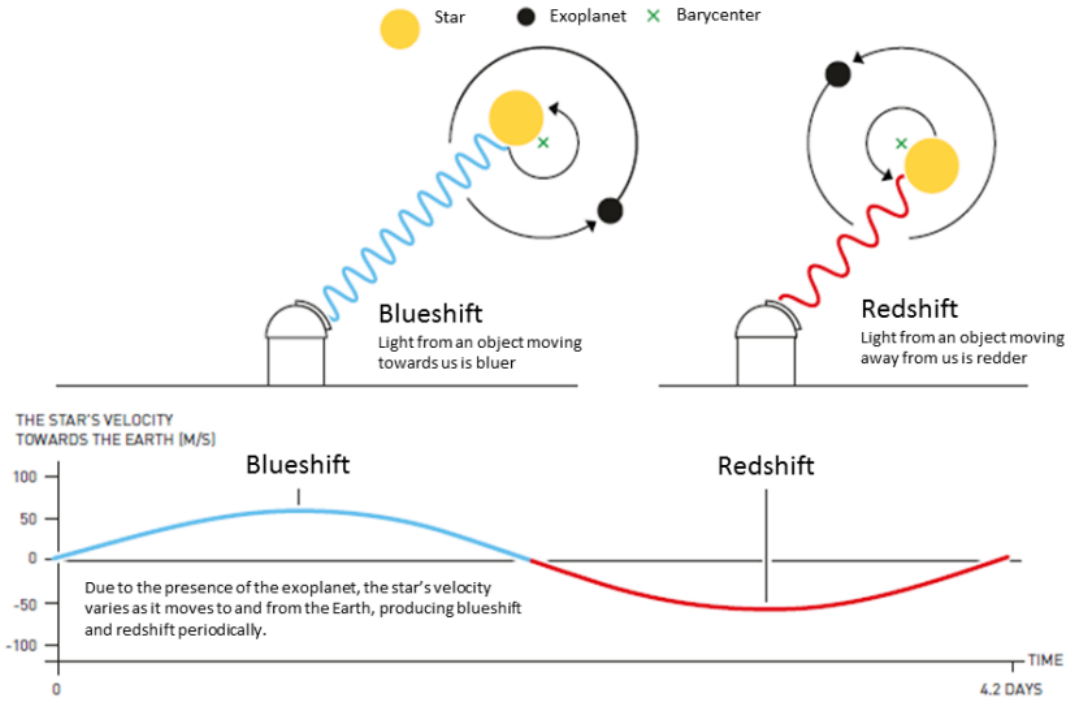


Figure 1.2: Schematic diagram of the RV method for exoplanet detection (Cardenas, 2019).

This method provides a more comprehensive view of planetary systems since it can also detect non-transiting exoplanets (Mayor and Queloz, 1995). It is also more effective at detecting planets with longer orbital periods (Cumming et al., 2008) and can be used to directly measure the mass of a planet, providing information about its composition and density (Butler and Marcy, 1996). However, the RV method is less sensitive to orbits with large inclinations (Fischer et al., 2007) and requires long observational periods to detect subtle periodic shifts in the stellar spectrum (Wright et al., 2012). This method also contains an ambiguity concerning the orbital plane and can be very sensitive to stellar activity which introduces noise into the RV measurements leading to false positive detections. (Dumusque et al., 2011).

Microlensing is another method, that observes the change in the gravitational lensing effect caused by a planet orbiting the lensing star. It has been used to discover 3.9 % (222) of confirmed exoplanets (Exoplanet Archive, 2023). This method is not affected by stellar activity (Poindexter et al., 2005) and is sensitive to low-mass planets (Gaudi and Gould, 1997). It does not depend on long-term observations and can also provide information on planets with larger orbits since it can probe farther away from the host star (Gaudi, 2011). However, microlensing events are rare which reduces the probability of detection. They also limit the information that can be obtained from the planetary system since the events are temporally short (Gaudi, 2000). This method is unable to confirm individual planets and also requires complex models which can be computationally demanding (Bennett et al., 2016).

The direct imaging method uses chronography to reduce the glare of the host star allowing for the direct imaging of the planet. It has been used to discover 1.4 % (82) of confirmed exoplanets (Exoplanet Archive, 2023). It is particularly useful for detecting planetary systems with multiple planets as well as young, self-luminous planets (Aime and Vakili, 2006). This method is also used in the spectroscopic characterization of exoplanet atmospheres and to determine the orbital parameters such as orbital inclination or its distance from the host star. Direct imaging is, therefore, a valuable tool for studying the formation and evolution of planets (Currie et al., 2022), however, it can be limited to young and nearby systems with larger planets further away from the host stars due to the challenges of brightness contrast between the host star and planet (Quanz et al., 2015). It can also be challenging to detect planets with short orbital periods since the direct imaging method requires long exposure times (Currie et al., 2022).

Each of these methods has advantages and disadvantages, however, the pursuit of exoplanet discovery and characterization continues to drive developments in technology and analytical techniques, expanding our understanding of planetary evolution and their distribution in the universe. Figure 1.1 shows that while massive exoplanets have been well explored, Earth-mass exoplanets and lower have only been discovered with short orbital periods of under 100 Earth days leaving a gap in our understanding of exoplanet distribution. This gap is a direct result of the bias and limits inherent in the methods and technology used to detect exoplanets. Some of these limits can be overcome with improved methods and technology as well as a better understanding of the data collected allowing us to obtain a more realistic distribution of exoplanets in the universe.

Additionally, the search for Earth-like planets orbiting sun-like stars is of particular interest since comparing properties such as orbital distance, size, and composition to our own Sun-Earth system provides the highest probability of finding habitable environments (Seager and Mallen-Ornelas, 2003).

1.2 The Radial Velocity Method

The RV method is particularly advantageous when searching for Earth-like planets around sun-like stars. As previously mentioned, the RV method is sensitive to planets with longer orbital periods which can help identify planets in the habitable zone of sun-like stars (Cumming et al., 2008). It can also detect non-transiting planets increasing the probability of detection (Mayor and Queloz, 1995). The ability to directly measure the planet's mass also provides a valuable constraint on its composition which can provide information on its habitability (Butler and Marcy, 1996).

Advancements in RV techniques such as improved instrumentation and data analysis methods have also significantly enhanced the precision and sensitivity of RV measurements enabling the detection of smaller, Earth-like exoplanets with greater accuracy (Fischer et al., 2016; Dumusque et al., 2015). The RV method is also less sensitive to stellar activity when compared to other detection methods which reduces the likelihood of false positives and increases the reliability of detections (Dumusque et al., 2011, 2015).

Additionally, the RV method complements other detection methods such as transit photometry and direct imaging, providing independent measurements of exoplanet properties such as mass and orbital parameters (Mayor and Queloz, 1995; Butler and Marcy, 1996). Combining RV measurements with transit photometry and direct imaging provides comprehensive characterizations of exoplanet systems (Souto et al., 2018; Brahm et al., 2020).

However, one of the main challenges of the RV method when searching for Earth-mass planets orbiting sun-like stars in the habitable zone is due to the apparent RV variations in the host star (Lanza et al., 2016). These variations are primarily attributed to various astrophysical phenomena such as convective blueshift, gravitational redshift, pressure-mode (p-mode) oscillations, and magnetic field activity. They can introduce systematic errors and noise into RV data thus affecting the precision.

Convective blueshift is a result of granulation on a star's surface. Hot plasma rises while cooler plasma descends creating a convective motion, resulting in a net blueshift of the star's spectrum. This can distort RV measurements by introducing velocity shifts that mimic the signal of an orbiting planet (Meunier et al., 2010; Chaplin et al., 2019). In contrast, gravitational redshift is a relativistic effect observed when light escapes the star's gravitational field, resulting in a systematic redshift to the observed spectrum (Madsen et al., 2003). Although this produces a systematic error, with a relatively constant effect for a given star, it should still be considered when making precise RV measurements.

P-mode oscillations are standing acoustic waves in the star's interior which can contribute to RV variability. They can cause the star's surface to oscillate up and down leading to periodic changes in the star's RV. Over short timescales, P-mode oscillations can introduce RV variations on the order of several meters per second for

sun-like stars (Dumusque et al., 2011), increasing the difficulty of detecting low-mass planets.

While the effects of p-mode oscillations and convective blueshift can be reduced by averaging observations on the same night (Dumusque et al., 2011) or on consecutive nights (Meunier et al., 2015) respectively, the effects of the magnetic field are much more complicated. The effects of the magnetic field can mimic those of exoplanet orbital periods as seen by Santos et al. (2014), presenting challenges in separating stellar activity and planetary signals.

It is therefore important to better understand and characterize stellar signals as they appear in RV measurements so that they can be disentangled from planetary signals in RV data. However, due to the timescales of these signals which range from minutes to years, standard RV observations do not have the sampling and time baselines needed to characterize these signals making it difficult to understand and remove stellar noise.

One proposed solution is to continuously monitor the Sun where we greatly understand the activity as a function of time and its astrophysical parameters with the same instrument used for exoplanet characterization (Dumusque et al., 2021) such as the High Accuracy Radial velocity Planet Searcher for the Northern hemisphere (HARPS-N) spectrograph. The Sun is also an ideal test case to study the correlation between RV measurements and stellar activity due to its proximity, detailed observational data, and extensive research about its activity patterns.

Unlike, distant stars, which are observed as point sources, the Sun has been observed in high resolution, allowing surface and atmospheric features to be observed in great detail. When observing other stars, these features that can affect RV measurements are not observable since all information is integrated into a single spectrum (Haywood et al., 2016). The Sun's well-studied 11-year activity cycle can also provide more accurate corrections for stellar noise in RV measurements by serving as a model to understand the magnetic activity cycle of other sun-like stars (Dumusque et al., 2011). Therefore, the sun and its various solar activity indicators can serve as an ideal proxy for sun-like stars and various stellar activity indicators.

1.3 Solar/Stellar Activity Indicators

The Sun is a G-type main-sequence star with distinct spectral features that are used to understand its characteristics when searching for exoplanets around sun-like stars. It is comprised primarily of hydrogen and helium and displays absorption lines across a wide range of wavelengths which provides valuable information on its composition, temperature, and atmospheric dynamics (Basu and Antia, 2008).

One of the key spectral features of the Sun includes the Fraunhofer lines which are absorption lines caused by the presence of hydrogen, helium, calcium, and iron in the solar atmosphere. The chemical composition and abundance of elements in the Sun or sun-like stars can be determined by analyzing the positions and intensities of these lines (Gray, 2021).

Another important feature of sun-like stars is the presence of Doppler shifts on the stellar spectrum due to the star's rotation and oscillatory motions, which provide additional information about its dynamics. This results in a periodic variation in the position of the spectral lines referred to as rotational and p-mode velocity variations which can affect the RV measurements of sun-like stars (Gray, 2021). As previously

mentioned, understanding and characterizing these velocity variations are important for distinguishing between stellar-induced and planetary signals in the RV data.

Additionally, solar activity phenomena such as magnetic activity cycles, sunspots, and solar flares can contribute to spectral variations in the shape and intensity of spectral lines further affecting RV measurements. Similar activity on sun-like stars can further complicate the search for planetary signals (Foukal, 2008). Understanding solar activity can help interpret RV data obtained from sun-like stars allowing for the distinction between stellar and planetary signals (Dumusque et al., 2015).

Sunspots are temporary phenomena in the Sun's photosphere associated with strong magnetic fields and are one of the most prominent solar activity indicators (Rajkumar et al., 2017). They can cause localized changes in the Sun's brightness and induce RV variations due to their influence on the convective motions and magnetic activity on the solar surface (Dumusque et al., 2011). Chromospheric emission lines such as Hα and Ca II H and K lines are also important solar activity indicators due to their sensitivity to magnetic activity. They can also serve as proxies for stellar activity levels (Meunier et al., 2015).

Other solar phenomena such as solar flares, prominences, and coronal mass ejections (CMEs) also influence the Sun's atmosphere and magnetic fields thus impacting its RV measurements. Heating and expansion of the Sun's atmosphere due to the sudden release of energy from solar flares can lead to changes in the density and velocity of plasma within the solar atmosphere (Hannah and Kontar, 2013). Similarly, prominences, which are large bright loops of plasma suspended within the solar atmosphere, can alter the dynamics of the surrounding plasma and introduce disturbances in the solar magnetic field (Schmieder et al., 2013). Additionally, CMEs can induce fluctuations in the Sun's magnetic field strength and can cause perturbations in the solar wind Kilpua et al. (2014). Understanding the complex relationship between these solar events and the RV variations they cause is important for disentangling stellar-induced and planetary signals around sun-like stars.

Monitoring the occurrences of these phenomena and their characteristics can improve our understanding of their impact on the Sun's atmospheric dynamics and magnetic fields which influence RV measurements. This knowledge can then be applied to the analysis of RV data of sun-like stars providing more accurate interpretations and improving the reliability of exoplanet detections around sun-like stars.

1.4 Overview of Thesis

This Master's thesis aims to investigate the influence of various solar/stellar activity features on RV signals. While several stellar activity proxies are already used to separate stellar-induced and planetary signals, increasing the number of proxies can improve the accuracy and confidence of this method, thus improving the reliability of exoplanet detections. Using the Sun as a star approach, solar RV measurements are compared to various solar activity indicators. Indicators that show good correlations to the solar RV are then used as proxies to model the solar RV and are compared to the actual RV measurements.

Chapter 2 describes the data used in this thesis, how the data is collected, and the instrumentation used to collect the data (where necessary). It also discusses the relationship of the data as a solar activity indication, the solar phenomena it

describes, and its potential impact on stellar spectra and RV measurements. This chapter also describes how the data is prepared for analysis, the methods used to ensure that the various data sets are compatible, and the methods used to fill gaps in the data.

Chapter 3 describes the methods of analysis and discusses the results obtained before moving on to chapter 4 which summarises the main results and their implications for exoplanet research. This chapter also highlights the methodological considerations, limitations, and possible future directions of this work. Finally, at the end of chapter 4, the internship carried out as a preparation for this thesis is described and discussed.

Chapter 2

Instrumentation and Data

To understand how solar proxies appear in RV observations, it is important to understand the observational data as well as how the data is collected and processed. This will allow for the distinction of features due to the instrument, the data collection method, the processing method, or the solar phenomena themselves.

This analysis utilizes data obtained from several instruments and different methods, thus it is important to understand them fully before any comparison or correlation is made. This chapter will describe the various data used in the analysis as well as the instruments and methods used to obtain the data. It will also describe how the data is prepared for analysis.

2.1 Solar Proxies

2.1.1 Daily sunspot number

The daily total sunspot number (Day_SSnum) is a measure of the number and size of sunspots visible on the solar surface each day. It is an important index for quantifying solar activity and is typically used for understanding solar dynamics (Hathaway, 2015), space weather (Pesnell, 2012), and climate studies. The Day_SSnum data used in this study was obtained from the Royal Observatory of Belgium's Solar Influences Data Analysis Center (SIDC). The SIDC uses the standard definition of the Wolf number, W_s to calculate the sunspot number (SN) for each individual observer:

$$W_s = k(10N_g + N_s), \tag{2.1}$$

where k is a scaling coefficient specific to each observer which accounts for observational conditions and equipment differences, N_g is the number of sunspot groups, and N_s is the total number of sunspots (Clette and Lefèvre, 2016). As of September 2011, a major revision of the Day_SSnum series was undertaken. The details are given by Clette and Lefèvre (2016), however, for this analysis only the Day_SSnum recorded between July 16, 2015, and July 18, 2018, was used.

As previously mentioned, sunspots are a photospheric feature of emerging magnetic field activity (Rajkumar and Haque, 2020) and therefore affect RV measurements. Higher magnetic activity associated with increased sunspot activity can induce RV variations due to the suppression of convective blueshift in active regions (Dumusque et al., 2015). Figure 2.1 shows the time series for the Day_SSnum.

As expected, as we approach the solar minimum, less magnetic activity leads to a decrease in Day_SSnum with some days having no spots on the solar surface.

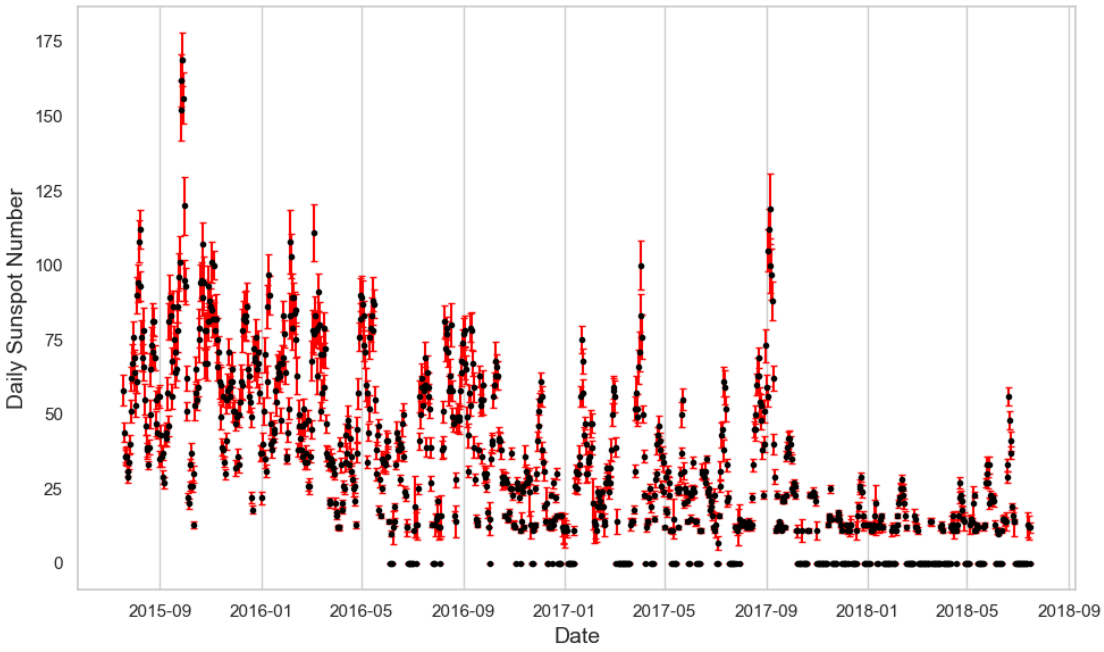


Figure 2.1: Time series for daily sunspot numbers with standard deviation.

2.1.2 Solar Flare Index

As mentioned, solar flares are sudden and intense bursts of radiation from the solar surface. Flares are suddenly released when magnetic energy has built up in the solar atmosphere and can release energy across the entire electromagnetic spectrum from radio waves to gamma rays (Emslie et al., 2012). While there are various ways to classify solar flares, this study uses data obtained from the Solar H-alpha Flare Index available via the National Centers for Environmental Information (NCEI) of the National Oceanic and Atmospheric Administration (NOAA).

The flare index data used was calculated by T. Atac and A. Ozguc from Bogazici University Kandilli Observatory, Istanbul, Turkey using the method introduced by Kleczek (1952). Kleczek's method of calculating the solar flare index, Q, uses Equation 2.2 to quantify the daily flare activity:

$$Q = i \times t, \tag{2.2}$$

where i is the intensity coefficient of the flare as seen in Table 2.1 and t is the duration of the flare in minutes (Kleczek, 1952). This equation assumes that the product of duration and intensity gives an estimate of the energy emitted by the flares. The daily calculations were determined by summing the index values for the northern and southern hemispheres as well as the total surface before dividing by the total observation time for that day (Ataç, 1987; Özgüç and Atac, 1989).

Table 2.1: Intensity coefficient i as described by Kleczek (1952) for determining solar flare index (Özgüç et al., 2003).

Class	i
SF, SN, SB	0.5
1F, 1N	1.0
1B	1.5
2F, 2N	2.0
2B	2.5
3N, 3F, 4F	3.0
3B, 4N	3.5
4B	4.0

Table 2.2: Classification of flare area (Kennewell, 2024).

Importance (Class)	**Corrected Area** (Millionths of Hemisphere)
0 (S)	$10 \leq$ area < 100
1	$100 \leq$ area < 250
2	$250 \leq$ area < 600
3	$600 \leq$ area < 1200
4	area ≥ 1200

Table 2.3: Classification of flare brightness (Kennewell, 2024).

Designator (Class)	**Brightness (% of Background)**	**Surrogate (total bandwidth)**
F (Faint)	$160\% \leq$ brightness $< 260\%$	0.08nm $\leq$ bandwidth $<$ 0.12 nm
N (Normal)	$260\% \leq$ brightness $< 360\%$	0.12nm $\leq$ bandwidth $<$ 0.2 nm
B (Brilliant)	brightness $\geq 360\%$	bandwidth $\geq$ 0.2 nm

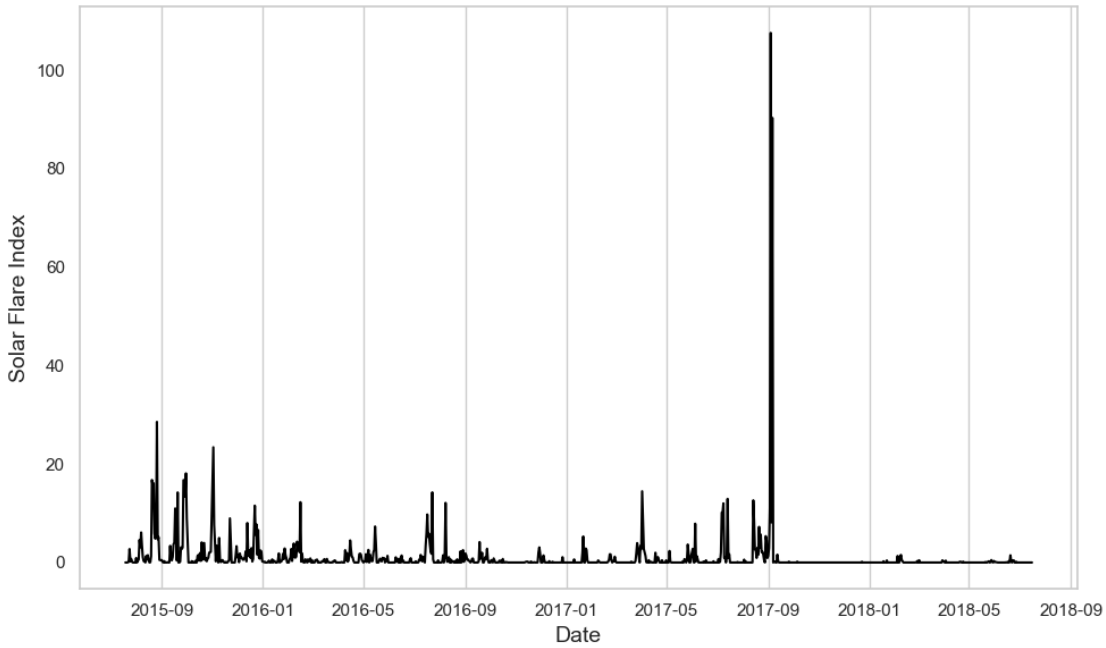


Figure 2.2: Time series for solar flare index.

Intense periods of solar activity, when there are many solar flare occurrences, are represented by a high solar flare index value. Solar flares also release significant amounts of energy into the stellar atmosphere causing localized heating and ionization which can lead to transient spectral line shifts and variations that affect RV measurements (Aigrain et al., 2016).

Figure 2.2 shows the solar flare index time series. We see that solar flare activity is not as frequent as other activity indicators however, it may still reveal a significant impact on the RV measurements.

