



## **Department of Space and Climate Physics**

**Individual Project – Final Report**

**Academic Year 2021 – 2022**

# **The Identification and Classification of Magnetic Switchbacks Using Statistical and Machine Learning Methods**

**Submitted in partial fulfilment of the requirements for the award of  
the degree of**

**MSc – Space Science and Engineering**

**Submitted by**

**21154540**

**Supervisor**

**Dr Daniel Verscharen**

**Dr Andy Smith**

**08/07/2022**

**Word count: 11 956**

# Abstract

The solar wind is a very complex and important solar phenomenon due to its associations with space weather and the interplanetary influence of the Sun. Upon its first close encounter with the Sun the in-situ magnetometers on board the Parker Solar Probe detected a significant amount of field reversals within the subtended interplanetary magnetic field. These events are known as magnetic switchbacks and are of vital importance to characterise as a result of their associated impacts on the surrounding plasma environment and subsequent interplanetary space. Thus due to this characterisation need, we aim to solve this problem by utilising statistical and machine learning techniques which have become increasingly popular in modern solar and plasma physics. In this report we investigate the appropriateness of using unsupervised machine learning as a statistical instrument to identify and characterise these solar events. We achieve this analysis by proposing novel evaluation techniques to investigate the benefits of classifying magnetic switchbacks in this manner. In this experimental analysis, these findings show that the density and probabilistic, algorithmic architectures perform best on the provided data. From these observations, we conclude that of the evaluated models and encounters, the DBSCAN and Bayesian Gaussian Mixture models are the most ideal algorithms for the application of switchback classification. These models produce predictions with accuracies in excess of 90 %, as well as possessing highly consistent and reliable predictions. The implications of our findings show significant potential to be implemented in an assortment of solar physical applications and in the field of space science as a whole.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Glossary</b>	<b>v</b>
<b>Nomenclature</b>	<b>vii</b>
<b>List of Abbreviations</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>3</b>
2.1 The Solar Wind . . . . .	3
2.2 Magnetic Switchbacks . . . . .	4
2.3 Parker Solar Probe . . . . .	6
2.4 Machine Learning in The Space Environment . . . . .	9
2.5 Machine Learning Methods . . . . .	10
<b>3 Materials and Methods</b>	<b>12</b>
3.1 Data Acquisition . . . . .	12
3.2 Feature Engineering . . . . .	14
3.3 Feature Scaling . . . . .	15
3.4 Statistical Modal Analysis . . . . .	16
3.5 Unsupervised Machine Learning Models . . . . .	17
3.5.1 K-Means . . . . .	17
3.5.2 DBSCAN . . . . .	19
3.5.3 Agglomerative Hierarchical Clustering . . . . .	20
3.5.4 Bayesian Gaussian Mixtures . . . . .	22
<b>4 Results and Discussion</b>	<b>25</b>
4.1 Prediction Evaluation . . . . .	25
4.2 Cross-Model Comparison . . . . .	26
4.3 Orbital Analysis . . . . .	29
4.4 Discussion . . . . .	33
<b>5 Conclusions</b>	<b>35</b>
<b>References</b>	<b>37</b>

<b>Appendix I - Comprehensive Data Capture</b>	<b>41</b>
Magnetic Field Data Acquisition Process . . . . .	41
Second and Fourth Encounter . . . . .	41
Proximity Data Acquisition Process . . . . .	44
Solar Wind Data Acquisition . . . . .	44
Extended Interpretation of Statistical Modes . . . . .	45
<b>Appendix II - Model Architecture &amp; Theory</b>	<b>47</b>
K-Means Architecture . . . . .	47
DBSCAN Architecture . . . . .	49