2.1.3 10.7 cm Radio Flux

The 10.7 cm radio flux (2800 MHz) or radio flux, is a measure of solar radio emission originating from the Sun's chromosphere and corona. The overall level of solar activity including sunspots and active regions influences this emission. It is often used to monitor the solar cycle and is one of the most widely used indicators of solar activity (Tapping, 2013).

The radio flux data used in this study were obtained from the Canadian Space Weather Forecast Center. The flux measurements are usually taken three times daily by the Dominion Radio Astrophysical Observatory (DRAO). From March to October, measurements are made at 1700, 2000 (local noon), and 2300 universal time (UT) while for the rest of the year, they are made at 1800, 2000, and 2200 UT. The values are then averaged to obtain the daily observed value.

However, this value is modulated by both solar activity and the changing distance between the Earth and Sun, therefore an adjusted value is provided after correcting for variations in the Earth-Sun distance. This adjusted value of the 10.7 cm radio flux (Adj_flux) was used for this analysis and is given in solar flux units (1 sfu = 10^{-22} Wm^{-2} Hz^{-1}).

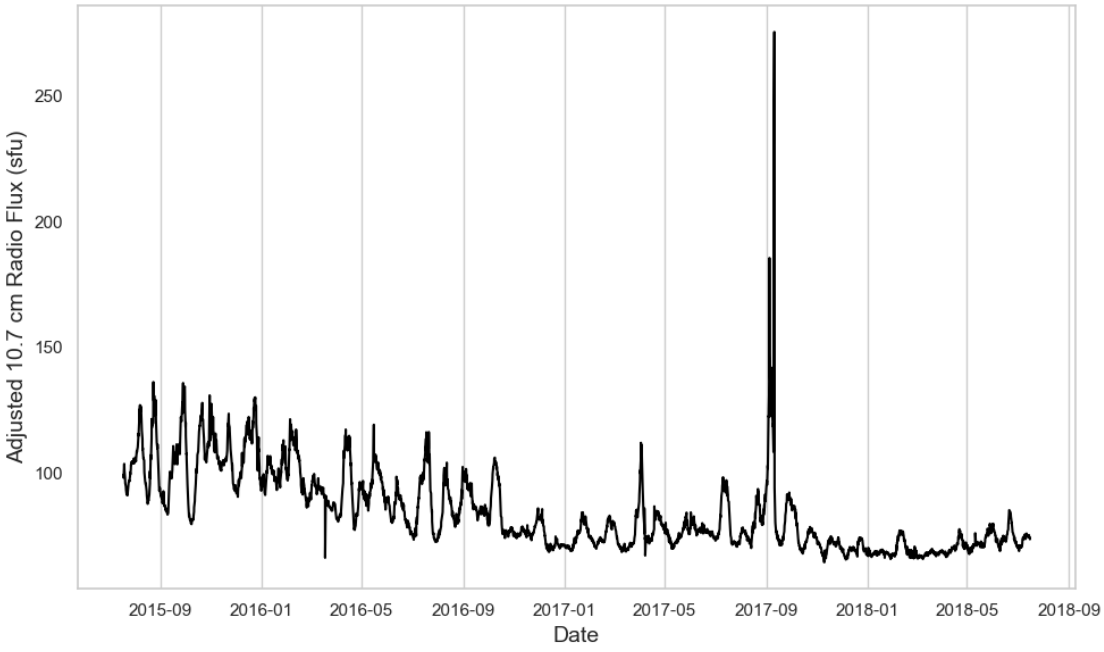


Figure 2.3: Time series for 10.7 cm radio flux

As mentioned, the radio flux is a good indicator of ionization levels in the solar corona with more active regions and increased solar activity generally leading to higher values. It also follows the solar cycle closely making it particularly useful for studying the long-term trends of solar activity (Covington, 1969). For this study, however, variations in the radio flux can indicate changes in the solar atmosphere's thermal and non-thermal emission processes that can affect the behavior and formation of spectral lines, potentially introducing variability in the RV measurements (Haywood et al., 2014). Greater magnetic activity which can lead to distortions in the RV signal due to suppression of convective blueshift in active regions, is usually associated with an increase in radio flux values (Dumusque et al., 2015).

Figure 2.3 shows the 10.7 cm radio flux time series. Similar to the Day_SSnum, we observe a decrease in radio flux activity as we approach the solar minimum. However, we also observe a spike in radio flux similar to that seen in the solar flare index.

2.1.4 Mg II index

The Mg II index (MGII_In) comes from the emission in the core of the Magnesium II (Mg II) h and k lines around 280 nm and is a measure of solar ultraviolet (UV) radiation. The Mg II index serves as a valuable proxy for solar chromospheric activity since this emission originates from the solar chromosphere. It is widely used in studies of solar variability and its effects on the Earth's atmosphere (Viereck et al., 2001).

The Mg II index data used in this study was obtained from the Institute of Environmental Physics (IUP) at the University of Bremen. It is derived from satellite measurements of the solar UV spectrum and is calculated as the ratio between the Mg II h and k line core flux and the nearby continuum flux. The composite Mg II index which was used is generated by combining data from the Global Ozone Monitoring Experiment (GOME), the Solar Backscatter Ultraviolet Radiometer (SBUV), and the Ozone Monitoring Instrument (OMI) among others. This ensures consistent and continuous long-term recording of the solar UV variability (Snow et al., 2014).

As mentioned, the Mg II index is a sensitive indicator of chromospheric activity which correlates well with other solar activity indicators such as Day_SSnum and the 10.7 cm radio flux. It can provide insights into the solar UV radiation changes since it follows the 11-year solar cycle and has proven crucial for understanding the effects of solar variability on the Earth's atmosphere (Heath and Schlesinger, 1986). Since the Mg II index can indicate changes in the chromospheric activity of the Sun, it can prove useful in characterizing changes in the RV measurements of sun-like stars caused by variations in their chromosphere. A high Mg II index value indicates increased chromospheric activity and can correlate to increased RV variability which can negatively affect exoplanet detection and characterizations (Dumusque et al., 2015).

Similar to the other solar activity indicators, we observe a decrease in the Mg II index as we approach the solar minimum, seen in Figure 2.4.

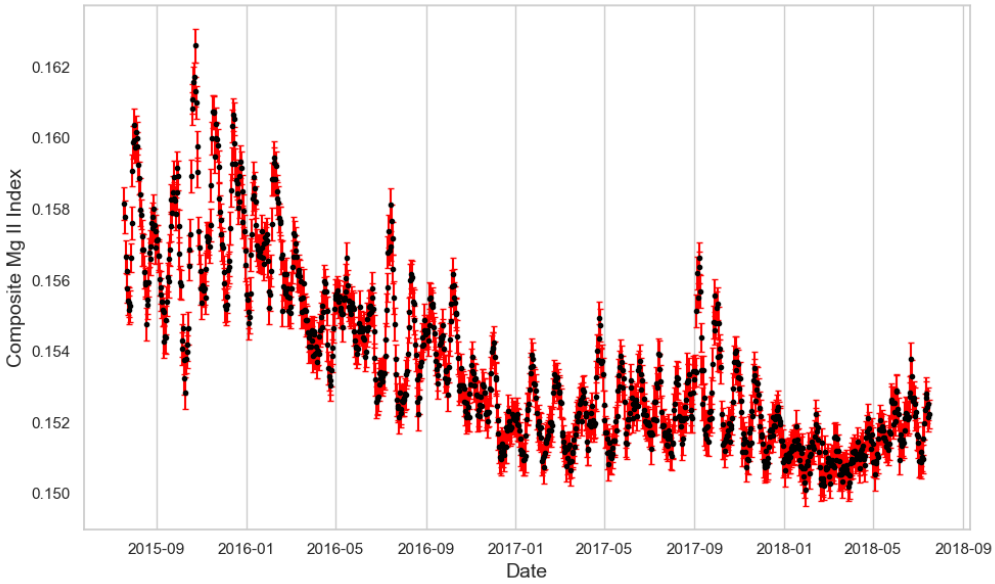


Figure 2.4: Time series for the Mg II index

2.1.5 Total Solar Irradiance

Total Solar Irradiance (TSI) is a representation of the total solar radiation output across all wavelengths. It is a measure of the solar energy received per unit area at the top of the Earth's atmosphere and has served as a fundamental parameter in understanding the Earth's climate making it crucial for climate modeling and solar physics (Kopp et al., 2016).

The data used in this study was obtained from the Solar Radiation and Climate Experiment (SORCE) mission, managed by the Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado Boulder. The SORCE mission uses the Total Irradiance Monitor (TIM) to measure the TSI with high precision and accuracy correcting for instrument drifts, degradation, and other systematic effects. The data is collected continuously and processed to provide average values every 6 hours along with the daily average values (Kopp and Lawrence, 2005). While a higher cadence would allow for short-term variations to be monitored, the 24-hour cadence was used to allow for comparison with other datasets in this study.

The TSI at 1 Astronomical Unit (AU) (TSI_1au) shows the total solar radiation output at the average distance of the Earth from the Sun while the TSI at Earth's distance (TSI_ed) provides the total solar radiation received by the Earth at its specific distance from the Sun. The TSI_ed considers the variability in solar radiation due to Earth's elliptical orbit (Kopp, 2019), however, for this study, the normalized TSI_1au is used for the analysis as it provides a better representation of the modulation of the TSI due to solar activity.

As mentioned, the TSI is a direct measure of the Sun's total energy output, therefore variations in the TSI can indicate changes in solar activity such as the emergence of sunspots or faculae on the solar surface which can modulate the solar irradiance. Changes in TSI have been linked to climate variations on Earth with higher TSI values leading to warming and vice-versa. Understanding TSI variations has been essential for climate studies and predicting climate trends (Lean et al., 1992). Variations in the TSI have also been correlated to stellar activity indica-

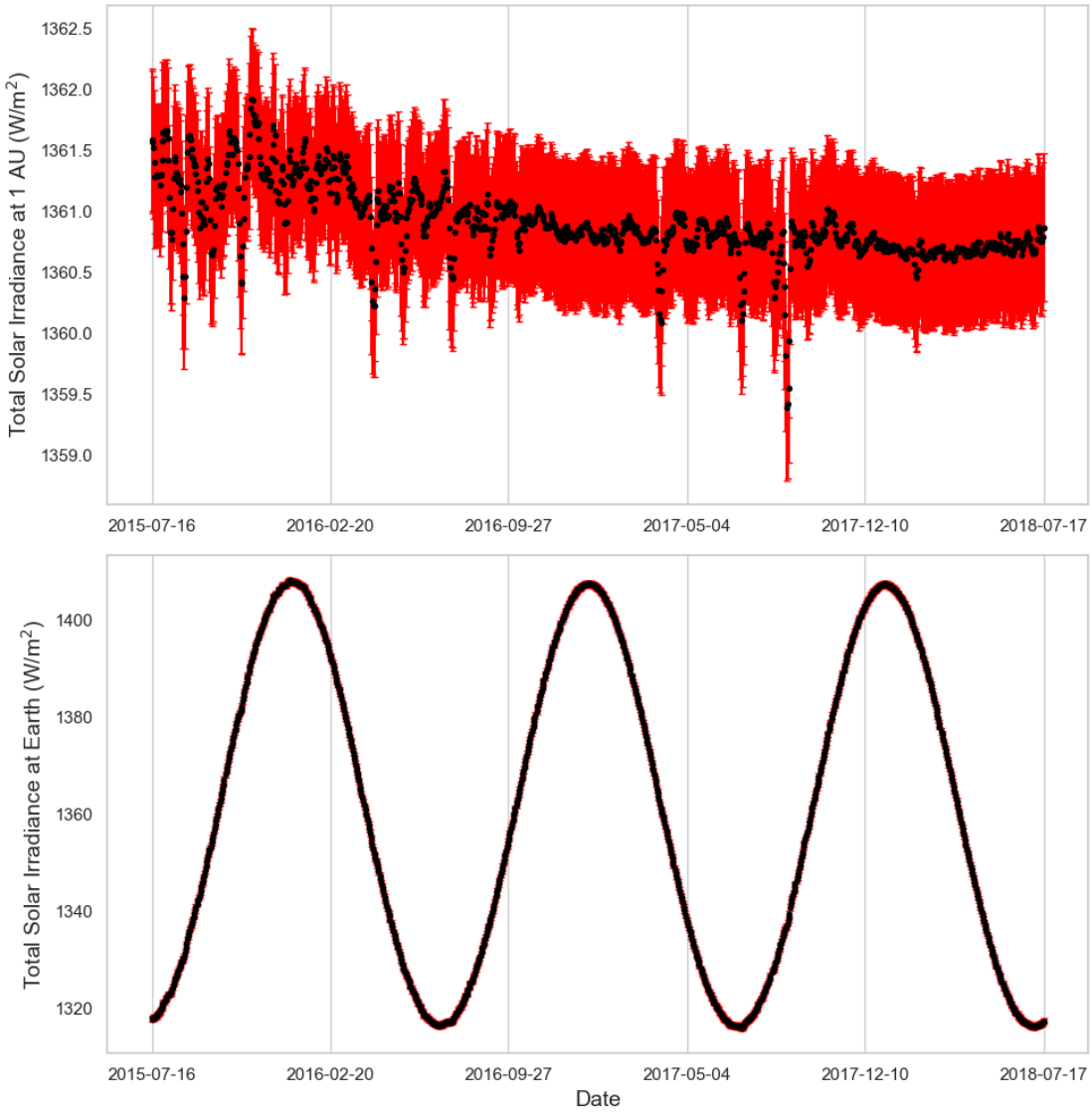


Figure 2.5: Time series for the TSI at 1 AU and Earth distance.

tors such as sunspots and magnetic activity which can affect RV measurements (Dumusque et al., 2015). Changes in the TSI can alter the stellar photosphere's structure and temperature. This can affect the formation and behavior of spectral lines used in RV measurements, potentially introducing variability (Haywood et al., 2014).

As seen in Figure 2.5, the TSI at 1AU shows a general decrease as we approach the solar minimum while the TSI at Earth distance is strongly modulated by the Earth's rotation around the Sun which can introduce a periodic signal in RV measurements taken on Earth with a frequency similar to that of an orbiting planet.

2.2 HARPS-N

High Accuracy Radial Velocity Planet Searcher (HARPS) is an echelle spectrograph attached to the ESO 3.6-meter telescope in La Silla, Chile used to detect exoplanets using high-accuracy RV measurements (Mayor et al., 2009). The High Accuracy Radial Velocity Planet Searcher for the Northern Hemisphere (HARPS-N) is a similar high-precision RV instrument installed on the 3.58 m Telescopio Nazionale Galileo (TNG) Telescope at Roque de los Muchachos Observatory in La Palma, Spain which is also used to search for exoplanets. However, during the day, light from the solar disk is integrated and passed to the spectrometer providing full-disk RV measurements of the sun(Al Moulla et al., 2023). This method, referred to as Sun as a star observation, allows comparisons between the full-disk RV measurements of the Sun and higher resolution observations (Dumusque et al., 2015).

HARPS observed its first light in 2003 while HARPS-N was commissioned in 2012. The HARPS-N's northern location also allows access to the targets of the K2 and TESS missions. While the main objective of HARPS-N has been to discover and characterize terrestrial exoplanets by achieving long-term 1 m/s median RV precision on V $<$ 11-mag stars (UniversiteofGeneve, 2014), its success in observing the Sun as a star has led to the installation of the HARPS Experiment for Light Integrated Over the Sun (HELIOS), a solar telescope which feeds to the HARPS instrument (information@eso.org, 2023). Each day, both instruments observe the Sun for as long as possible while its altitude is higher than 10 degrees.

HARPS-N is a fiber-fed, cross-dispersed spectrograph with a resolving power of $R = 115000$ and a spectral range of 390 nm to 690 nm. It is fed calibration and stellar light by two fibers, the reference, and science fibers. The fiber entrance is re-imaged onto a 4k4k CCD by the spectrograph optics, where echelle spectra of 69 orders are formed for each fiber. The spectrograph contains no moving parts and is mounted on a nickel-plated stainless steel mount. Its temperature and pressure are also accurately controlled to avoid spectral drifts (Phillips et al., 2016).

To obtain sub 10 cm/s calibration of the HARPS-N spectrograph, required for detecting terrestrial exoplanets in the habitable zone of sun-like stars, a laser frequency comb (LFC) based calibrator (astro-comb) is used to obtain short-term stability of less than 2 cm/s with a long-term drift of approximately 10 cm/s per day (Phillips et al., 2016). However, it is important to note that this level of precision is not needed for this analysis since the majority of the other data sets used will only provide daily averages. This will be discussed further in later sections.

A custom-built solar telescope is connected to HARPS-N providing a near-continuous stream of disk-integrated solar spectra during the day (Milbourne et al., 2019; Dumusque et al., 2015). This instrument works with the HARPS-N spectrograph to observe the Sun as a star with a temporal coverage of approximately one exposure every 5 minutes and a daily coverage of about 6 hours. The solar telescope converts solar images into an equivalent point source using a 3" lens which feeds an integrating sphere. The sphere scrambles all angular information allowing for observations of the Sun as a star (Milbourne et al., 2019). Reports show that the solar telescope captures the full disk of the Sun with a RV precision below 10 cm/s comparable to independent SDO/HMI images Phillips et al. (2016).

The HARPS-N data used in this analysis was obtained from the Data & Analysis Center for Exoplanets (DACE). It consists of 745 observed days between July 16,

2015, and July 18, 2018, with a cadence of 5.5 min (5 min exposure and 30s readout time). This data was reduced using version 2.2.2 of the Echelle SPectrograph Rocky Exoplanet and Stable Spectroscopic Observations (ESPRESSO) Data Reduction Software (DRS) (Dumusque et al., 2021). It is also important to note that this data contains gaps due to various reasons such as observational constraints or equipment maintenance which will be discussed in future sections. Table 2.4 shows the description of the time series data obtained from the HARPS-N data set used in this analysis.

Table 2.4: Description of the time series data obtained from the HARPS-N data set.

Name	**Description**	**Unit**
RV	RV of the Sun in the heliocentric rest frame, corrected for differential extinction	m s^{-1}
RHK	$\log(R'_{HK})$ calcium activity index	dex
S_{MW}	S Mount Wilson calcium activity index	-
Bis Span	Bisector span of the CCF	m s^{-1}
FWHM	FWHM of the CCF, corrected for the solar ecliptic obliquity and Earth orbit eccentricity	m s^{-1}
Contrast	Contrast of the CCF, corrected for the FWHM correction so that the equivalent width of the CCF is conserved	%

RHK or the $\log(R'_{HK})$ calcium activity index is a measure of chromospheric activity in stars, including the Sun (Vaughan and Preston, 1980). It is derived from the emission-line cores of the Ca II H and K spectral lines. This index is a normalized measure that corrects for the photospheric contribution to the Ca II H and K line emission, providing a better indication of chromospheric activity (Noyes et al., 1984). The RHK is obtained by measuring the flux in the cores of the Ca II H and K lines from the observed spectra. The measurements are then calibrated against a set of standard stars with known chromospheric activity levels (Wright et al., 2004).

The Ca II H and K lines are widely used as a proxy for magnetic activity and rotational period of sun-like stars (Mittag et al., 2018). Dark spots and bright faculae are formed on photospheres of late-type stars due to concentrated magnetic fields that extend to the chromosphere causing additional nonthermal heating. This can lead to the formation of chromospheric magnetic features which cause variability in the Ca II H and K spectral lines (Sowmya et al., 2023). RHK can provide insights into the magnetic activity of a star which affects the observed RV by inducing changes in the stellar surface. Strong magnetic activity is indicated by high RHK values and correlates to increased RV variability (Dumusque et al., 2015).

The S Mount Wilson index or S_{MW} is the precursor of the RHK index and another measure of chromospheric activity derived from the Mount Wilson Observatory's long-term stellar activity monitoring program. The S_{MW} can be quantified by comparing the relative strength of the Ca II H and K line emission to the nearby continuum (Martin et al., 2017). The S_{MW} is calculated using the measured flux in narrow bands centered on the Ca II H and K lines normalized by the flux of two nearby reference bands and scaled using a calibration derived from the Mount Wilson survey data (Vaughan and Preston, 1980).

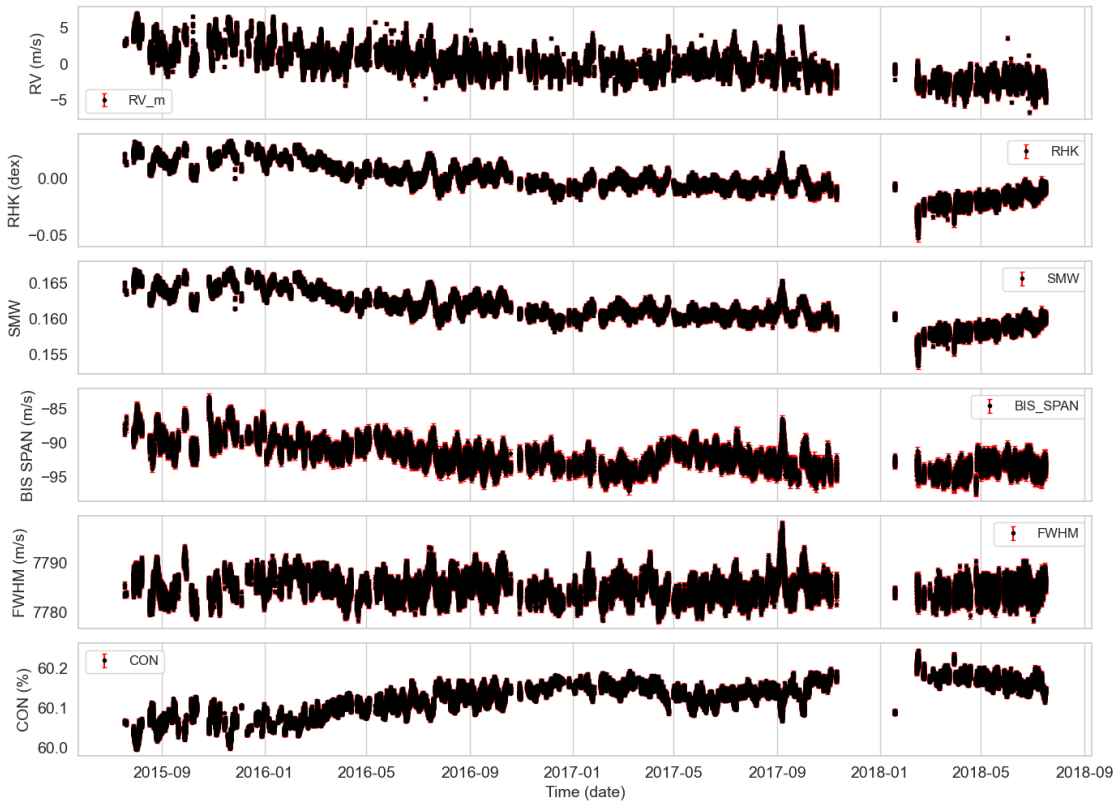


Figure 2.6: Time series for HARPS-N data set.

Unlike the RHK index, the S_{MW} index provides a proxy for photospheric plus chromospheric Ca II strength and also serves as an indicator of magnetic field activity (Tinney et al., 2002). It is used to identify periods of high magnetic activity which can affect RV measurements by inducing spectral line variations and asymmetries.

The cross-correlation function (CCF) is a mathematical tool that measures the correlation between two spectra (observed and reference spectra) as a function of velocity shift applied to one of the spectra. The RV of the object corresponds to the peak of the CCF where the velocity shift allows for the observed spectra to best match the reference spectra (Simola et al., 2019).

The shape of the CCF can be distorted by asymmetries in the spectral lines caused by stellar activity which can affect RV measurements. The bisector span or Bis Span (BIS_SPAN) is a measure of this asymmetry (Queloz et al., 2001).

The width of the CCF measured by the full width at half maximum (FWHM) can provide information on broadening mechanisms such as stellar rotation, macro-turbulence, or instrumental effects (Simola et al., 2019). The FWHM is measured by fitting a suitable function to the CCF and measuring the width of the profile at half of its peak height.

The depth or contrast of the CCF also provides information about the stellar activity as it indicates the sharpness and strength of the spectral lines Simola et al. (2019). The contrast is calculated by measuring the difference between the peak and the baseline of the CCF and then normalizing it to the baseline.

Figure 2.6 shows the time series data obtained from HARPS-N. We see the RV time series for the Sun along with the RHK and S_{MW} activity index time series followed by the time series of the components of the CCF. These solar activity

indicators can be used to understand how solar activity affects the RV measurements of the Sun.

2.2.1 Data Preparation

It was first noted that the cadence of the data sets is different. While HARPS-N provides a high-resolution time series with a cadence on the order of 5 minutes, the solar indicator datasets have cadences of 1 day. Although a high temporal resolution is very useful for classifying short-period stellar signals that can later be removed, due to the limits of the other data sets, the daily averages of the HARPS-N RV data were determined as seen in Figure 2.7. For this analysis, the daily averages are sufficient since we are interested in signals that mimic Earth analogs around solar-type stars.

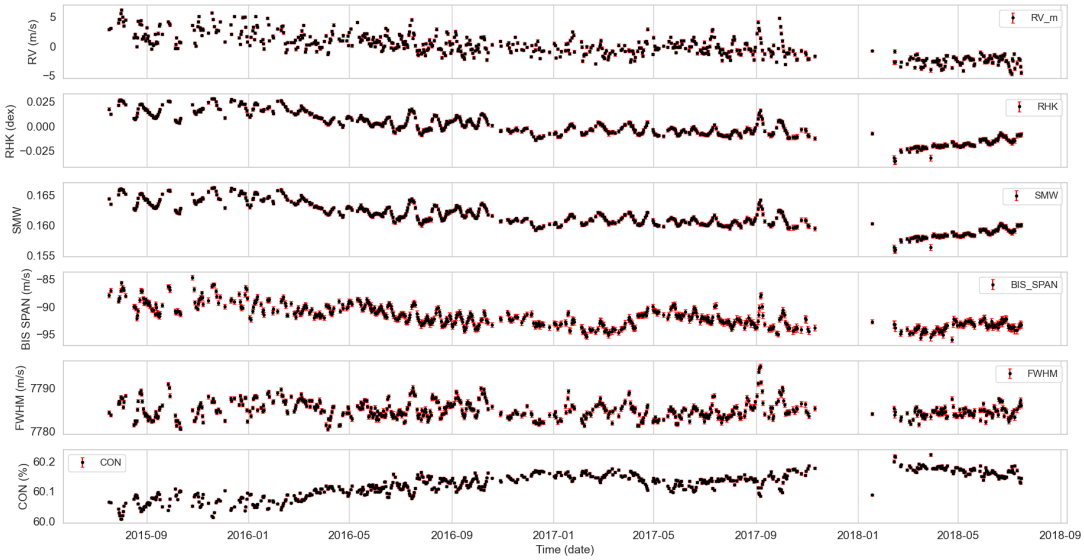


Figure 2.7: Daily Averaged Time Series for HARPS-N data set.

Gaps in the data sets were also observed. This missing data can be caused by various issues such as:

- Poor weather conditions which can affect ground-based observations.
- Scheduled instrument maintenance.
- Technical failures such as mechanical failures, software glitches, or electronic malfunctions can lead to gaps in the data while repairs or re-calibrations are being performed.
- Data rejection during calibration due to poor data quality.

Handling missing data is important in time series analysis since missing data can skew results and affect the correlation analysis. Missing data in time series are usually treated with:

- Forward or backward fill techniques which fill missing values with the previous or the next observed value, respectively (Giobbi, 2023). These methods are

valid when the data changes slowly over time or is assumed to remain constant over some time. It can also be used for short gaps such as daily data where the gap is on the order of a few days (Little and Rubin, 2019).

- Linear regression-based imputation estimates the missing data by creating a linear model between one or more predictor (independent) variables and the (dependent) variable with the missing values by finding a linear relationship between them. It involves training the model on known data before applying it to the data with the gaps. This method assumes that there is a linear relationship between variables with no sudden shifts or discontinuities. It also needs sufficient data for the mode to be trained (Seber and Lee, 2012).

While gaps in the data for the solar activity indicators were limited to a few days, the HARPS-N data contained an approximate two-month gap around December 2017. This was due to damage to the fiber guiding light from the solar telescope to the spectrometer which led to observations with very low S/N. This led to low precisions and the data was rejected during processing (Dumusque et al., 2021).

This study used a combination of forward and backward fill methods to fill short gaps in the solar activity indicator datasets, however for the HARPS-N data set which contained a large gap, the linear regression-based imputation was used which required a predictor variable without missing values. It was decided that HARPS-N RV data simulated with SolAster described in Section 2.2.2 would be used as the predictor variable. Section 2.2.3 describes the linear regression-based reputation in more detail when applied to the HARPS-N dataset.

2.2.2 Simulated Sun as a star RVs: SolAster

Unfortunately, there were no space-based RV measurements of the Sun during the observation period. However, it was possible to estimate disk-integrated RV measurements from high-spatial-resolution data obtained from the Helioseismic and Magnetic Imager (HMI) aboard the Solar Dynamics Observatory (SDO). SolAster is a publicly distributed pipeline created to estimate the Sun as a star disk-integrated RVs from high-spatial-resolution data. This tool was made to study the activity-induced temporal variation of the Sun in great detail allowing the Sun to serve as a proxy for understanding similar processes in other sun-like stars (Ervin et al., 2022).

The high-resolution data products provided by the SDO/HMI include dopplergrams, magnetograms, and continuum filtergrams. Dopplergrams measure the Doppler shift of the solar photospheric lines and provide information about the velocity across the solar disk with a cadence of 45 seconds. Magnetograms capture the line-of-sight magnetic flux density providing maps of the solar surface's magnetic field with a cadence of 45 seconds. With a similar cadence of 45 seconds, continuum filtergrams provide insights into the brightness distribution across the solar disk by providing intensity measurements in the continuum near the photospheric absorption line. The raw HMI data are processed to correct for instrumental effects including flat fielding, cosmic ray removal, and image alignment ensuring accurate and consistent full disk images (Pesnell, 2012; Scherrer et al., 2012).

SolAster determines the disk-integrated RV by integrating the velocity information from Dopplergrams over the entire solar disk. This is done by summing the contribution from all regions of the solar disk while considering the effects of solar

rotation and limb darkening. Next, it models the impact of magnetic activity on the RV signal using continuum images and magnetograms. It considers the suppression of convective blueshift in active regions as well as other magnetic effects that distort the RV measurements. The pipeline also applies corrections for foreshortening, where regions near the solar limb appear compressed due to the projection effect, and limb-darkening, where the solar intensity is decreased toward the limb. Applying these corrections ensures that the RV signal is an accurate representation of the different contributions across the solar disk and that the velocity and intensity distributions are accurately modeled across the disk Ervin et al. (2022).

While SolAster can provide simulated RV measurements with the same cadence of SDO/HMI, 45 seconds, for this study, and due to limited time and resources, a cadence of 1 day was used. Similar to the HARPS-N data, a high temporal resolution would allow the capture of short-term variations in solar activity and their impact on RV measurements, however, other data sets used in this study do not have the luxury of such high temporal resolutions.

It should also be noted that the pipeline includes methods to reduce noise in the RV time series such as filtering out high-frequency noise and correcting for systematic errors which can enhance the precision of the simulated RV measurements. Other errors to consider are gaps in the simulated measurements due to missing SDO/HMI data. If any of the three full disk images is missing or fails to download, no calculation is performed leaving a gap in the simulated data. However, these gaps are limited to a few days and can be easily filled using forward or backward-filling techniques.

SolAster applies a scaling factor to the simulated SDO RV measurements to allow for more accurate comparisons. The HARPS-N scaling factor was used to adjust the amplitude of the RV variations, making them consistent with the scale and sensitivity of the HARPS-N measurements. This scaling factor was obtained by comparing the output RV measurements from SolAster to actual HARPS-N data. Table 2.5 shows the resulting outputs of the SolAster pipeline along with a short description and their respective units. However, only the rv_model data will be used as the predictor variable when performing the linear regression-based imputation of the HARPS-N data.

SolAster was developed to provide a robust method for simulating RV measurements from full-disk SDO images. It can be used to better understand the relationship between RV variations and solar magnetic activity which is important for improving the detection of Earth-like planets around sun-like stars. By analyzing the convective and photometric velocity components derived from the SDO/HMI data, SolAster allows for correlating solar magnetic activity indicators with RV variations (Ervin et al., 2022). By comparing spaced-based (SolAster) to ground-based (HARPS-N) RV measurements, systematic differences in the HARPS-N data can be identified and corrected leading to more accurate and reliable RV data (Kopp et al., 2016). Figure 2.8 shows the time series for rv_model data estimated from the SDO data using SolAster.

Table 2.5: Description of the simulated data obtained from SolAster.

Result	**Description**	**Unit**
date_obs	Observation timestamp in UT from SDO metadata.	Date-Time
date_jd	Calculated JD from observation timestamp.	JD
rv_model	Modeled RV variation from Doppler shifts across the solar disk.	m/s
v_quiet	Velocity component from quiet Sun regions, measured from Doppler shifts in minimal activity areas.	m/s
v_disk	Overall disk velocity calculated by averaging Doppler shifts across the solar disk.	m/s
v_phot	Photometric velocity component from intensity variations correlated with Doppler shifts.	m/s
v_conv	Convective velocity derived from plasma motion models within the Sun's convective zone.	m/s
f_bright	Filling factor for bright regions determined by segmenting faculae and calculating their fractional area.	%
f_spot	Filling factor for sunspots determined by segmenting dark regions and calculating their fractional area.	%
f	Total filling factor for all observed features on the solar disk.	%
Bobs	Unsigned magnetic flux calculated from magnetogram data, representing total magnetic field strength.	Gauss
vphot_bright	Photometric velocity component from bright regions' Doppler shifts.	m/s
vphot_spot	Photometric velocity component from sunspots' Doppler shifts.	m/s
f_small	Filling factor for small sunspots and faculae determined by segmenting and measuring their fractional area.	%
f_large	Filling factor for large sunspots and ARs determined by segmenting and measuring their fractional area.	%
f_network	Filling factor for the magnetic network determined by identifying dense magnetic field line regions.	%
f_plage	Filling factor for plage regions determined by segmenting bright areas around sunspots.	%
quiet_flux	Magnetic flux for quiet-Sun regions calculated from magnetogram data.	Gauss
ar_flux	Magnetic flux for ARs calculated from magnetogram data.	Gauss
conv_flux	Magnetic flux for large ARs calculated from magnetogram data.	Gauss
pol_flux	Polarized magnetic flux derived from polarization analysis of light emissions.	Gauss
pol_conv_flux	Polarized magnetic flux from large ARs calculated from polarization data.	Gauss
vconv_quiet	Convective velocity for quiet regions modeled from Doppler measurements.	m/s
vconv_large	Convective velocity for large ARs modeled from Doppler measurements.	m/s
vconv_small	Convective velocity for small ARs modeled from Doppler measurements.	m/s

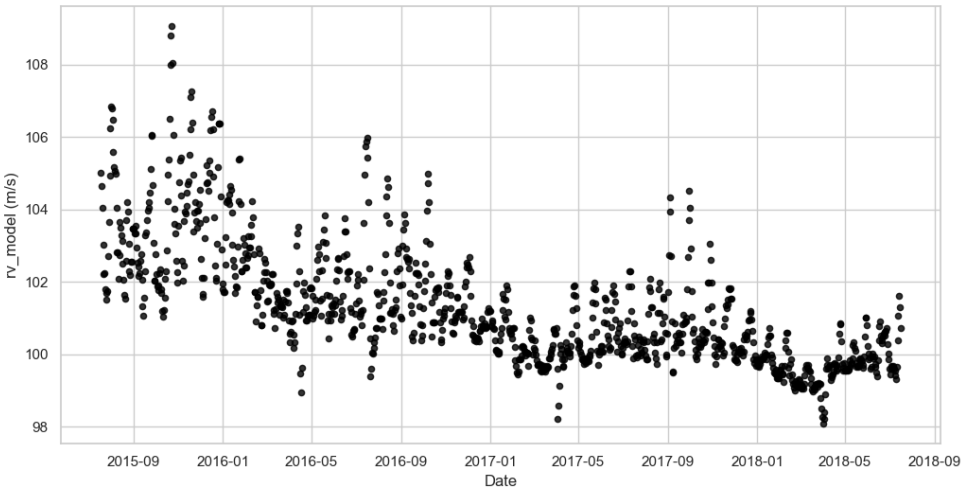


Figure 2.8: Time series for rv_model derived from SDO data using SolAster.

2.2.3 Linear Regression-based imputation applied to HARPS-N data

As previously mentioned, the linear regression-based imputation was used to fill the missing values in the HARPS-N data set. This method uses the relationship between the predictor or reference variable, (X_{ref} = rv_model), and the target variables, y, to predict missing values.

The target variables that contained missing data were first identified. These are the only columns that would be subjected to the imputation process. The reference variable is then scaled, X_{scaled}, to match the statistical properties (mean, μ, and standard deviation, σ) of the target column as seen in Equation 2.3. This ensures that the predicted values are on a comparable scale with the original data of the target column.

$$X_{scaled} = \left(X_{ref} - \mu_{ref}\right) \times \frac{\sigma_y}{\sigma_{ref}} + \mu_y. \tag{2.3}$$

Next, a linear regression model is defined as:

$$y = \beta_0 + \beta_1 X + \epsilon, \tag{2.4}$$

where y is the independent variable, X is the dependent variable, β_0 and β_1 are the intercept and slope coefficients, and ϵ is the error term. The model is then trained using non-missing data from the target column (y = y_{train}). The corresponding values of the scaled reference variable are treated as the dependent variable (X = X_{train}). The model was trained on the available data by minimizing the residual sum of squares (RSS):

$$RSS = \sum_i (y_i - (\beta_0 + \beta_1 X_{train,i}))^2. \tag{2.5}$$

The optimal parameters, $\hat{\beta_0}$ and $\hat{\beta_1}$ were obtained using the least squares method, which gives:

$$\hat{\beta}_1 = \frac{\sum_i (X_{train,i} - \bar{X}_{train})(y_i - \bar{y})}{\sum_i (X_{train,i} - \bar{X}_{train})^2}, \tag{2.6}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}_{train}. \tag{2.7}$$

Once the model was fitted, the scaled reference variable was used to predict the missing values in the target data:

$$\hat{y}_{missing} = \hat{\beta}_0 + \hat{\beta}_1 X_{scaled,missing}. \tag{2.8}$$

Figure 2.9 compares the filled (red) and unfilled (black) HARPS-N data where we see all gaps in the data are filled. It is important to note that the underlying assumption of the linear regression-based imputation is that the reference variable and the target variables share a linear relationship. This assumption is justified since the reference variable used was simulated HARPS-N data derived from SDO images, however, further statistical tests will be performed to observe the effects of filling all datasets.

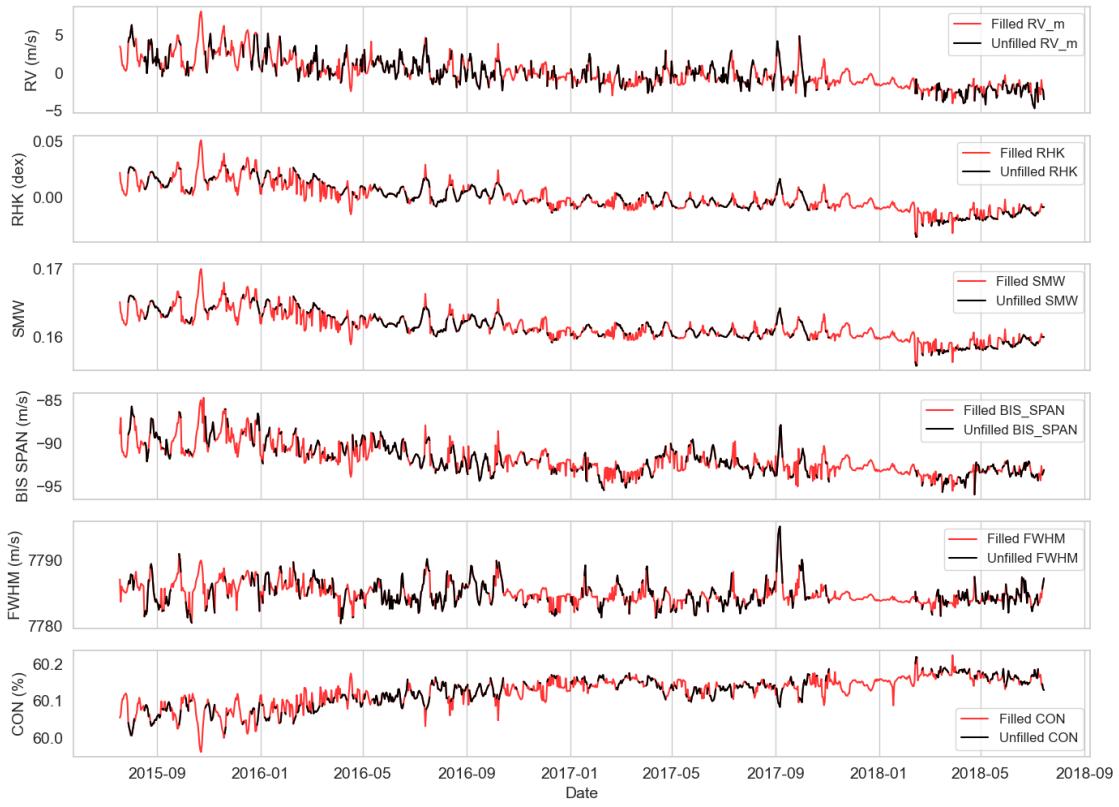


Figure 2.9: Time series comparing filled and unfilled HARPS-N data.

2.2.4 Comparing data before and after filling gaps

As previously mentioned, incorrectly filling the gaps in the data can lead to biases in the analysis performed. It is therefore important to check if filling in the missing values has affected the descriptive statistics of the data sets. Figure 2.9 provides a visual idea of how much data was filled in the HARPS-N dataset. Figure 2.10 compares the filled and unfilled data for the other solar indicators. It was observed that the flare index, radio flux, Mg II index, and TSI had little to no gaps in the data while the Day_SSnum had a few small gaps which needed to be filled.

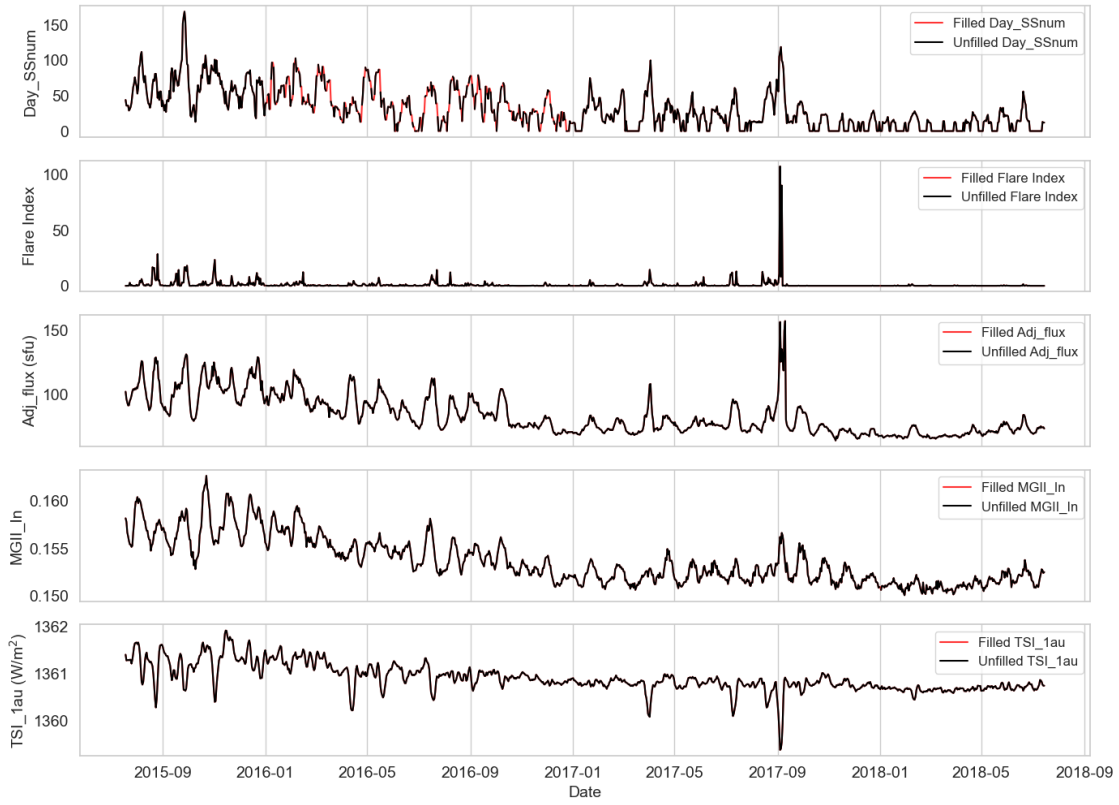


Figure 2.10: Time series comparing filled and unfilled solar indicator data.

Tables 2.6 and 2.7 provide the descriptive statistics of the unfilled and filled data respectively. Boxplots (Figure 2.11) for each solar indicator were then plotted to help visualize this information. From these plots, we see no significant change between the means of the unfilled and filled data. Histograms (Figure 2.12) were also prepared which showed no significant change in the distribution of the data for each solar indicator. This shows that the filled data is a good representation of the observed data and can be used for our analysis as described in chapter 3.

Table 2.6: Descriptive statistics of unfilled data sets.

Unfilled Data								
Indicators	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
RV_m	630	0.03	1.96	-4.67	-1.32	-0.12	1.18	6.28
RHK	630	0	0.01	-0.04	-0.01	0	0.01	0.03
S_{MW}	630	0.16	0	0.16	0.16	0.16	0.16	0.17
BIS_SPAN	630	-91.79	1.94	-95.94	-93.2	-92.05	-90.56	-84.64
FWHM	630	7784.96	2.04	7780.47	7783.57	7784.62	7786.08	7795.17
CON	630	60.13	0.04	60.01	60.1	60.13	60.15	60.22
Day_SSnum	995	31.18	28.17	0	11	25	49	169
Flare Index	1090	1.07	4.92	0	0	0	0.52	107.54
Adj_flux	1093	84.39	15.43	64.57	72.33	79.1	93.97	157.03
MGII_In	1091	0.15	0	0.15	0.15	0.15	0.16	0.16
TSI_1au	1093	1360.91	0.29	1359.39	1360.73	1360.84	1361.05	1361.92

Table 2.7: Descriptive statistics of filled data sets.

Filled Data								
Indicators	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
RV_m	1093	0.02	1.92	-4.67	-1.33	-0.31	1.04	8.06
RHK	1093	0	0.01	-0.04	-0.01	0	0.01	0.05
S_{MW}	1093	0.16	0	0.16	0.16	0.16	0.16	0.17
BIS_SPAN	1093	-91.79	1.8	-95.94	-93.06	-92.16	-90.83	-84.64
FWHM	1093	7784.96	1.73	7780.47	7783.91	7784.62	7785.77	7795.17
CON	1093	60.13	0.04	59.96	60.1	60.13	60.15	60.22
Day_SSnum	1093	32.03	27.97	0	12	26	50	169
Flare Index	1093	1.07	4.91	0	0	0	0.52	107.54
Adj_flux	1093	84.39	15.43	64.57	72.33	79.1	93.97	157.03
MGII_In	1093	0.15	0	0.15	0.15	0.15	0.16	0.16
TSI_1au	1093	1360.91	0.29	1359.39	1360.73	1360.84	1361.05	1361.92

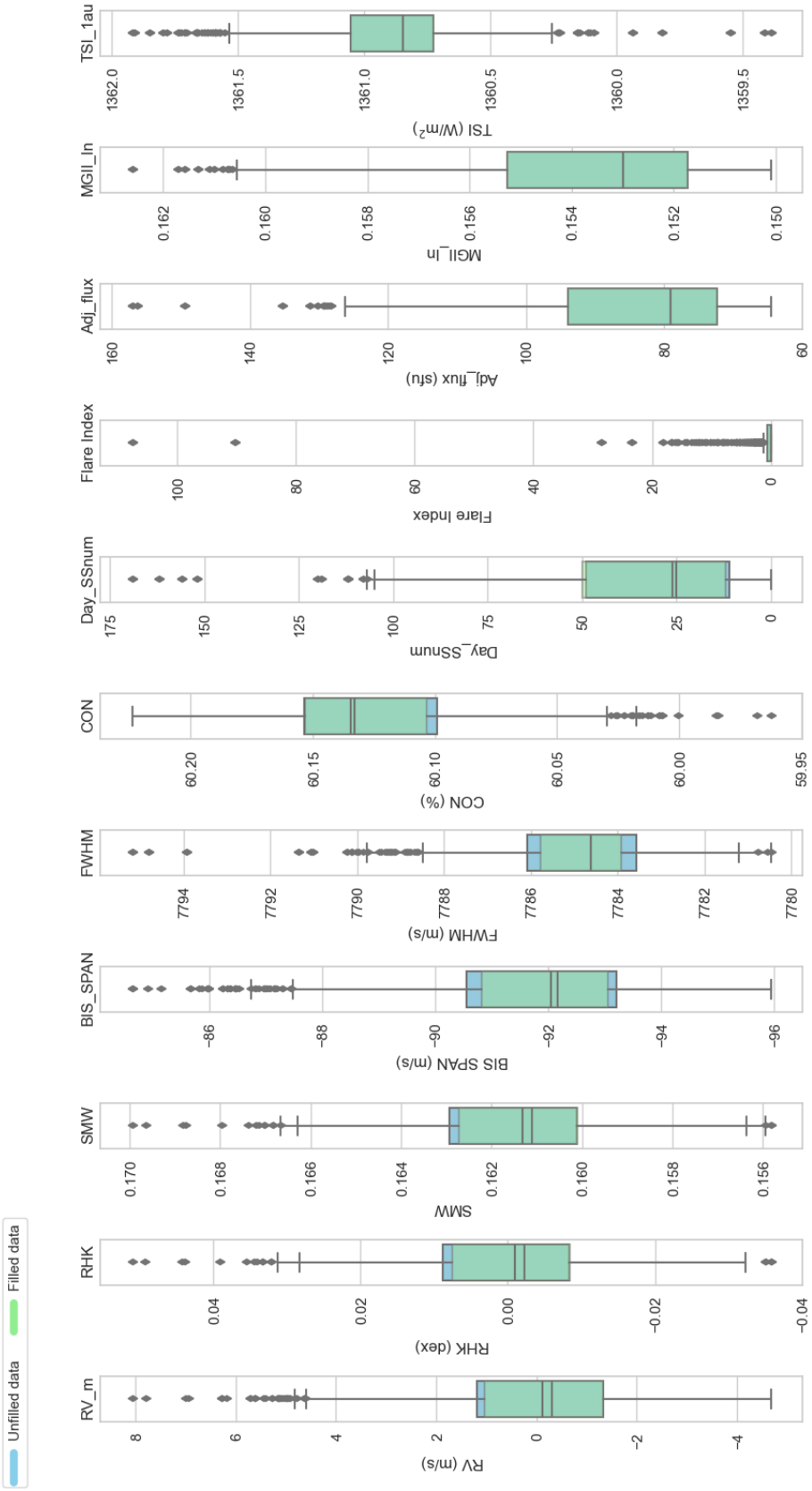


Figure 2.11: Boxplots showing comparison of unfilled and filled solar indicators.

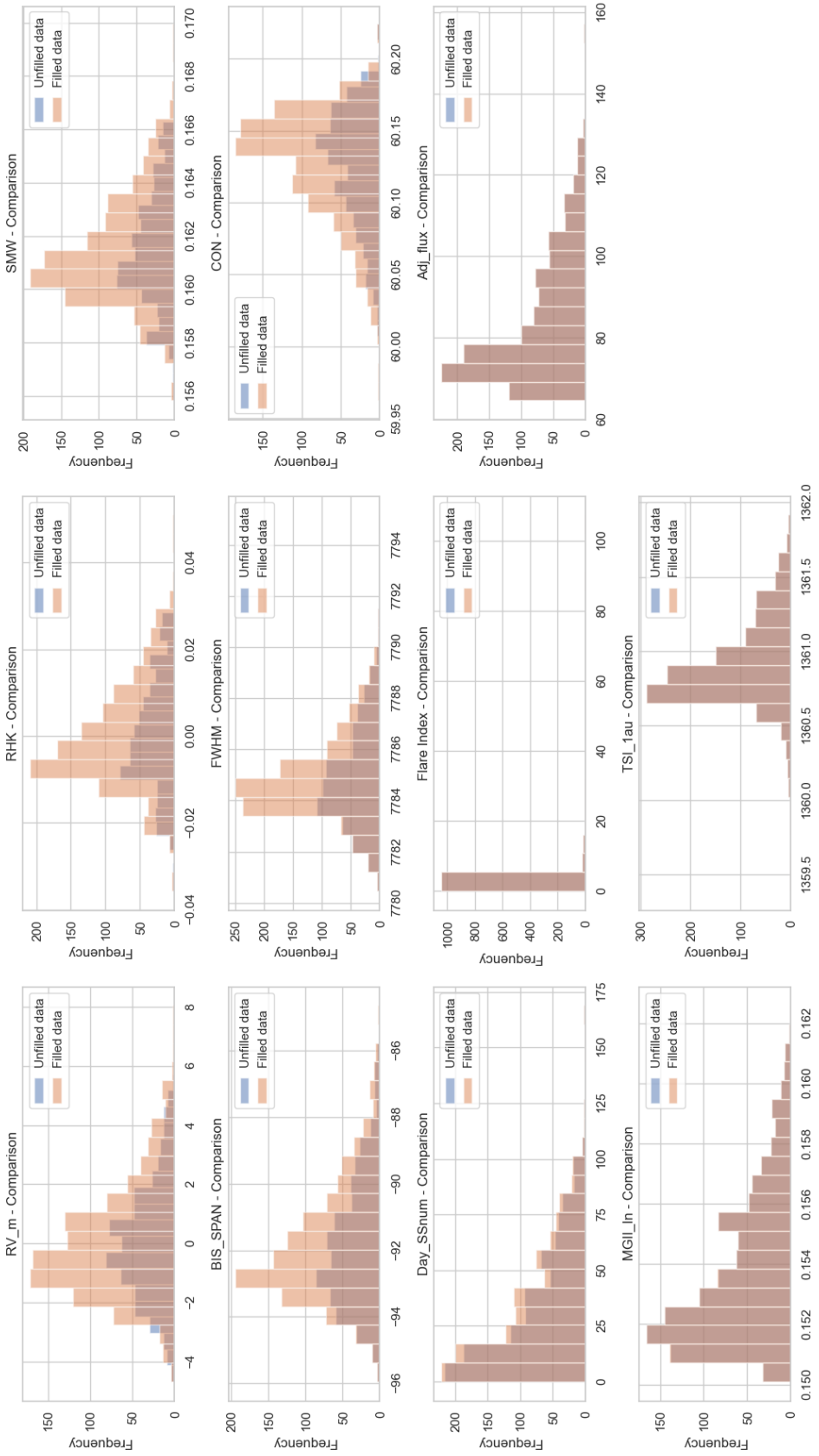


Figure 2.12: Histograms comparing unfilled and filled solar indicators.

Chapter 3

Analysis and Results

3.1 Correlation of Solar Indicators

To examine the relationships between the solar indicators, a simple correlation matrix was computed using Pearson's correlation coefficient to measure the linear relationship between each pair of variables:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}, \tag{3.1}$$

where x_i and y_i are individual sample points, and $\bar{x}$ and $\bar{y}$ are the sample means.

The correlation matrix was visualized using a heat map as seen in Figure 3.1. From this, we see that the RV measurements have strong correlations with the RHK index, S_{MW} index, Bis Span, 10.7 cm radio flux, and the Mg II index. It also has a strong anti-correlation with the contrast and weak correlations with the FWHM, daily sunspot number, and total solar irradiance.

Of all the solar indicators tested, the flare index is the only solar activity indicator with a poor correlation to the RV measurements. This is most likely because solar flares are sporadic events. While they are connected to solar magnetic activity, their occurrences are sudden, intense, and less frequent and do not contribute significantly to the RV signal of their host stars.

Figure 3.2 shows a scatter plot matrix of the solar indicators to better visualize their relationships. This allows for the exploration of linear and non-linear correlations between indicators. Each scatter plot represents the relationship between two indicators while the histograms along the diagonal show the distribution of the indicators.

The HARPS RV measurements show a clear positive linear relationship with the RHK index, S_{MW} index, and the bisector span as expected from the correlation matrix. This strong correlation is likely due to these indicators' direct connection to chromospheric and stellar activity which also affects the RV measurements.

The FWHM and MG II index also show a clear positive linear relationship, however, the data points are more dispersed compared to the previously mentioned indicators. This may suggest that the broadening of the spectral lines caused by rotation and macroturbulence effects directly affect the RV measurements, however, there may be some additional variability that is not immediately clear.

The strong negative linear relationship between the RV measurements and the contrast stands out with lower contrast values being associated with larger RVs.

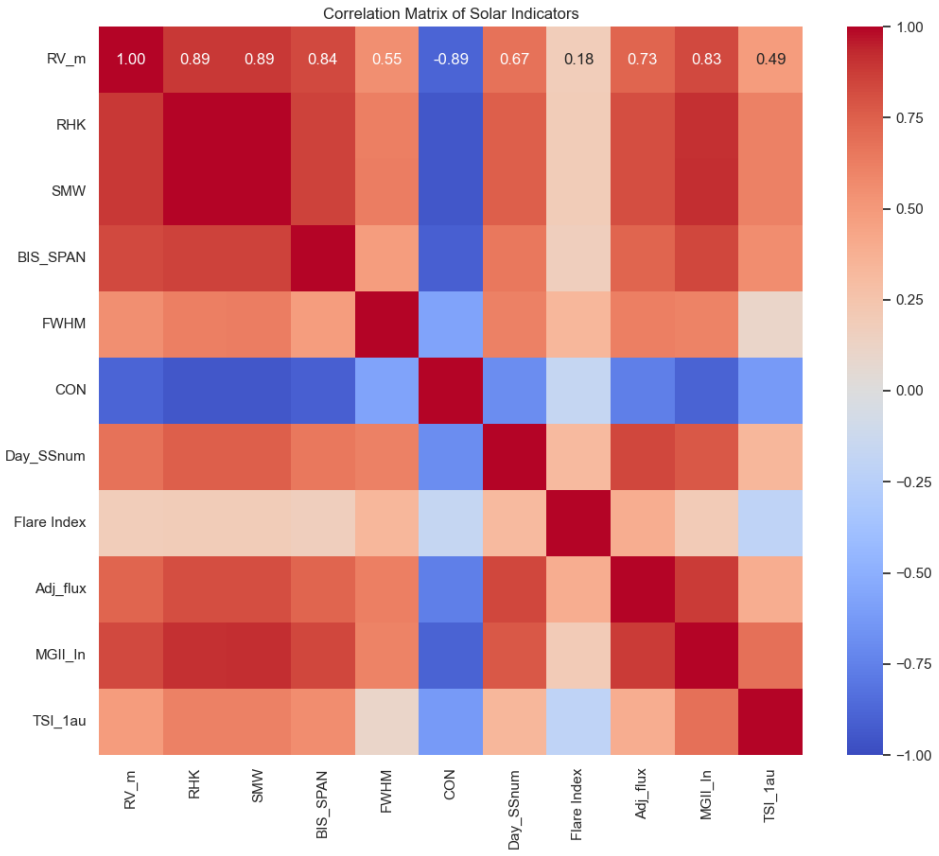


Figure 3.1: Heat map visualizing the correlation matrix of solar indicators.

As previously mentioned, the contrast indicates the sharpness and strength of the spectral lines. It makes sense that as the RV measurements increase, the sharpness of the spectral lines would appear to decrease due to the shifting of the spectral lines.

The relationship between the RV measurements and the daily sunspot numbers, the 10.7 cm radio flux, and the TSI are more dispersed showing weaker and more scattered patterns. This indicates that these solar indicators may have less influence on the RV measurements. It is also possible that the relationship may be more complex and not strictly linear.

Finally, the flare index data was observed to be slightly dispersed about the 0 value with no clear linear relationship with the RV or any of the other solar indicators. This is again most likely due to the sporadic nature of solar flares and its complex relationship with the solar magnetic field.

Combining these relationships with the distribution of each solar indicator (histograms) in the sample can provide more insight into the impact of the solar indicators on the RV measurements. The histogram of the RV measurements shows a relatively symmetric distribution centered on zero. This is because the RV signal is dominated by stellar rotation. Similarly, the bisector span has a relatively symmetric distribution showing no overall asymmetry in this data set.

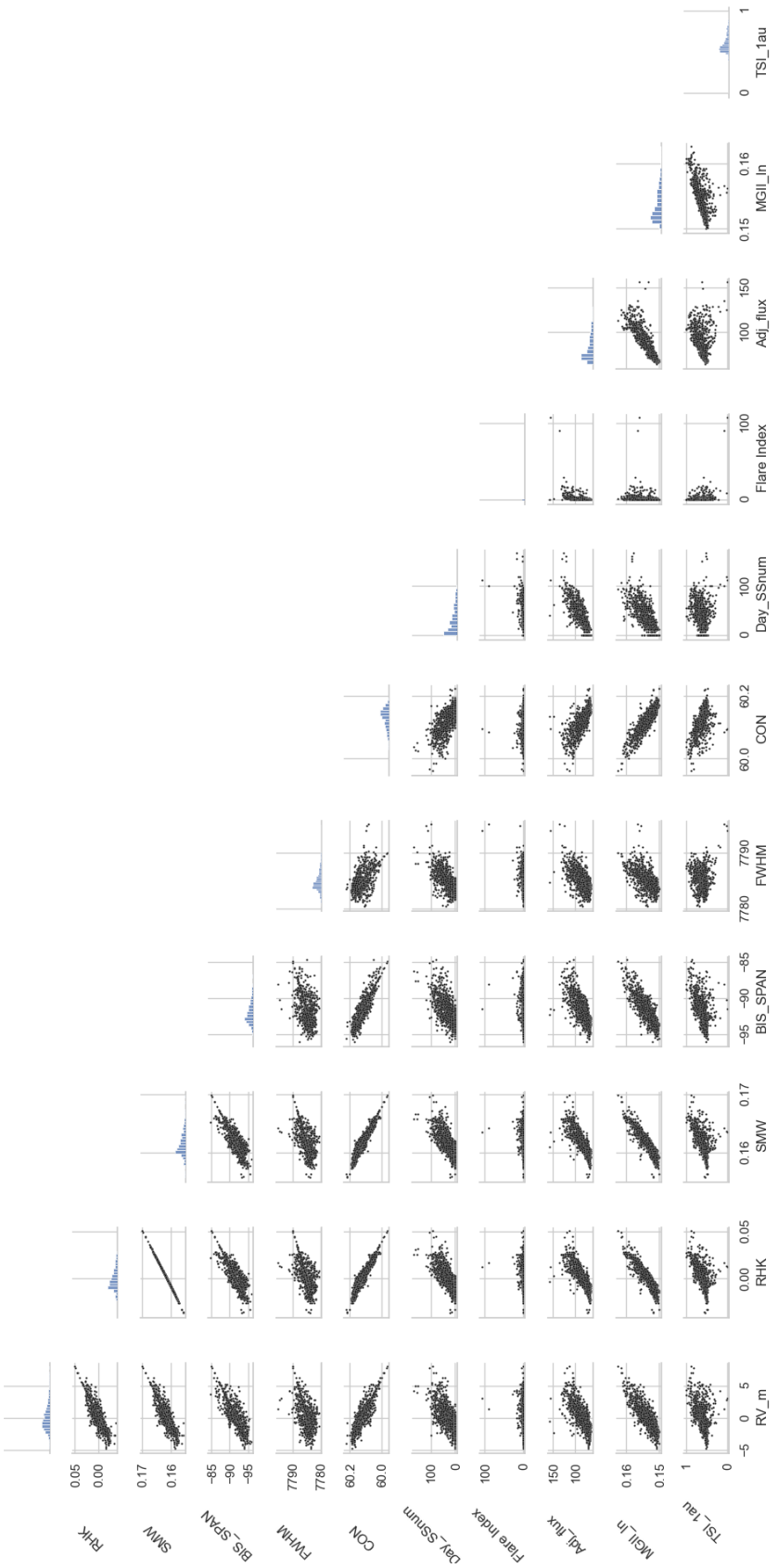


Figure 3.2: Scatter plot matrix of solar indicators with histograms along the diagonal.

The distribution of data for the RHK and S_{MW} are both skewed toward lower values. This suggests that low to moderate levels of chromospheric activity are common in this dataset. Combined with the strong correlations to the RV measurements, it implies that periods of increased activity, while less frequent, have significant impacts on the RV signal. Although they have weaker correlations with the RV measurements, the MG II indicator, 10.7 cm radio flux, daily sunspot numbers, and FWHM histograms are all skewed to lower measurements. The increased frequency of lower values may be attributed to data collection during the Sun's approach to solar minimum.

This analysis shows that the solar flare index was the only solar activity indicator tested without a significant linear relationship with the HARPS-N RV measurements. It is suggested that the remaining solar activity indicators can serve as proxies of stellar activity and be used to model the stellar RV of sun-like stars. This hypothesis is further discussed and tested in Section 3.2.

3.2 Modeling Solar RVs using Solar Proxies

As stated in the previous section, it was hypothesized that the solar activity indicators with good correlations can be used as proxies of stellar activity to model the stellar RV of sun-like stars.

To test this hypothesis, three groups of solar indicators were first defined to evaluate their ability to predict the HARPS-N RV measurements using a linear regression analysis:

- HARPS-N Indicators: The set of indicators usually provided with the HARPS-N data comprised of RHK, S_{MW}, Bis Span, FWHM, and contrast (Dumusque et al., 2015),
- Good Indicators: This includes all the tested solar indicators that correlate well with the RV measurements such as RHK, S_{MW}, Bis Span, FWHM, contrast, Day_SSnum, 10.7 cm radio flux, Mg II index, and the normalized TSI,
- Suggested Indicators: The group contained the tested solar indicators which can also be observed for other sun-like stars including RHK, S_{MW}, Bis Span, FWHM, contrast, 10.7 cm radio flux, Mg II index, and the normalized TSI.

The HARPS-N RV dataset was then randomly split 80/20 into training and testing sets. The training set was used to train the linear regression model while the testing set was used to evaluate the model's performance.

The linear regression model used is expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon, \tag{3.2}$$

where y is the RV measurements, X_1, X_2, ..., X_n are the selected solar indicators, β_0, β_1, ..., β_n are the regression coefficients, and ϵ is the error term. This model assumes a linear relationship between the solar indicators and the RV measurements.

The mean squared error (MSE), the R-squared (R^2) statistic, and the moments of the residuals between the observed and predicted data were then estimated to evaluate the performance of the model. The performance metrics for each model can be seen in Table 3.1.

The MSE was calculated using Equation 3.3 and measures the averaged squared difference between the observed and predicted values with a lower value indicating a better model performance (Shumway and Stoffer, 2006; Kedem and Fokianos, 2005):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_m - y_p)^2, \tag{3.3}$$

where y_m and y_p are the measured and predicted RV values.

The R^2 statistic given by 3.4 measures the proportion or variance in the dependent variable that is predictable from the independent variables with a R^2 value closer to 1 indicating a better fit (Shumway and Stoffer, 2006; Kedem and Fokianos, 2005).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_m - y_p)^2}{\sum_{i=1}^{n}(y_m - \bar{y})^2}, \tag{3.4}$$

where $\bar{y}$ is the mean of the measured values.

The moments of the residuals include the mean (1st moment), standard deviation (2nd moment), skewness (3rd moment), and kurtosis (4th moment). The mean of the residuals (μ_r) provides the average error, given by Equation 3.5, and should be zero if there are no systematic biases in the predictions. Equation 3.6 gives the standard deviation of residuals (σ_r) which measures the spread of the residuals around the mean. A smaller value represents more consistent predictions and less variability in the prediction errors. The skewness is a measure of the asymmetry of the distribution of residuals with 0 representing a symmetric distribution while positive or negative values indicate that the residuals were more frequently underpredicted or overpredicted respectively. The skewness is given by Equation 3.7. Finally, Kurtosis measures the "tailedness" of the distribution compared to a normal distribution. Estimated using Equation 3.8, a positive kurtosis suggests the presence of outliers while a negative kurtosis suggests fewer extreme residuals than a normal distribution (Hazelton, 2011).

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n} r_i, \tag{3.5}$$

where $r_i = y_m - y_p$,

$$\sigma_r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \mu_r)^2}, \tag{3.6}$$

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i - \mu_r}{\sigma_r}\right)^3, \tag{3.7}$$

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i - \mu_r}{\sigma_r}\right)^4 - 3. \tag{3.8}$$

Table 3.1: Table showing model's performance metrics using different solar indicator groups.

Solar Indicator Group	**MSE**	R^2	**Moments of the Residuals**			
			μ_r	σ_r	**Skewness**	**Kurtosis**
HARPS-N	0.65	0.80	0.00	0.83	0.21	2.68
Good	0.58	0.82	0.00	0.80	0.10	2.12
Suggested	0.58	0.82	0.00	0.80	0.10	2.10

Figure 3.3 visually compares predicted and measured RV data of different models using scatter plots with a "perfect fit" line showing the ideal scenario where the models accurately predict the measured data.

The time series plots of the predicted and measured RV data for different models were also visually compared in Figure 3.4.

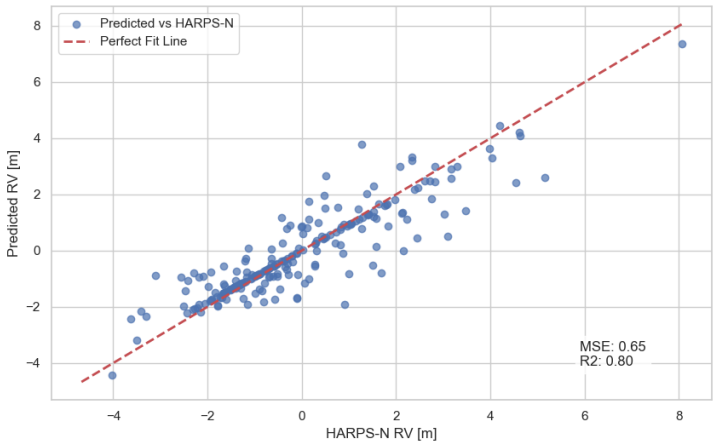
From Table 3.1, it is observed that the MSE for the HARPS-N indicator model is slightly higher (0.65) than the good and suggested indicator models (0.58), suggesting an increase in the predictive performance of the linear model when the Daily sunspot number, 10.7 cm radio flux, Mg II index, and the normalized TSI are included in the model. The R^2 values support this with both Good and Suggested indicator models achieving a R^2 of 0.82 which is a small improvement from the HARPS-N indicator model R^2 of 0.80.

The residual means of all three models are zero (to two decimal places) indicating the lack of systematic bias in their predictions. The standard deviation of the residuals for the HARPS-N indicator models is 0.83 with the models for good and suggested indicators having a slightly smaller σ_r of 0.8 suggesting that the new indicators improve the variability of the model predictions. This means the predicted points are less dispersed as seen when comparing Figure 3.3a to Figures 3.3b and 3.3c.

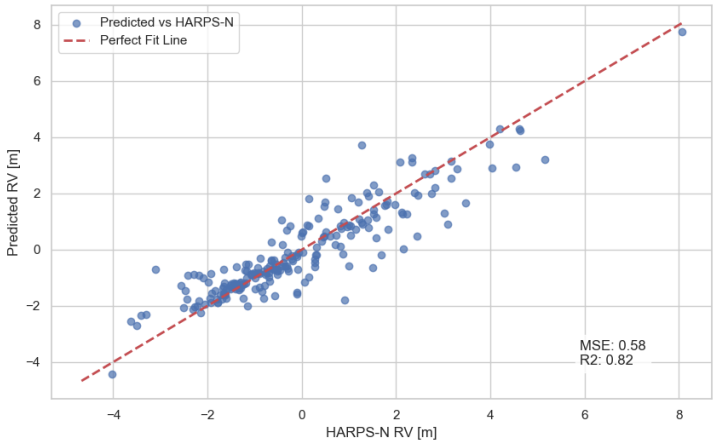
The skewness values of the good and suggested indicator models were both 0.1 which is an improvement from the HARPS-N indicator model with a positive skewness of 0.21. This indicates that all models tested tended to underpredict RV values which can be seen in Figures 3.4a to 3.4c. However, including the new indicators reduced the frequency of underpredicted values.

Kurtosis was the only moment that showed any significant difference between the good and suggested indicator models. As previously mentioned, a higher kurtosis value indicates a greater presence of extreme residuals or outliers in that model. Therefore, the HARPS-N indicator model produced the largest number of outliers with a kurtosis value of 2.68, followed by the good and then suggested indicator models with kurtosis values of 2.12 and 2.10 respectively.

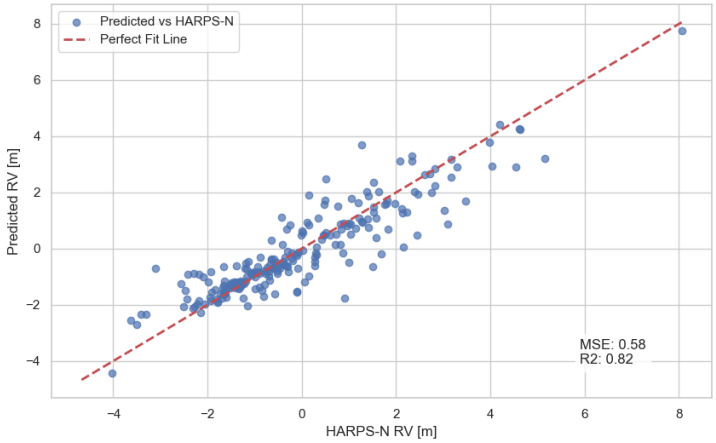
In summary, it was observed that including the additional solar activity indicators in the linear regression model provided more accurate and consistent predictions of RV variations. They also improve the reliability of the models by generating fewer outliers and a more balanced error distribution.



(a) HARPS-N indicators.

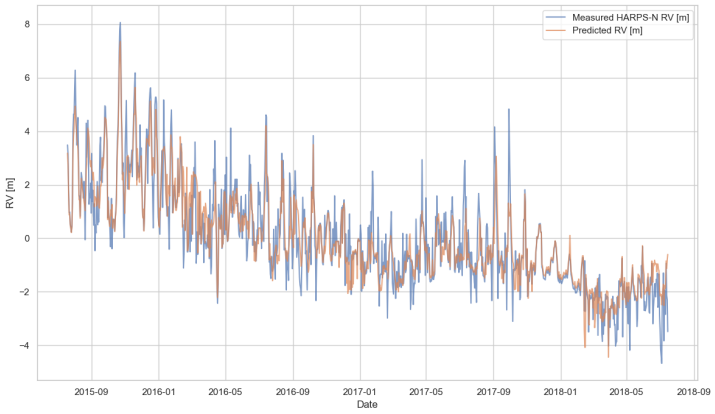


(b) Good indicators.

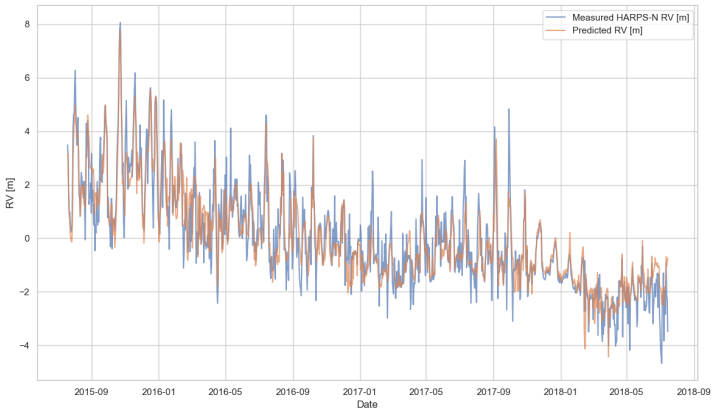


(c) Suggested indicators.

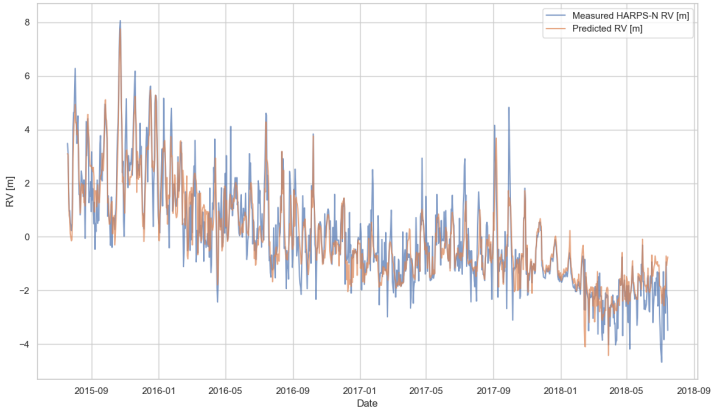
Figure 3.3: Scatter plots of predicted vs measured RV data for various solar indicators groups.



(a) HARPS-N indicators.



(b) Good indicators.



(c) Suggested indicators.

Figure 3.4: Time series comparing predicted vs measured RV data for various solar indicators groups.

Chapter 4

Discussion

The main objective of this thesis was to explore the influence of stellar activity on RV signals, focusing on how stellar activity indicators can be used to model stellar noise. Using the Sun as a proxy for other sun-like stars, an attempt was made to improve the modeling of the solar activity signal using solar activity indicators. The analysis revealed several important results, which are discussed in the following sections.

4.1 Main Results and Their Implications

The complex interplay between stellar magnetic activity and RV measurements was highlighted by the correlations between solar activity indicators and the RV signals observed in this study. Several solar activity indicators were tested leading to various implications for their use in stellar activity modeling based on their varying levels of correlation with the RV measurements.

Particularly, the chromospheric activity indices such as RHK and S_{MW} had strong positive correlations with the RV measurements. This is consistent with previous studies which have identified chromospheric activity as a major contributor to RV variations (Dumusque et al., 2015; Meunier et al., 2010). Mittag et al. (2018) suggests that the calcium activity index, RHK, in particular, can serve as a reliable proxy for magnetic activity in sun-like stars which is supported by the strong positive correlation with the RV measurements observed in this study.

The daily sunspot number and the 10.7 cm radio flux also had good correlations with the RV measurements. This also supports earlier research which shows that sunspots affect convective motions and hence modulate RV measurements by suppressing convective blueshift in active regions (Dumusque et al., 2014; Haywood et al., 2014). Similarly, the 10.7 cm radio flux is a measure of solar radio emission originating from the chromosphere and corona and follows the solar cycle closely and can therefore serve as an effective proxy for solar magnetic activity (Tapping, 2013). It is important to note, that it is much more difficult to observe starspots on sun-like stars, thus, sunspot numbers were not included in the suggested activity indicators.

The Mg II index, which measures solar UV radiation, also has a positive correlation with the RV measurements. This index has been shown to track chromospheric activity similar to RHK and has been used widely in solar variability studies (Heath and Schlesinger, 1986). The positive correlation supports the use of the Mg II index

as a proxy for stellar activity in sun-like stars, particularly when high-resolution UV data is available.

The TSI showed a weaker correlation compared to the chromospheric indicators. While it is modulated by both sunspots and faculae, it appears to have a more subtle effect on RV measurements (Kopp et al., 2016). However, the TSI is still important for understanding the overall energy output of the Sun and can potentially have indirect effects on the RV, therefore it can still be considered as a proxy for stellar activity and used to model stellar RV.

In contrast, a poor correlation was observed between the solar flare index and the RV variations. While solar flares are energetic and impactful in the short term, they appear to have less influence on the long-term RV signal. This can be accounted for by the short-lived and sporadic nature of solar flares, which have a lower contribution to the overall magnetic activity cycle compared to more persistent phenomena such as sunspots (Emslie et al., 2012). While solar flare activity is important for space weather, their limited impact on RV signals makes them less relevant as proxies in RV-based exoplanet searches.

The observations of this analysis further support the importance of using multiple stellar activity indicators to model stellar activity. The strong correlations observed for indicators such as RHK, S_{MW}, Bisector span, FWHM, Contrast, 10.7 cm radio flux, Mg II index, and the normalized TSI support the potential of using these indicators in modeling stellar RV signals, thereby filtering noise caused by stellar activity when searching for exoplanets around sun-like stars.

4.2 Methodological Considerations

Linear regression-based imputation described in section 2.2.3 was one of the key methodological approaches used in this study to address the gaps in the HARPS-N data. The accuracy of this method was tested by comparing the statistical properties of the data before and after the imputation. While the results of these tests showed no significant change, the simulated RV data from the SolAster pipeline (Ervin et al., 2022) may contain uncertainties that were not considered. While the simulated data was effective in filling gaps, it may also contain its own limitations regarding assumptions about solar magnetic activity and its effects on RV. One way to avoid this uncertainty would be to utilize analysis methods that are not sensitive to gaps in the data such as periodogram-based analysis. One such example is briefly discussed in Appendix A

The relatively high R^2 values highlighted the success of using the linear regression models described in section 3.2 to predict RV variations based on solar activity indicators. However, it is important to note that this method is inherently limited in the ability to capture the full complexity of the relationships between solar activity and RV signals. Linear models cannot account for the non-linear interactions such as those caused by the interplay of different magnetic phenomena (Lanza et al., 2016). More advanced modeling techniques, such as machine learning can be more beneficial in capturing these complex dynamics (Haywood et al., 2014).

4.3 Broader Implications for Exoplanet Research

The outcomes of this research have several broad implications for exoplanet research. Specifically for improving the reliability and accuracy of RV methods for detecting Earth-like exoplanets around sun-like stars. The need for careful treatment of stellar noise in exoplanet research is highlighted by the correlations between solar activity indicators and the RV measurements in this study. This section discusses the implications of these results, their impact on current exoplanet detection methods, ongoing research, and future prospects.

4.3.1 Enhancing the Precision of RV Measurements

As previously mentioned, one of the major challenges of the RV method of exoplanet detection is differentiating between stellar-induced signals and planetary signals. This study highlights the idea that solar activity indicators such as RHK, Mg II, and the TSI among others, can be used effectively to model the RV measurements due to stellar activity. This is important for improving the ability to detect smaller, Earth-like exoplanets, which are usually hidden by the stellar activity signal (Meunier and Lagrange, 2013).

Advancements in instrumentation, such as the ESPRESSO spectrograph, can reach sub-meter-per-second precision, a significant improvement to RV measurements (Pepe et al., 2021). However, stellar activity noise is still dominant, even with these improvements. Including more solar activity proxies in the analysis, can help further reduce this noise, allowing for the detection of planets with longer orbital periods and lower masses, typically found in the habitable zones of their host stars (Dumusque et al., 2021).

Using the RV method alongside other methods such as the transit photometry method can also enhance the precision of exoplanet detection. The precise determination of planetary masses and radii is accomplished by combining RV data with transit observations from missions such as the Transiting Exoplanet Survey Satellite (TESS) (Chontos et al., 2022; Huang et al., 2018). Similarly, using stellar proxies to correct for stellar activity can improve the reliability of RV measurements and thus a more powerful complement to transit photometry methods, particularly for non-transiting planets that cannot be detected with photometry alone (Rajpaul et al., 2020, 2015).

4.3.2 Expanding Applications to Other Stellar Types

This study focused on using the Sun as a proxy for sun-like stars, however, the methods applied here can also be used for a broader range of stellar types. Smaller, cooler, M-dwarf stars which are more magnetically active than sun-like stars, are ideal targets for exoplanet searches due to the increased likelihood of detecting habitable planets around them and their prevalence in the galaxy (Bentley, 2021). As previously mentioned, this high level of magnetic activity can pose challenges for RV measurements. Applying the techniques used in this study to M-dwarfs can help disentangle stellar activity-induced RV signals from actual planetary signals, thus improving the accuracy of exoplanet detections around M-dwarfs (Robertson et al., 2014).

Future missions such as the PLATO mission, which will search for exoplanets around a variety of stellar types, will also benefit from improved stellar activity models. PLATO's broad mission to study stellar variability, as well as its focus on bright sun-like stars provides an opportunity to test the techniques developed in this study to a larger sample of stars (Rauer et al., 2014). PLATO aims to improve the understanding of planetary system architectures and identify more Earth-like planets in habitable zones by refining the detection of planetary signals among stellar noise.

4.3.3 Implications for Habitable Zone Exoplanet Searches

Another major objective of exoplanet research involves the search for exoplanets in the habitable zone where liquid water could exist. However, due to their faint signals and long orbital periods, high-precision measurements over long periods are needed for detection. Using solar activity indicators to model and correct for stellar activity as seen in this study can potentially improve the likelihood of detecting these habitable zone planets, particularly those around sun-like stars where the potential for habitability is the highest.

Better understanding and modeling of RV signals of stellar magnetic cycles can be achieved using indicators such as RHK and Mg II indices and used to refine RV measurements of Earth analogs, which can be obscured by stellar activity cycles, and increase the possibility of detecting small, potentially habitable planets (Meunier and Lagrange, 2019).

The ability to correct for stellar noise will be critical as observation precision improves, to confirm and access the potential habitability of habitable zone exoplanets. The importance of improving RV techniques to detect these planets is emphasized by the discovery of habitable zone planets such as Proxima Centauri b (Anglada-Escudé et al., 2016).

Additionally, the stellar activity indicators discussed in this research are not only important for mitigating stellar noise but can also provide valuable information about the habitability of exoplanets orbiting sun-like stars since magnetic field driven stellar activity plays an important role in determining the habitability of exoplanets.

Intense stellar flares and CMEs, which are linked to indicators such as the Mg II index and the 10.7 cm radio flux, have been shown to strip atmospheres from planets, especially those without strong magnetic fields. This can lead to atmospheric erosion and loss of surface water and volatiles which are important for habitability (Tarter et al., 2007; Lammer et al., 2007).

The Mg II index also serves as an indicator of solar UV radiation which can be harmful to planetary habitability. Molecules in a planet's atmosphere can be broken down by high levels of UV radiation, depleting valuable protective layers such as ozone, thus leaving the surface vulnerable to harmful radiation (Segura et al., 2010). This can also alter the atmospheric chemistry of planets orbiting active stars leading to difficulties for the development or persistence of life.

Moreover, stellar activity cycles, similar to the Sun's 11-year cycle, can also lead to long-term changes in planetary climates. Variability in TSI measurements of sun-like stars can provide information on the changes in stellar output which can lead to periods of warming or cooling which can affect a planet's capacity to maintain stable, life-supporting conditions Nandy et al. (2021).

While this analysis is focused on the mitigation of stellar noise, it is also impor-

tant to note the significance of solar activity indicators in determining exoplanet habitability.

4.4 Limitations and Future Directions

While this study has shown the potential of using solar activity indicators as proxies for stellar activity, several limitations should be acknowledged. As mentioned in section 4.2, a linear regression-based imputation was used to fill gaps in the HARPS-N data. While this is necessary for a continuous time series, it can introduce uncertainties that may affect the final analysis. It may be useful to consider periodogram-based analysis techniques as discussed in Appendix A which are not sensitive to gaps in the data so to avoid introducing unnecessary uncertainties. The study should also be expanded to include other sun-like stars to validate these methods on a broader range of stellar systems.

The linear regression approach, while effective, may not capture the full complexity of the relationships between solar activity and the RV signals, particularly with long-term phenomena such as magnetic cycles (Lanza et al., 2016). Future research should aim to incorporate non-linear techniques to improve the models used to predict RV variations. As previously mentioned, machine learning algorithms can model complex non-linear relationships between solar activity indicators and RV measurements more effectively than linear models (Jones et al., 2022). There is therefore great potential to further reduce stellar noise by using machine learning techniques in the analysis of RV data. Recent studies by Perger et al. (2023) and de Beurs et al. (2022) have shown that machine learning techniques can improve precision RV measurements by predicting and removing RV variations due to stellar activity. Further development of machine learning techniques could improve how stellar activity noise is removed from RV measurements, thus enhancing the detection of small exoplanets.

Additionally, the next generation of spectrographs will push the limits of RV precision to detect Earth-like exoplanets. Spectrographs being planned for the Thirty Meter Telescope (TMT) and the Extremely Large Telescope (ELT) will aim for precisions on the order of centimeters per second (Larkin et al., 2016; Jurgenson et al., 2016). As these new spectrographs come online, techniques to mitigate stellar noise need to be developed using more comprehensive methods.

Finally, space-based RV measurements such as those from the upcoming PLATO mission should be integrated into the analysis. The higher precision data with fewer observation constraints will allow for a more comprehensive review of the impact of stellar activity on RV signals and further improve the accuracy of the RV method (Rauer et al., 2014).

Chapter 5

Conclusion

This thesis focused on the influence of solar activity on RV measurements by utilizing the Sun as a star approach, where the Sun serves as a proxy for understanding stellar activity in sun-like stars. Using the correlations between various solar activity indicators and RV measurements, this study aimed to investigate the influence of stellar activity features on RV signals using the Sun as a star approach. This chapter summarizes the main findings from my Master's thesis work.

One of the main results was significant correlations identified between solar RV measurements and several solar activity indicators such as $\log(R'_{HK})$ index, S_{MW}, sunspot number, 10.7 cm radio flux, Mg II index, along with cross-correlation metrics such as the bisector span, FWHM, and contrast. These indicators, closely tied to solar magnetic activity, were found to reflect variations in RV measurements strongly, making them good stellar indicators to use for mitigating stellar noise in RV data.

Conversely, the solar flare index was found to have a poor correlation with the RV measurements suggesting no clear linear relationship between them. This can be attributed to the sporadic nature of solar flares which contribute less to stellar-induced noise.

This thesis also demonstrated the effectiveness of using a multi-proxy approach to increase accuracy when modeling stellar activity. Increasing the range of indicators in the HARPS-N dataset to include data such as the 10.7 cm radio flux and Mg II index was found to improve models of stellar noise. This improved multi-proxy method increased the accuracy and reliability of models used to predict solar activity and remove stellar noise which can enhance the precision of exoplanet searches.

The results of the internship, discussed in Appendix A, showed that time series detrending techniques and the Generalized Lomb-Scargle periodogram were effective at removing stellar activity trends and identifying significant periodicities in RV measurements. These techniques allowed for the isolation of planetary signals from stellar noise thereby helping to identify periodicities associated with planetary signals.

This research has broader implications that extend beyond the study of our Sun. The methods developed here can be applied to other sun-like stars, providing improved methods to disentangle stellar noise from planetary signals. This can enhance the potential for detecting Earth-like exoplanets in the habitable zones of their host stars. The indicators analyzed such as the Mg II index and TSI can also provide information on the UV output and variability in the stellar output, respectively, which be useful in determining the potential habitability of exoplanets.

Overall, the results highlight the need for careful consideration of stellar activity when analyzing RV data, particularly when searching for small, Earth-like exoplanets, where stellar noise can easily hide planetary signals. This research also highlights the importance of a multi-proxy approach in stellar activity modeling, demonstrating that combining different stellar activity indicators can improve the accuracy of corrections for stellar noise, therefore improving the reliability of exoplanet detections. The methods highlighted in this study can be applied to other sun-like stars to help remove the effects of stellar activity in RV measurements and increase the likelihood of detecting Earth-like planets in the habitable zone.

In conclusion, this thesis contributes to the ongoing effort to refine RV methods for exoplanet detection by improving the understanding of how stellar activity affects RV measurements. Using the Sun as a test case, this research provides valuable insights that can be applied to other sun-like stars, ultimately advancing the ability to detect and characterize Earth-like exoplanets in the habitable zone.

Bibliography

Aigrain, S., Angus, R., Barstow, J., Rajpaul, V., Gillen, E., Parviainen, H., Pope, B., Roberts, S., McQuillan, A., Gibson, N., et al. (2016). The effects of stellar activity on detecting and characterising planets. In *19th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun (CS19)*, page 12.

Aime, C. and Vakili, F. (2006). Direct imaging of exoplanets: Science & techniques. In *IAU Colloq. 200: Direct Imaging of Exoplanets: Science & Techniques*.

Al Moulla, K., Dumusque, X., Figueira, P., Curto, G. L., Santos, N., and Wildi, F. (2023). Stellar signal components seen in harps and harps-n solar radial velocities. *Astronomy & Astrophysics*, 669:A39.

Anglada-Escudé, G., Amado, P. J., Barnes, J., Berdiñas, Z. M., Butler, R. P., Coleman, G. A., de La Cueva, I., Dreizler, S., Endl, M., Giesers, B., et al. (2016). A terrestrial planet candidate in a temperate orbit around proxima centauri. *nature*, 536(7617):437–440.

Ataç, T. (1987). Time variation of the flare index during the 21st solar cycle. *Astrophysics and space science*, 135:201–205.

Basu, S. and Antia, H. (2008). Helioseismology and solar abundances. *Physics Reports*, 457(5-6):217–283.

Bennett, D., Rhie, S., Udalski, A., Gould, A., Tsapras, Y., Kubas, D., Bond, I., Greenhill, J., Cassan, A., Rattenbury, N., et al. (2016). The first circumbinary planet found by microlensing: Ogle-2007-blg-349l (ab) c. *The Astronomical Journal*, 152(5):125.

Bentley, J. (2021). *Aspects Associated with the Use of M-dwarf Stars in Exoplanet Searches*. PhD thesis, UNSW Sydney.

Brahm, R., Nielsen, L. D., Wittenmyer, R. A., Wang, S., Rodriguez, J. E., Espinoza, N., Jones, M. I., Jordán, A., Henning, T., Hobson, M., et al. (2020). Toi-481 b and toi-892 b: Two long-period hot jupiters from the transiting exoplanet survey satellite. *The Astronomical Journal*, 160(5):235.

Brenner, P., Walbolt, K., and Schirner, T. (2024). Discoveries Dashboard — Discovery – Exoplanet Exploration: Planets Beyond our Solar System — exoplanets.nasa.gov. https://exoplanets.nasa.gov/discovery/discoveries-dashboard/. [Accessed 30-01-2024].

Butler, R. P. and Marcy, G. W. (1996). A planet orbiting 47 ursae majoris. *The Astrophysical Journal*, 464(2):L153.

Cardenas, R. (2019). Finding new worlds: Didier queloz and michel mayor's nobel prize in physics 2019. *Revista Cubana de Fisica*, 36(2).

Catala, C. and Consortium, P. (2009). Plato: Planetary transits and oscillations of stars. *Experimental Astronomy*, 23:329–356.

Chaplin, W. J., Cegla, H. M., Watson, C. A., Davies, G. R., and Ball, W. H. (2019). Filtering solar-like oscillations for exoplanet detection in radial velocity observations. *The Astronomical Journal*, 157(4):163.

Chontos, A., Murphy, J. M. A., MacDougall, M. G., Fetherolf, T., Van Zandt, J., Rubenzahl, R. A., Beard, C., Huber, D., Batalha, N. M., Crossfield, I. J., et al. (2022). The tess-keck survey:* science goals and target selection. *The Astronomical Journal*, 163(6):297.

Clette, F. and Lefèvre, L. (2016). The new sunspot number: assembling all corrections. *Solar Physics*, 291:2629–2651.

Covington, A. E. (1969). Solar radio emission at 10.7 cm, 1947-1968. *Journal of the Royal Astronomical Society of Canada, Vol. 63, p. 125*, 63:125.

Cumming, A., Butler, R. P., Marcy, G. W., Vogt, S. S., Wright, J. T., and Fischer, D. A. (2008). The keck planet search: detectability and the minimum mass and orbital period distribution of extrasolar planets. *Publications of the Astronomical Society of the Pacific*, 120(867):531.

Currie, T., Biller, B., Lagrange, A.-M., Marois, C., Guyon, O., Nielsen, E., Bonnefoy, M., and De Rosa, R. (2022). Direct imaging and spectroscopy of extrasolar planets. *arXiv preprint arXiv:2205.05696*.

de Beurs, Z. L., Vanderburg, A., Shallue, C. J., Dumusque, X., Cameron, A. C., Leet, C., Buchhave, L. A., Cosentino, R., Ghedina, A., Haywood, R. D., et al. (2022). Identifying exoplanets with deep learning. iv. removing stellar activity signals from radial velocity measurements using neural networks. *The Astronomical Journal*, 164(2):49.

Dumusque, X., Boisse, I., and Santos, N. (2014). Soap 2.0: a tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. *The Astrophysical Journal*, 796(2):132.

Dumusque, X., Cretignier, M., Sosnowska, D., Buchschacher, N., Lovis, C., Phillips, D., Pepe, F., Alesina, F., Buchhave, L., Burnier, J., et al. (2021). Three years of harps-n high-resolution spectroscopy and precise radial velocity data for the sun. *Astronomy & Astrophysics*, 648:A103.

Dumusque, X., Glenday, A., Phillips, D. F., Buchschacher, N., Cameron, A. C., Cecconi, M., Charbonneau, D., Cosentino, R., Ghedina, A., Latham, D. W., et al. (2015). Harps-n observes the sun as a star. *The Astrophysical Journal Letters*, 814(2):L21.

Dumusque, X., Udry, S., Lovis, C., Santos, N. C., and Monteiro, M. (2011). Planetary detection limits taking into account stellar noise-i. observational strategies

to reduce stellar oscillation and granulation effects. *Astronomy & Astrophysics*, 525:A140.

Emslie, A., Dennis, B., Shih, A., Chamberlin, P., Mewaldt, R., Moore, C., Share, G., Vourlidas, A., and Welsch, B. (2012). Global energetics of thirty-eight large solar eruptive events. *The Astrophysical Journal*, 759(1):71.

Ervin, T., Halverson, S., Burrows, A., Murphy, N., Roy, A., Haywood, R. D., Rescigno, F., Bender, C. F., Lin, A. S., Burt, J., et al. (2022). Leveraging space-based data from the nearest solar-type star to better understand stellar activity signatures in radial velocity data. *The Astronomical Journal*, 163(6):272.

Exoplanet Archive (2023). Exoplanet and candidate statitics.

Fischer, D. A., Anglada-Escude, G., Arriagada, P., Baluev, R. V., Bean, J. L., Bouchy, F., Buchhave, L. A., Carroll, T., Chakraborty, A., Crepp, J. R., et al. (2016). State of the field: extreme precision radial velocities. *Publications of the Astronomical Society of the Pacific*, 128(964):066001.

Fischer, D. A., Vogt, S. S., Marcy, G. W., Butler, R. P., Sato, B., Henry, G. W., Robinson, S., Laughlin, G., Ida, S., Toyota, E., et al. (2007). Five intermediate-period planets from the n2k sample. *The Astrophysical Journal*, 669(2):1336.

Foukal, P. V. (2008). *Solar astrophysics*. John Wiley & Sons.

Gallagher, N. B. (2020). Savitzky-golay smoothing and differentiation filter. *Eigenvector Research Incorporated*.

Gaudi, B. S. (2000). *Microlensing and the search for extrasolar planets*. The Ohio State University.

Gaudi, B. S. (2011). Microlensing by exoplanets. *Exoplanets*, pages 79–110.

Gaudi, B. S. and Gould, A. (1997). Planet parameters in microlensing events. *The Astrophysical Journal*, 486(1):85.

Giobbi, P. (2022/2023). *Machine Learning Algorithms for Understanding Solar Radial Velocity: Implications for Exoplanet Detection*. PhD thesis, University of Rome (Tor Vergata).

Gray, D. F. (2021). *The observation and analysis of stellar photospheres*. Cambridge university press.

Hannah, I. and Kontar, E. (2013). Multi-thermal dynamics and energetics of a coronal mass ejection in the low solar atmosphere. *Astronomy & Astrophysics*, 553:A10.

Hathaway, D. H. (2015). The solar cycle. *Living reviews in solar physics*, 12(1):4.

Haywood, R., Collier Cameron, A., Queloz, D., Barros, S., Deleuil, M., Fares, R., Gillon, M., Lanza, A., Lovis, C., Moutou, C., et al. (2014). Planets and stellar activity: hide and seek in the corot-7 system. *Monthly notices of the royal astronomical society*, 443(3):2517–2531.

Haywood, R. D., Collier Cameron, A., Unruh, Y., Lovis, C., Lanza, A. F., Llama, J., Deleuil, M., Fares, R., Gillon, M., Moutou, C., et al. (2016). The sun as a planet-host star: proxies from sdo images for harps radial-velocity variations. *Monthly Notices of the Royal Astronomical Society*, 457(4):3637–3651.

Hazelton, M. L. (2011). *Methods of Moments Estimation*, pages 816–817. Springer Berlin Heidelberg, Berlin, Heidelberg.

Heath, D. F. and Schlesinger, B. M. (1986). The mg 280-nm doublet as a monitor of changes in solar ultraviolet irradiance. *Journal of Geophysical Research: Atmospheres*, 91(D8):8672–8682.

Hippke, M., David, T. J., Mulders, G. D., and Heller, R. (2019). Wōtan: Comprehensive time-series detrending in python. *The Astronomical Journal*, 158(4):143.

Huang, C. X., Burt, J., Vanderburg, A., Günther, M. N., Shporer, A., Dittmann, J. A., Winn, J. N., Wittenmyer, R., Sha, L., Kane, S. R., et al. (2018). Tess discovery of a transiting super-earth in the pi mensae system. *The Astrophysical Journal Letters*, 868(2):L39.

information@eso.org (2023). Harps sees sunshine for the first time. https://www.eso.org/public/announcements/ann18033/. [Accessed 28-12-2023].

Jiang, J., Cameron, R. H., and Schuessler, M. (2015). The cause of the weak solar cycle 24. *The Astrophysical Journal Letters*, 808(1):L28.

Jones, D. E., Stenning, D. C., Ford, E. B., Wolpert, R. L., Loredo, T. J., Gilbertson, C., and Dumusque, X. (2022). Improving exoplanet detection power: Multivariate gaussian process models for stellar activity. *The Annals of Applied Statistics*, 16(2):652–679.

Jurgenson, C., Fischer, D., McCracken, T., Sawyer, D., Szymkowiak, A., Davis, A., Muller, G., and Santoro, F. (2016). Expres: a next generation rv spectrograph in the search for earth-like worlds. In *Ground-based and airborne instrumentation for astronomy vi*, volume 9908, pages 2051–2070. SPIE.

Kedem, B. and Fokianos, K. (2005). *Regression models for time series analysis*. John Wiley & Sons.

Kennewell, J. (2024). Solar flare patrol. Accessed: 2024-05-31.

Kilpua, E., Mierla, M., Zhukov, A., Rodriguez, L., Vourlidas, A., and Wood, B. (2014). Solar sources of interplanetary coronal mass ejections during the solar cycle 23/24 minimum. *Solar Physics*, 289:3773–3797.

Kleczek, J. (1952). Publ. inst. centr. astron., no. 22.

Kopp, G. (2019). Sorce level 3 total solar irradiance daily means v018. *Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD*.

Kopp, G., Krivova, N., Wu, C., and Lean, J. (2016). The impact of the revised sunspot record on solar irradiance reconstructions. *Solar Physics*, 291:2951–2965.

Kopp, G. and Lawrence, G. (2005). The total irradiance monitor (tim): instrument design. *Solar Physics*, 230:91–109.

Lammer, H., Lichtenegger, H. I., Kulikov, Y. N., Grießmeier, J.-M., Terada, N., Erkaev, N. V., Biernat, H. K., Khodachenko, M. L., Ribas, I., Penz, T., et al. (2007). Coronal mass ejection (cme) activity of low mass m stars as an important factor for the habitability of terrestrial exoplanets. ii. cme-induced ion pick up of earth-like exoplanets in close-in habitable zones. *Astrobiology*, 7(1):185–207.

Lanza, A. F., Molaro, P., Monaco, L., and Haywood, R. (2016). Long-term radial-velocity variations of the sun as a star: The harps view. *Astronomy & Astrophysics*, 587:A103.

Larkin, J. E., Moore, A. M., Wright, S. A., Wincentsen, J. E., Anderson, D., Chisholm, E. M., Dekany, R. G., Dunn, J. S., Ellerbroek, B. L., Hayano, Y., et al. (2016). The infrared imaging spectrograph (iris) for tmt: instrument overview. In *Ground-based and Airborne Instrumentation for Astronomy VI*, volume 9908, pages 582–594. SPIE.

Lean, J., Skumanich, A., and White, O. (1992). Estimating the sun's radiative output during the maunder minimum. *Geophysical Research Letters*, 19(15):1591–1594.

Lefebvre, S., García, R., Jiménez-Reyes, S., Turck-Chièze, S., and Mathur, S. (2008). Variations of the solar granulation motions with height using the golf/soho experiment. *Astronomy & Astrophysics*, 490(3):1143–1149.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Madsen, S., Dravins, D., Ludwig, H.-G., and Lindegren, L. (2003). Intrinsic spectral blueshifts in rapidly rotating stars? *Astronomy & Astrophysics*, 411(3):581–586.

Martin, J., Fuhrmeister, B., Mittag, M., Schmidt, T., Hempelmann, A., González-Pérez, J., and Schmitt, J. (2017). The ca ii infrared triplet's performance as an activity indicator compared to ca ii h and k-empirical relations to convert ca ii infrared triplet measurements to common activity indices. *Astronomy & Astrophysics*, 605:A113.

Mayor, M. and Queloz, D. (1995). A jupiter-mass companion to a solar-type star. *Nature*, 378(6555):355–359.

Mayor, M., Udry, S., Lovis, C., Pepe, F., Queloz, D., Benz, W., Bertaux, J.-L., Bouchy, F., Mordasini, C., and Ségransan, D. (2009). The harps search for southern extra-solar planets-xiii. a planetary system with 3 super-earths (4.2, 6.9, and 9.2 m). *Astronomy & Astrophysics*, 493(2):639–644.

Meunier, N., Desort, M., and Lagrange, A.-M. (2010). Using the sun to estimate earth-like planets detection capabilities-ii. impact of plages. *Astronomy & Astrophysics*, 512:A39.

Meunier, N. and Lagrange, A.-M. (2013). Using the sun to estimate earth-like planets detection capabilities-iv. correcting for the convective component. *Astronomy & Astrophysics*, 551:A101.

Meunier, N. and Lagrange, A.-M. (2019). Unexpectedly strong effect of supergranulation on the detectability of earth twins orbiting sun-like stars with radial velocities. *Astronomy & Astrophysics*, 625:L6.

Meunier, N., Lagrange, A.-M., Borgniet, S., and Rieutord, M. (2015). Using the sun to estimate earth-like planet detection capabilities-vi. simulation of granulation and supergranulation radial velocity and photometric time series. *Astronomy & Astrophysics*, 583:A118.

Milbourne, T., Haywood, R., Phillips, D., Saar, S., Cegla, H., Cameron, A., Costes, J., Dumusque, X., Langellier, N., Latham, D., et al. (2019). Harps-n solar rvs are dominated by large, bright magnetic regions. *The Astrophysical Journal*, 874(1):107.

Mittag, M., Schmitt, J., and Schröder, K.-P. (2018). Revisiting the connection between magnetic activity, rotation period, and convective turnover time for main-sequence stars. *Astronomy & Astrophysics*, 618:A48.

Nandy, D., Martens, P. C., Obridko, V., Dash, S., and Georgieva, K. (2021). Solar evolution and extrema: current state of understanding of long-term solar variability and its planetary impacts. *Progress in Earth and Planetary Science*, 8:1–21.

Newville, M., Otten, R., Nelson, A., Ingargiola, A., Stensitzki, T., Allan, D., Fox, A., Carter, F., Osborn, R., Pustakhod, D., et al. (2021). lmfit/lmfit-py: 1.0. 3. *Zenodo*.

Noyes, R., Hartmann, L., Baliunas, S., Duncan, D., and Vaughan, A. (1984). Rotation, convection, and magnetic activity in lower main-sequence stars. *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 279, April 15, 1984, p. 763-777. Research supported by the Smithsonian Institution and National Geographic Society.*, 279:763–777.

Özgüç, A. and Atac, T. (1989). Periodic behavior of solar flare index during solar cycles 20 and 21. *Solar Physics*, 123:357–365.

Özgüç, A., Ataç, T., and Rybák, J. (2003). Temporal variability of the flare index (1966–2001). *Solar Physics*, 214:375–396.

Pepe, F., Cristiani, S., Rebolo, R., Santos, N., Dekker, H., Cabral, A., Di Marcantonio, P., Figueira, P., Curto, G. L., Lovis, C., et al. (2021). Espresso at vlt-on-sky performance and first results. *Astronomy & Astrophysics*, 645:A96.

Perger, M., Anglada-Escudé, G., Baroch, D., Lafarga, M., Ribas, I., Morales, J., Herrero, E., Amado, P. J., Barnes, J., Caballero, J., et al. (2023). A machine learning approach for correcting radial velocities using physical observables. *Astronomy & Astrophysics*, 672:A118.

Pesnell, W. D. (2012). Solar cycle predictions (invited review). *Solar Physics*, 281:507–532.

Phillips, D. F., Glenday, A. G., Dumusque, X., Buchschacher, N., Cameron, A. C., Cecconi, M., Charbonneau, D., Cosentino, R., Ghedina, A., Haywood, R., et al. (2016). An astro-comb calibrated solar telescope to search for the radial velocity signature of venus. In *Advances in Optical and Mechanical Technologies for Telescopes and Instrumentation II*, volume 9912, pages 2163–2170. SPIE.

Poindexter, S., Afonso, C., Bennett, D. P., Glicenstein, J.-F., Gould, A., Szymański, M. K., and Udalski, A. (2005). Systematic analysis of 22 microlensing parallax candidates. *The Astrophysical Journal*, 633(2):914.

Price-Whelan, A. M., Sipőcz, B., Günther, H., Lim, P., Crawford, S., Conseil, S., Shupe, D., Craig, M., Dencheva, N., Ginsburg, A., et al. (2018). The astropy project: Building an open-science project and status of the v2. 0 core package. *The Astronomical Journal*, 156(3):123.

Quanz, S. P., Crossfield, I., Meyer, M. R., Schmalzl, E., and Held, J. (2015). Direct detection of exoplanets in the 3–10 μm range with e-elt/metis. *International Journal of Astrobiology*, 14(2):279–289.

Queloz, D., Henry, G. W., Sivan, J.-P., Baliunas, S. L., Beuzit, J.-L., Donahue, R. A., Mayor, M., Naef, D., Perrier, C., and Udry, S. (2001). No planet for hd 166435. *Astronomy & Astrophysics*, 379(1):279–287.

Rajkumar, B. and Haque, S. (2020). A fractal analysis of magnetograms within active regions. *Solar Physics*, 295(2):10.

Rajkumar, B., Haque, S., and Hrudey, W. (2017). Fractal dimensions of umbral and penumbral regions of sunspots. *Solar Physics*, 292(11):170.

Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., and Roberts, S. (2015). A gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291.

Rajpaul, V. M., Aigrain, S., and Buchhave, L. A. (2020). A robust, template-free approach to precise radial velocity extraction. *Monthly Notices of the Royal Astronomical Society*, 492(3):3960–3983.

Rauer, H., Catala, C., Aerts, C., Appourchaux, T., Benz, W., Brandeker, A., Christensen-Dalsgaard, J., Deleuil, M., Gizon, L., Goupil, M.-J., et al. (2014). The plato 2.0 mission. *Experimental Astronomy*, 38:249–330.

Robertson, P., Mahadevan, S., Endl, M., and Roy, A. (2014). Stellar activity masquerading as planets in the habitable zone of the m dwarf gliese 581. *Science*, 345(6195):440–444.

Roweis, S. (1996). Levenberg-marquardt optimization. *Notes, University Of Toronto*, 52.

Santos, N., Mortier, A., Faria, J., Dumusque, X., Adibekyan, V. Z., Delgado-Mena, E., Figueira, P., Benamati, L., Boisse, I., Cunha, D., et al. (2014). The harps search for southern extra-solar planets-xxxv. the interesting case of hd 41248: stellar activity, no planets? *Astronomy & Astrophysics*, 566:A35.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Scherrer, P. H., Schou, J., Bush, R., Kosovichev, A., Bogart, R., Hoeksema, J., Liu, Y., Duvall, T., Zhao, J., Title, A., et al. (2012). The helioseismic and magnetic imager (hmi) investigation for the solar dynamics observatory (sdo). *Solar Physics*, 275:207–227.

Schmieder, B., Malherbe, J.-M., and Wu, S. T. (2013). *Nature of Prominences and their role in Space Weather*. Cambridge University Press.

Seager, S. and Mallen-Ornelas, G. (2003). A unique solution of planet and star parameters from an extrasolar planet transit light curve. *The Astrophysical Journal*, 585(2):1038.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Segura, A., Walkowicz, L. M., Meadows, V., Kasting, J., and Hawley, S. (2010). The effect of a strong stellar flare on the atmospheric chemistry of an earth-like planet orbiting an m dwarf. *Astrobiology*, 10(7):751–771.

Shumway, R. H. and Stoffer, D. S. (2006). Time series regression and exploratory data analysis. *Time series analysis and its applications: With R examples*, pages 48–83.

Simola, U., Dumusque, X., and Cisewski-Kehe, J. (2019). Measuring precise radial velocities and cross-correlation function line-profile variations using a skew normal density. *Astronomy & Astrophysics*, 622:A131.

Snow, M., Weber, M., Machol, J., Viereck, R., and Richard, E. (2014). Comparison of magnesium ii core-to-wing ratio observations during solar minimum 23/24. *Journal of Space Weather and Space Climate*, 4:A04.

Souto, D., Unterborn, C. T., Smith, V. V., Cunha, K., Teske, J., Covey, K., Rojas-Ayala, B., García-Hernández, D., Stassun, K., Zamora, O., et al. (2018). Stellar and planetary characterization of the ross 128 exoplanetary system from apogee spectra. *The Astrophysical Journal Letters*, 860(1):L15.

Sowmya, K., Shapiro, A., van der Voort, L. R., Krivova, N., and Solanki, S. (2023). Modeling stellar ca ii h and k emission variations: Spot contribution to the s-index. *The Astrophysical Journal Letters*, 956(1):L10.

Sulis, S., Crossfield, I., Santerne, A., Saillenfest, M., Sousa, S., Mary, D., Aguichine, A., Deleuil, M., Mena, E. D., Mathur, S., et al. (2024). A low-mass sub-neptune planet transiting the bright active star hd 73344. *arXiv preprint arXiv:2405.17155*.

Tapping, K. (2013). The 10.7 cm solar radio flux (f10. 7). *Space weather*, 11(7):394–406.

Tarter, J. C., Backus, P. R., Mancinelli, R. L., Aurnou, J. M., Backman, D. E., Basri, G. S., Boss, A. P., Clarke, A., Deming, D., Doyle, L. R., et al. (2007). A reappraisal of the habitability of planets around m dwarf stars. *Astrobiology*, 7(1):30–65.

Tinney, C., McCarthy, C., Jones, H. R., Paul Butler, R., Carter, B. D., Marcy, G. W., and Penny, A. J. (2002). Echelle spectroscopy of ca ii hk activity in southern hemisphere planet search targets. *Monthly Notices of the Royal Astronomical Society*, 332(3):759–763.

UniversiteofGeneve (2014). Harps-n white paper. https://plone.unige.ch/HARPS-N/science-with-harps-n. [Accessed 28-12-2023].

Vaughan, A. H. and Preston, G. W. (1980). A survey of chromospheric ca ii h and k emission in field stars of the solar neighborhood. *Publications of the Astronomical Society of the Pacific*, 92(548):385.

Viereck, R., Puga, L., McMullin, D., Judge, D., Weber, M., and Tobiska, W. K. (2001). The mg ii index: A proxy for solar euv. *Geophysical Research Letters*, 28(7):1343–1346.

Winn, J. N. (2010). *Exoplanet transits and occultations*, volume 55. University of Arizona Press Tucson, AZ.

Wright, J., Marcy, G., Howard, A., Johnson, J. A., Morton, T., and Fischer, D. (2012). The frequency of hot jupiters orbiting nearby solar-type stars. *The Astrophysical Journal*, 753(2):160.

Wright, J. T., Marcy, G. W., Butler, R. P., and Vogt, S. S. (2004). Chromospheric ca ii emission in nearby f, g, k, and m stars. *The Astrophysical Journal Supplement Series*, 152(2):261.

Zechmeister, M. and Kürster, M. (2009). The generalised lomb-scargle periodogram-a new formalism for the floating-mean and keplerian periodograms. *Astronomy & Astrophysics*, 496(2):577–584.

Appendix A

Internship

In preparation for this thesis, an internship was carried out to become familiar with the HARPS-N Sun as a star RV data and to explore some of the techniques used to analyze time series data. It was decided that this aim could be achieved by reproducing some of the results of the paper titled "Stellar signal components seen in HARPS and HARPS-N solar radial velocities." by Al Moulla et al. (2023). The internship focused on periodogram-based analysis techniques working in the frequency domain while this study, focused on linear regression techniques working in the time domain. While very different, the analysis performed during the internship provided insight into the HARPS-N Sun as a star RV data, and how the data could be manipulated. It also provided a contrasting analysis method that could be compared to the linear regression methods used in this study. The following sections provide an overview of the attempt to reproduce some of the results by Al Moulla et al. (2023) and other exercises performed as part of the internship.

A.1 Internship Data

This internship utilized data from both the HARPS and HARPS-N instruments introduced in section 2.2. Table A.1 shows a comparison between the properties of the HARPS and HARPS-N data (Al Moulla et al., 2023). The key difference between the two data sets is the cadences. HARPS observes the Sun every minute while HARPS-N observes the Sun every 5.5 minutes which does not detect the p-mode oscillations of the Sun. Additionally, (Al Moulla et al., 2023) also reports that the HARPS-N data was reduced with the recent ESPRESSO data reduction software while HARPS was not which may be responsible for the significant discrepancy between the average daily rms values between the data sets.

Table A.1: Comparison of properties of HARPS and HARPS-N data.

Properties	HARPS (HELIOS)	HARPS-N
Observed days	376	745
Observation Period	08/09/2018 to 20/03/2020	16/07/2015 - 15/07/2018
Cadence	1 min	5.5 min
Exposure Time	0.5 min	5 min

The detrended HARPS data was obtained from Al Moulla et al. (2023) while the HARPS-N data was obtained from DACE as mentioned in section 2.2. Figure A.1 shows a side-by-side comparison of the HARPS and HARPS-N RV time series.

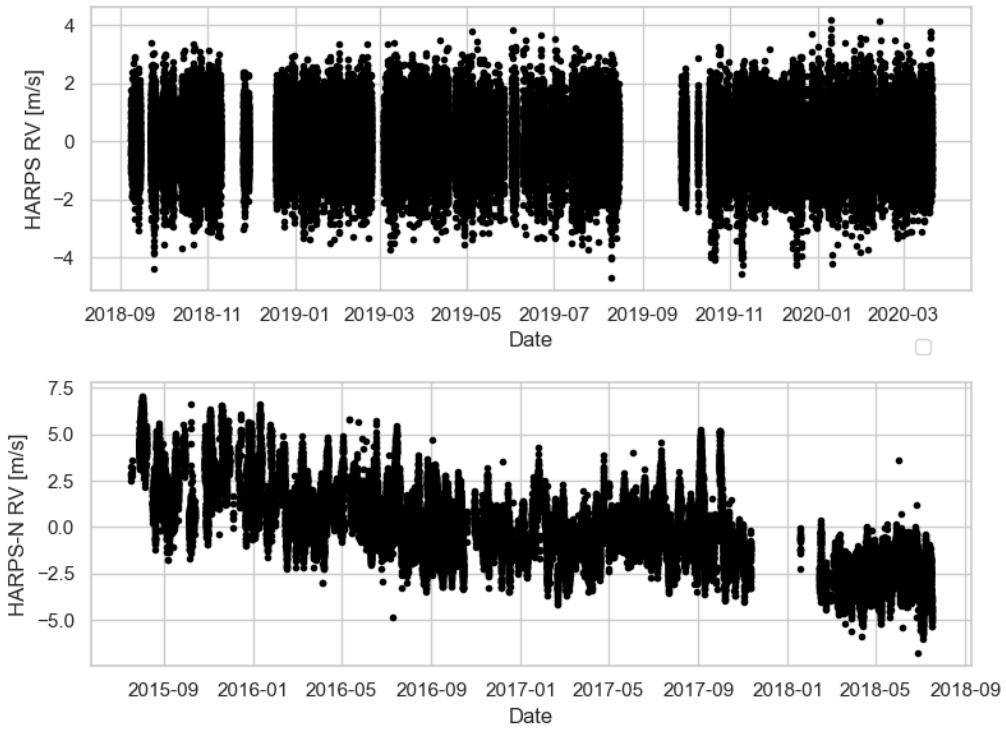


Figure A.1: Side-by-side comparison of HARPS and HARPS-N RV time series.

A.2 Time-series Detrending

When preparing the data for this analysis, the first step was to remove unwanted trends using time series detrending. This technique helps eliminate systematic effects or other trends of a specific frequency, allowing for clearer identification of planetary signals. As Hippke et al. (2019) explains, there are numerous methods available for detrending time series data, each with its strengths and limitations. The HARPS RV data obtained has already been detrended using a mean over 1 day. (Al Moulla et al., 2023) used a 100-day rolling mean to detrend the HARPS-N RV data, however, this result was not successfully reproduced. Instead, the *Wotan* python package was used to detrend the HARPS-N data using the Savitzky-Golay (savgol) filter.

The savgol filter is a popular technique for smoothing time series data. It works by fitting a polynomial of a specified degree to a subset (or window) of data points, using the method of least squares. The window then slides through the entire dataset, and the fitted polynomial is evaluated at each point within the window. The central data point of each window is replaced by the value of the fitted polynomial at that position, resulting in a smoothed time series (Gallagher, 2020; Savitzky and Golay, 1964).

For each window of data, a polynomial, $y(x)$, of degree, p, is fitted to the data points:

$$y(x) = a_0 + a_1 x + a_2 x^2 + ... + a_p x^p, \tag{A.1}$$

where a_i are the coefficients of the polynomial. Once the polynomial is fitted to the data within the window of size N, it is evaluated at the central point of the window giving the smoothed value at the central point, $y_{smoothed}(x_c)$:

$$y_{smoothed}(x_c) = a_0 + a_1 x_c + a_2 x_c^2 + ... + a_p x_c^p. \tag{A.2}$$

The choice of polynomial degree and window size are important. A higher degree polynomial allows for a closer fit to the data within the window which can capture more complex variations but can also capture more noise. A lower-degree polynomial, however, results in a smoother trend but can miss finer details. Similarly, a larger window size captures broader trends but risks over-smoothening, while a smaller window size can be overly sensitive to noise but can preserve more detailed structures.

For this analysis, the goal of the time-series detrending was to remove the long-term or low-frequency trend which would correspond to the 11-year magnetic activity cycle and instrumental long-term drifts (Al Moulla et al., 2023). Therefore, the savgol filter was applied to the HARPS-N data with a polynomial degree of 1 and different window lengths (51,501,5001 data points). The resulting trend line, <RV>, is then subtracted from the RV data. Figure A.2 shows the results of the HARPS-N time series detrending. As expected, the larger window length successfully captured the long-term trend with minimal effect on the stellar variations.

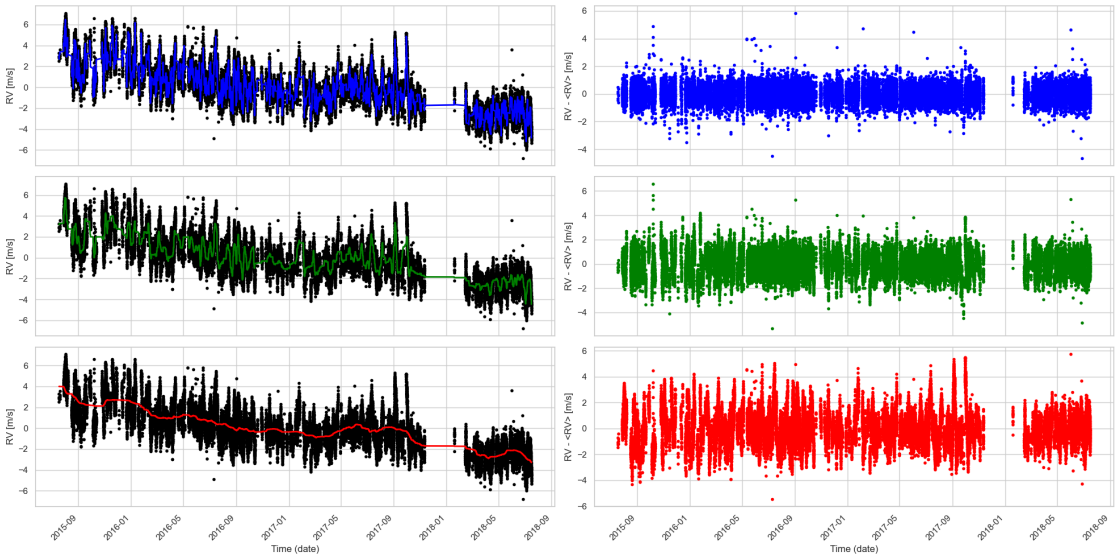


Figure A.2: Resulting trend lines (left) and detrended HARPS-N data (right) for window lengths 51, 501, 5001 (blue, green, and red respectively) overlaid on the HARPS-N time series (black)

A.3 Periodogram Analysis

As previously mentioned, due to the gaps in the HARPS-N data, it may be more favorable to use a periodogram-based method to analyze the data. When working with RV measurements for exoplanet detection, it is also often necessary to analyze the periodicities within the data to detect signals that could indicate the presence of

exoplanets. In many cases, however, traditional discrete Fourier transform (DFT) methods are not suitable due to gaps in the data or irregular sampling. For irregularly sampled data, the General Lomb-Scargle (GLS) periodogram provides a powerful alternative for determining the power spectrum density (PSD) of the time series (Zechmeister and Kürster, 2009).

The GLS periodogram identifies periodic signals in unevenly spaced observations by fitting a model composed of a linear combination of sinusoids and a constant offset which can be expressed by Equation A.3:

$$v(t) = Acos(2\pi ft) + Bsin(2\pi ft) + C, \tag{A.3}$$

where $v(t)$ is the RV data at time t, f is the frequency, A and B are the amplitudes of the cosine and sine components, and C is the constant offset.

The GLS periodogram aims to minimize the residuals between the observed data and this sinusoidal model by varying the frequency f and determining the optimal values of A, B, and C that provide the best fit. The periodogram then computes the power at each frequency, where the power is defined as the reduction in the sum of the squared residuals when the model is applied compared to a constant model as seen in Equation A.4:

$$P(f) = \frac{1}{2}\left(\frac{(\sum_i w_i(v_i - C)cos(2\pi ft_i))^2}{\sum_i w_i cos^2(2\pi ft_i)} + \frac{(\sum_i w_i(v_i - C)sin(2\pi ft_i))^2}{\sum_i w_i sin^2(2\pi ft_i)}\right), \tag{A.4}$$

where $P(f)$ is the power at frequency f, and w_i are the weights associated with each data point v_i at time t_i.

The velocity power spectrum (VPS) can then be determined using the summed squares of the sine and cosine terms at each frequency. Finally, the velocity power spectrum density (VPSD) is calculated by multiplying the power by the total length of the observation run (Al Moulla et al., 2023). This provides a measure of the strength of the periodic signal in the data as a function of frequency.

The GLS periodogram was applied to the HARPS and HARPS-N RV data using the 'LombScargle' function from the Astropy library (Price-Whelan et al., 2018). From the VPS seen in Figure A.2, we observe the p-mode oscillations in the HARPS VPS, which has a cadence of 1 minute, while it does not appear in the HARPS-N VPS, which has a cadence of 5.5 minutes. This is because the p-mode oscillations of the Sun have a 5-minute period which isn't properly sampled by the HARPS-N observations. Granulation accounts for the slope seen in both VPS, however, the HARPS-N VPS also contains signals due to rotation and aliasing caused by the day/night cycle which account for the increased power in the lower frequencies. Due to the different cadences, higher-frequency signals were better captured in the HARPS data while the HARPS-N data was better at capturing lower-frequency signals. Following Al Moulla et al. (2023), the frequencies in the gray areas of A.3 were combined and then divided by the length of the observation window to produce the VPSD seen in Figure A.4.

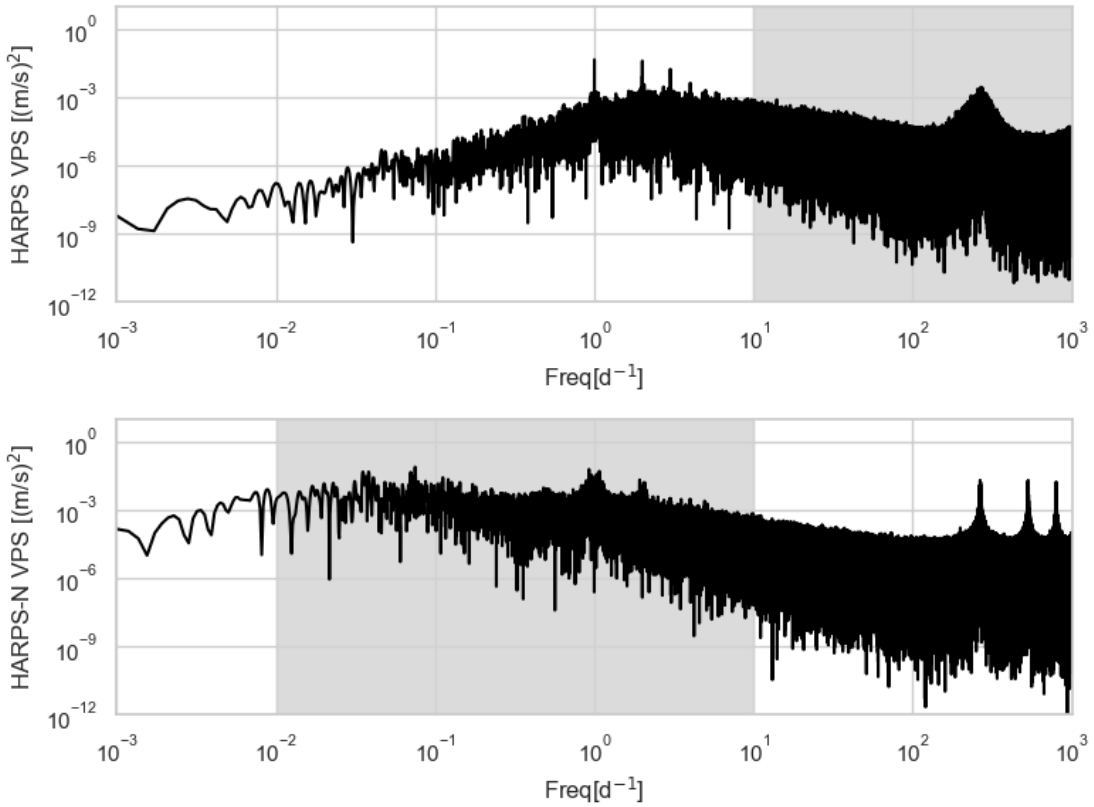


Figure A.3: The VPS obtained from HARPS and HARPS-N data.

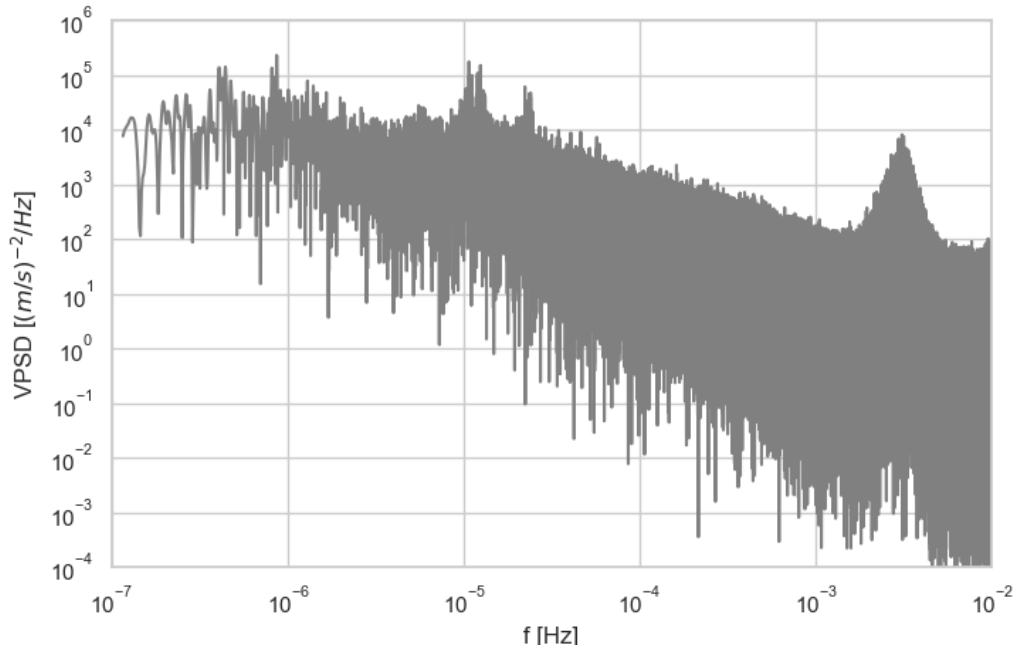


Figure A.4: The VPSD obtained from the combined HARPS and HARPS-N data.

A.4 Modeling the VPSD

Following Al Moulla et al. (2023), two types of analytical functions were chosen to fit the stellar signal components. A Lorentzian function in the form:

$$VPSD_L = A_L \frac{\Gamma^2}{(\nu - \nu_0)^2 + \Gamma^2}, \tag{A.5}$$

was chosen to represent the periodic components where A_L is the amplitude, Γ is the full-width at half maximum (FWHM) and ν_0 is the central frequency. Additionally, a Harvey function in the form:

$$VPSD_H = \frac{A_H}{1 + (\tau\nu)^\alpha}, \tag{A.6}$$

was chosen to fit the granulation phenomena where A_H is the amplitude, τ is the characteristic timescale and α is the power-law slope. To constrain the Harvey functions, α was set to 2 as in Lefebvre et al. (2008). The photon noise was also included as a constant.

Using the reported coefficients from Al Moulla et al. (2023), the analytical functions were overplotted on the VPSD and then adjusted by eyes. They were then combined to produce an initial model for the VPSD. Figure A.5 shows the fitted-by-eyes functions and the modeled VPSD overplotted on the GLS VPSD.

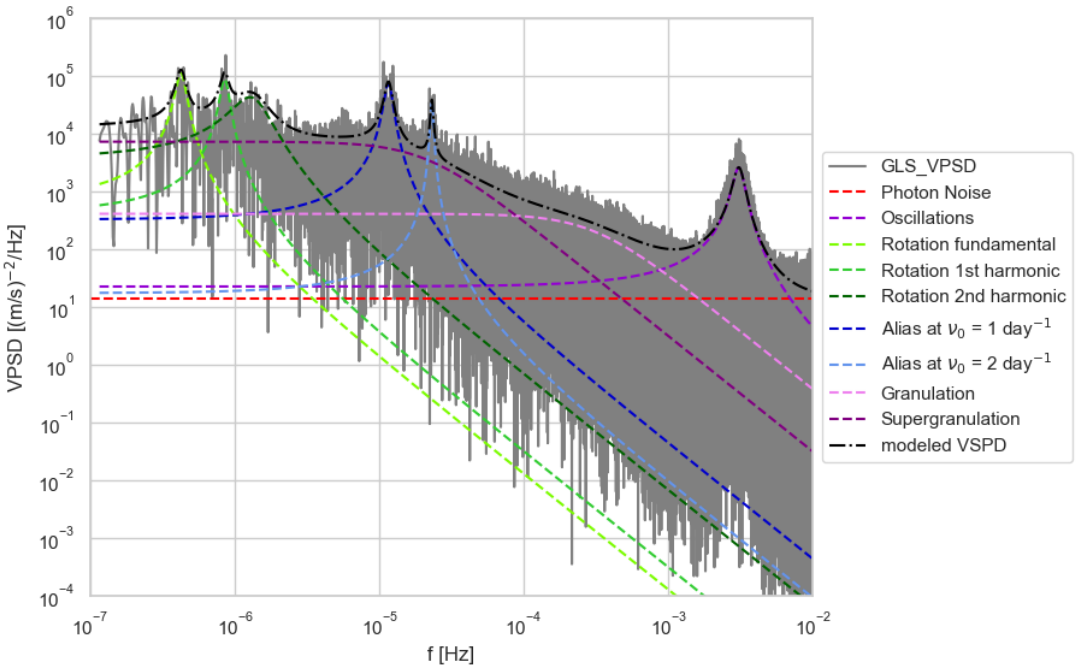


Figure A.5: Over plot of modeled and GLS VPSD.

An attempt was made to optimize the model by fitting the fitted-by-eyes VPSD to the GLS VPSD using a Levenberg-Marquardt (LM) minimization from the lmfit python module (Newville et al., 2021). The LM minimization is an optimization method used primarily for solving nonlinear least squares problems (Roweis, 1996).

Following Al Moulla et al. (2023), the model was optimized with respect to the GLS VPSD averaged on logarithmically equidistant frequency steps. The fitted-by-eyes VPSD was used as the initial guess and max iterations of 50,100,500 was tested. The results are seen in Figure A.6.

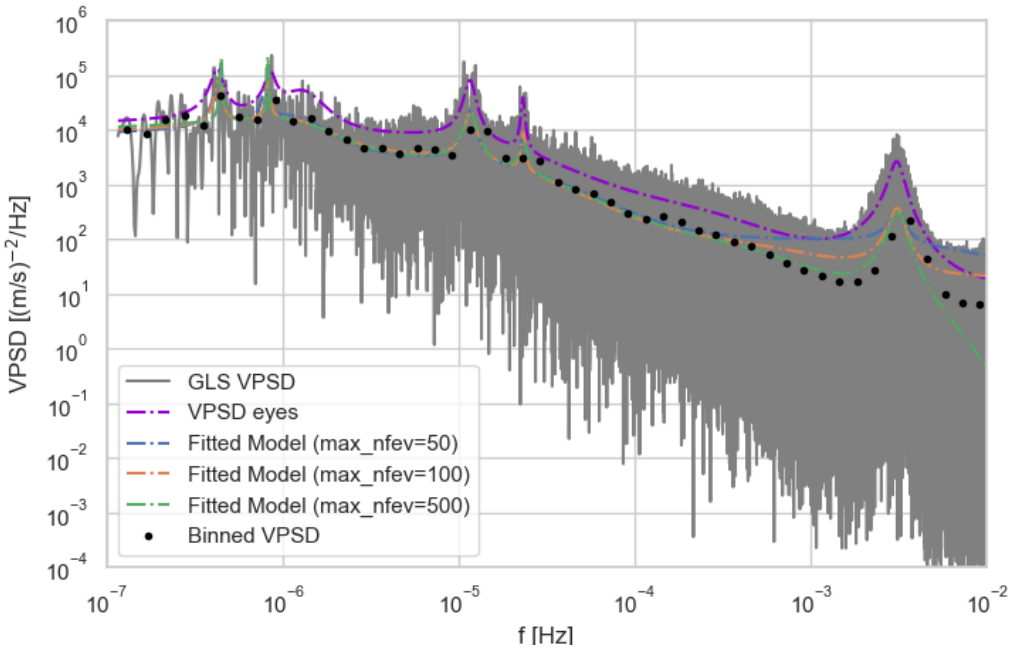


Figure A.6: Log-log plot of VPSD vs Frequencies. The fitted-by-eyes VPSD (purple), log binned VPSD (dots), and the fitted VPSD models with different max iterations can be seen overplotted on the GLS VPSD (grey).

From Figure A.6, it was observed that if the max iterations are set too low, the fitted model is not optimized, however, if too high, the fitted model begins to diverge. The max iterations were therefore set to 100 which produced the most visually accurate fit to the GLS VPSD. Additionally, the validity of using the binned GLS VPSD, represented by the black dots, in the fitting procedure is questioned since it does not appear to follow the bump of the p-mode oscillations. Unfortunately, it was not possible to explore this further during the internship.

A.5 Experimenting with the HARPS-N Solar RV Data

During the internship, two short experiments were performed with the HARPS-N data set, to learn to manipulate and become more familiar with the data set.

As previously discussed, the HARPS-N data set contains two types of stellar activity indicators. Photospheric indicators such as Bis Span, FWHM, and contrast which are derived from the CCF between the stellar spectrum and the binary-weighted line mask, and photospheric indicators such as RHK and S_{MW}. The first experiment involved a visual comparison of the long-term trends due to stellar magnetic activity of the observed RV measurements and the RHK indices. The aim was

to show, similar long-term variability in both sets of data. This was done by normalizing the data and extracting the slow variation due to the 11-year magnetic cycle as outlined in Section A.2. The trends were then normalized to 1 and overplotted along with their difference to visualize the differences seen in Figure A.7.

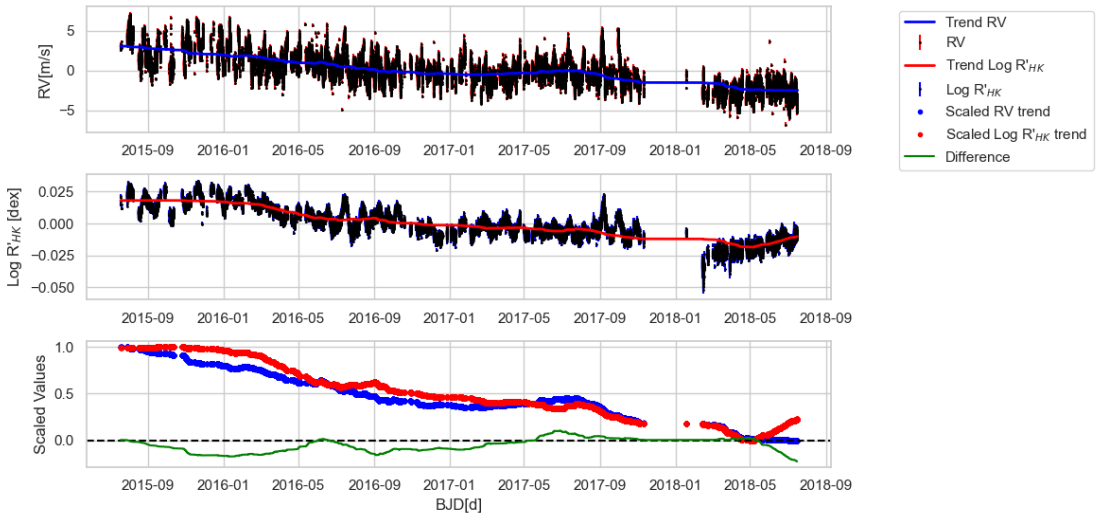


Figure A.7: Comparison of HARPS-N RV and the $\log(R'_{HK})$ solar indicator.

Here, it was observed that large-scale trends in the HARPS-N RV and RHK indices are similar with small variations at low scales. These small variations are expected since the RV measurements are based on photospheric lines while the RHK is based on chromospheric lines, where the changes in the solar magnetic activity between the photosphere and chromosphere would result in small-scale or high-frequency variations. It was also observed that the difference between the trends shows little variation with a standard deviation of 0.08 about a mean of -0.05, reiterating the usefulness of the RHK indices as a solar activity indicator.

As previously discussed, the detrending method can be modified to capture higher frequency variations by increasing the polynomial degree or reducing the window length. Similar to the thesis analysis, other methods can also be applied where multiple indicators are jointly considered thus capturing more variations due to solar activity. One such tool is the modified GLS periodogram where a linear combination of multiple filters along with the sinusoidals and constant is fitted to the data. This method is briefly discussed in Section A.6.

The second experiment performed, aimed to observe the effect of the 11-year magnetic cycle on the GLS periodogram. The HARPS-N data release contains 3 years of RV measurements from 07/2015 to 07/2018, approximately 1/4 of the 11-year solar magnetic cycle 24 (Jiang et al., 2015). It was therefore hypothesized that if the long-term variability due to the magnetic cycle is not removed during the detrending process, it should have a measurable effect on the GLS VPSD. This was tested by splitting the HARPS-N data into three 1-year sets, calculating the GLS VPSD for each, and comparing the estimated parameters that describe the GLS VPSD. The process was repeated for the RHK data to verify that any changes were due to the solar magnetic activity cycle.

From Figure A.8 it was observed that some models were not properly fitted.

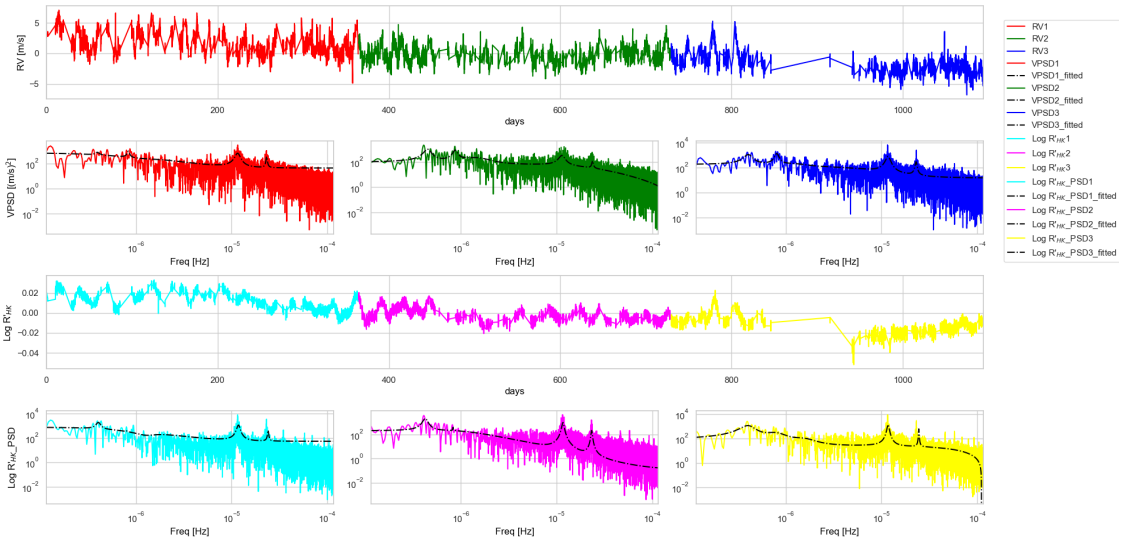


Figure A.8: The first row shows the HARPS-N RV time series divided into 3 data sets, RV1 (red), RV2 (green), and RV3(blue), each 1 year long. Row 2 contains their corresponding VPSD with an over plot of the fitted modeled VPSD. Rows 3 and 4 are the RHK equivalent to rows 1 and 2.

This was due to limitations in the fitting procedure described in Section A.4 such as its sensitivity to the initial model as well as the tendency to diverge with too many iterations. Unfortunately, these limitations could not addressed during the internship, however, an attempt was still made to compare the fitted models as seen in Figure A.9.

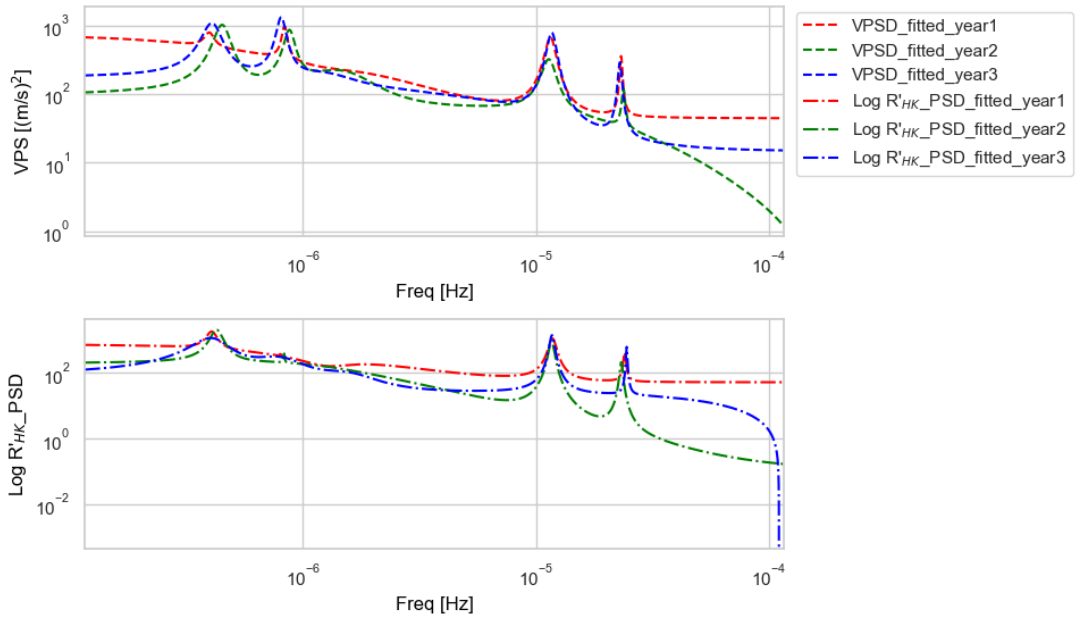


Figure A.9: Comparison of fitted model VPSD and RHK PSD.

The lack of clear trends in Figure A.9 over the three years may be due to several factors including the limitations in the fitting procedure. Since the available

HARPS-N data was collected as the solar cycle was entering a minimum, it is possible the long-term variations were too small to be captured by this analysis. It would be better to test the largest long-term variations during the solar maximum to minimum transition or vice versa. The upcoming release of three more years of HARPS-N data (2018-2021) will provide greater coverage of the 11-year solar magnetic cycle which can increase the possibility of capturing long-term variations thus improving this analysis.

A.6 Contributions to Sulis et al. (2024)

The final part of the internship involved a small contribution to "A low-mass sub-Neptune planet transiting the bright active star HD 73344" by Sulis et al. (2024). HD 73344 is a bright F star with a sub-Neptune planet candidate orbiting at ≈ 15 days. Sulis et al. (2024) proposes the detection of a second planetary candidate with a period of 66.45 days. My contribution involved using the tools learned during the internship to perform various tests on SOPHIE RV data to rule out false confirmation of this second candidate due to stellar activity and other processes. The data provided included the SOPHIE RV data and several stellar activity indicators such as ΔFWHM, Area, Bisector, log R'_{HK}, and H-α.

The first test compared the GLS periodograms of the unbinned and binned data to observe the effect of binning on the GLS periodograms. This was done for all the binned and unbinned data provided as well as an additional linear trend as seen in Figure A.10. From this, it was observed that binning the data introduced a small power in the GLS periodogram at the period of interest, PoI, however, the power introduced was too small to account for the peak at the PoI. It was also noted that binning introduced power fluctuations at shorter periods which can lead to false detections. Therefore, it is recommended to avoid binning when possible to avoid contamination of the periodogram.

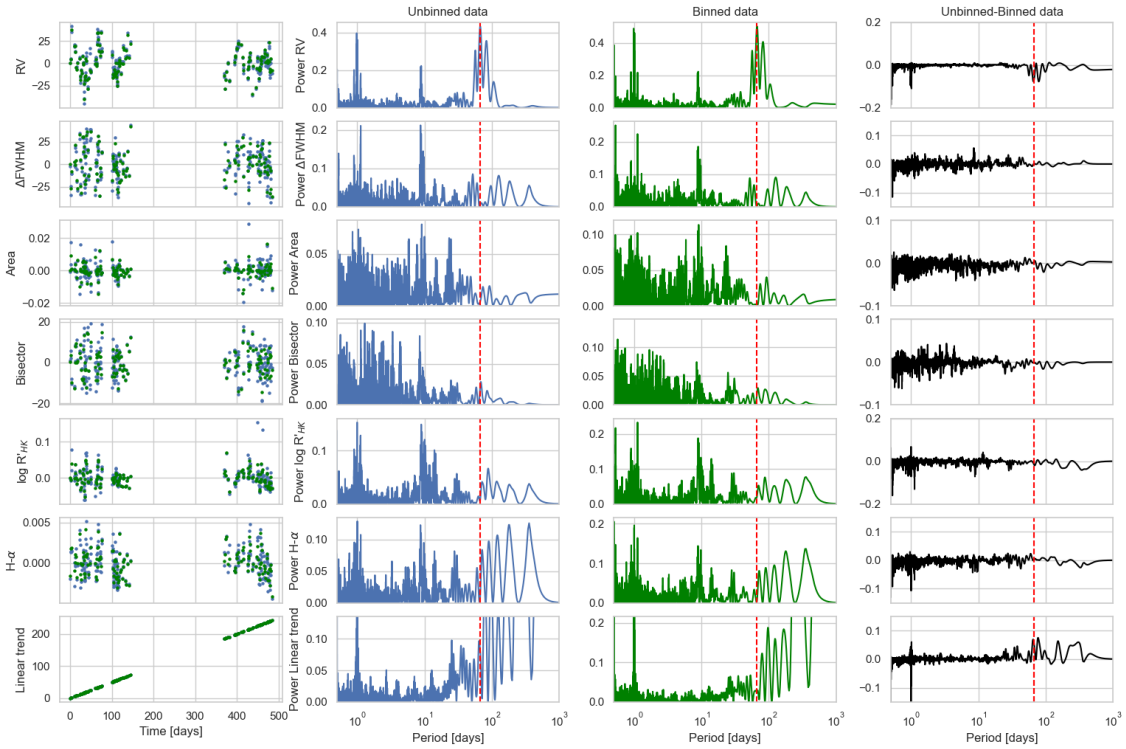


Figure A.10: The first column displays the unbinned and binned time series plots. The second and third columns contain the corresponding GLS periodograms for the unbinned and binned data respectively. The final column contains the difference between the unbinned and binned GLS periodograms. The red vertical line indicates the period of the suspected planet.

As seen in the first column of Figure A.10, the data was obtained in two campaigns (C1 and C2) some time apart. The second test aimed to verify the coherence of the periodic signals between the campaigns. In Figure A.11, extracted from Sulis et al. (2024), a small peak can be seen at the PoI in some of the stellar indicators, and the linear trend in C1 and C2, however, it is no longer present when the campaigns are joined. This suggests that the phase of the signal observed in C1 and C2 may be out of phase and therefore not from the same source (planet). This theory was tested by calculating the phase of the signal at the period of interest for C1, C2, and C1 + C2. The periodic signal for each phase was then overplotted on the RV measurements for a visual comparison as seen in Figure A.12. While the calculated phases appear to be different, visually, the signals appear to be coherent confirming that they are from the same source. The small variation seen may be due to instrumental noise.

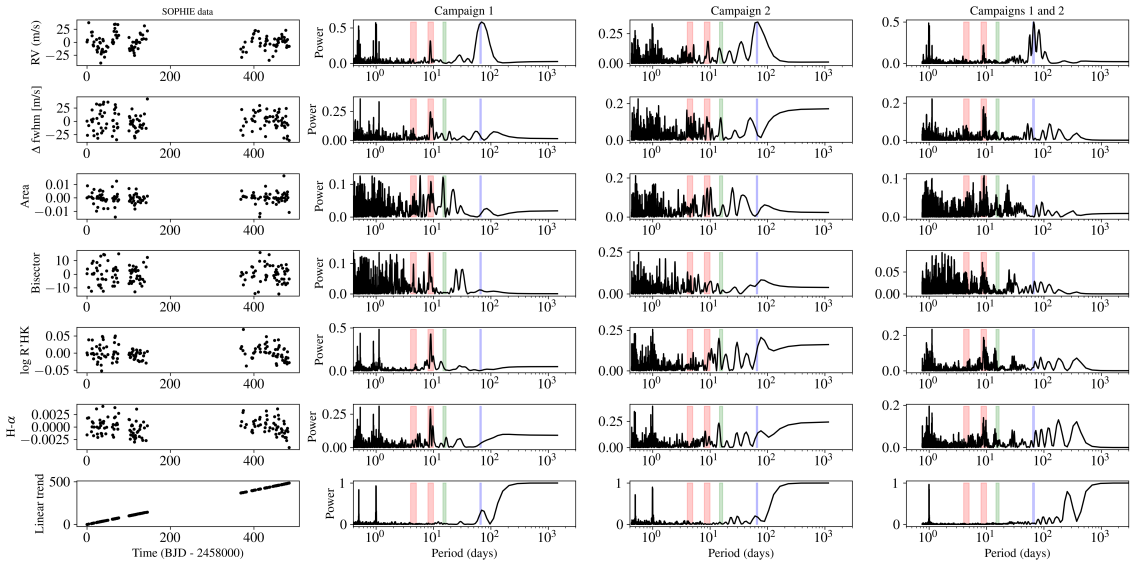


Figure A.11: Temporal evolution of the stellar activity. From top to bottom: RV data, FWHM, the Area of the Gaussian fit to the CCF, Bisector, $\log(R'_{HK})$, Hα lines, and linear trend. From left to right: Time series, GLSP for the first campaign, second campaign, and both campaigns. In all the time series, the long-term variation has been eliminated by a two-degree polynomial fit. The large red, green, and blue vertical lines indicate the stellar rotation period (and half it's period), the orbital period of the transiting planet (b), and the orbital period of the candidate planet (c), respectively (Sulis et al., 2024).

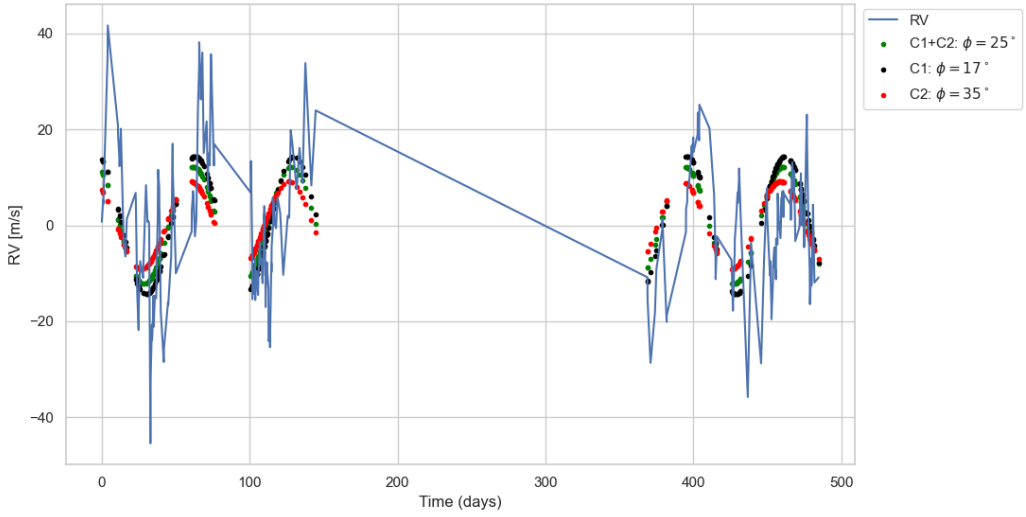


Figure A.12: Periodic signal with calculated phases from C1, C1, and C1 + C2 overplotted on SOPHIA RV measurements.

The final test was more closely related to the work done in this thesis. Using four different models, the VPS was calculated for the SOPHIE RV using a modified GLS periodogram and compared as seen in Figure A.13. The models used contained a linear combination of:

1. Sinusoids
2. Sinusoids + solar activity indicators
3. Sinusoids + a constant + solar activity indicators
4. Sinusoids + a constant + a linear trend + solar activity indicators

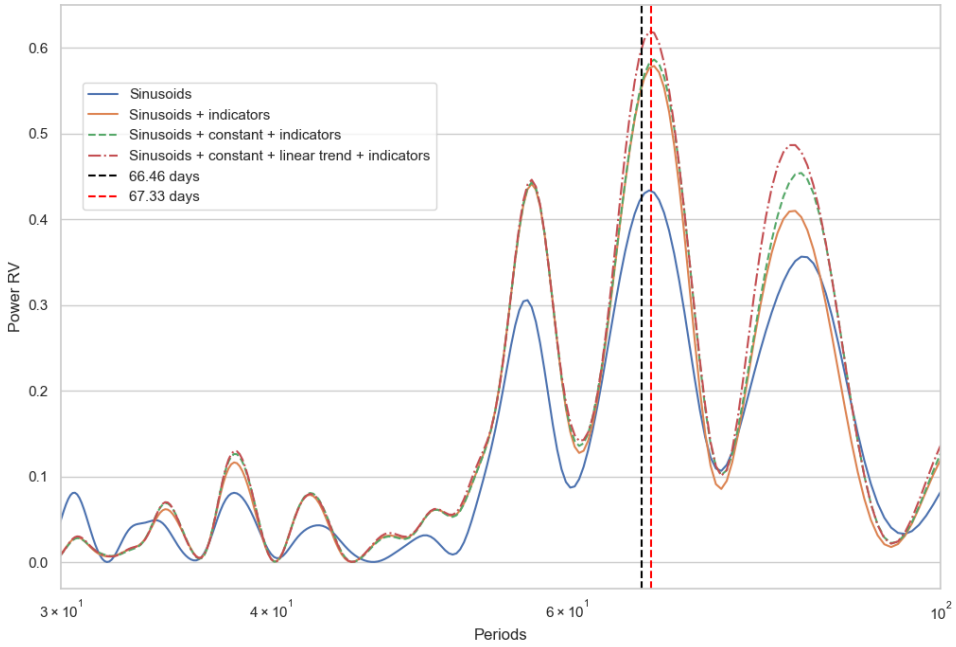


Figure A.13: Comparison of the VPS of the SOPHIE RV calculated with a modified GLS periodogram with four different models. The black vertical line indicates the PoI (66.46 days) as suggested by Sulis et al. (2024) while the red line is the period (67.33 days) where peaks were observed.

From Figure A.13, we observe the peak in all the VPS, which indicates that the signal at this period exists in the RV measurements but not in any of the models tested including the stellar indicator models. While this does not confirm the presence of an exoplanet, it confirms that the signal at this PoI is not due to any systematic error or stellar activity.

Acknowledgements

Firstly, I would like to thank the universe for all the twists and turns, challenges, and adventures that have led me to this exact moment in space and time. Even if I complain profusely, I would not do anything differently if it were to change the outcome.

To the EMJM MASS Selection committee and anyone else involved in awarding these scholarships, I thank you for seeing potential in me and providing not only the opportunity to study and advance my career but also for giving me the chance to explore new cultures, expand my community, and develop as a global citizen. The MASS program has not only helped me grow academically but also as a person and for that, I am grateful.

There are many persons without whom I would not have been able to complete this research. I would like to express my gratitude to my supervisors, Prof. Francesco Berrilli, University of Rome (Tor Vergata), and Dr. Lionel Bigot, UMR Lagrange, Observatoire de la Côte d'Azur, for guiding me through my thesis research and internship, for being understanding, encouraging, and allowing me the freedom to work at my own pace when needed. I would also like to thank Dr. David Mary, Laboratoire Lagrange, Université Côte d'Azur, who co-supervised my internship, for his patience, understanding, and knowledge. I would also like to express a special thank you to Dr. Mary and Dr. Sophia Sulis, Université Aix Marseille, for the opportunity to contribute to an academic publication during my internship.

To my friends and family, who have encouraged and kept me motivated through this program; those in my home country of Trinidad and Tobago, who've kept sharing a little slice of our paradise so I feel a little less homesick; to the MASS OG's, edition 1 students whom I was lucky enough to struggle, persevered, and celebrate alongside, and all the other friends I've made over the last two years, I thank you.

I would also like to acknowledge the efforts of the MASS Project Office for all their work and efforts in making our time as part of MASS as smooth and unproblematic as possible; particularly, Dr. Adalia Caroli who from the moment we were accepted into the program, has been in contact with us, answering all our questions, guiding us through administrative hurdles even while in other countries. I am truly grateful for all the kindness and support I have received throughout this program.

Finally, I would like to dedicate this thesis to my grandmother, who unfortunately returned to the universe earlier this semester. She played an important role in my upbringing and I would not be here today without her, Thank you.

Brandon Rajkumar acknowledges support through an Erasmus Mundus Joint Master (EMJM) funded by the European Union in the framework of the Erasmus+, Erasmus Mundus Joint Master in Astrophysics and Space Science – MASS. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority European Education and Culture Executive Agency (EACEA). Neither the European Union nor the granting authority can be held responsible for them.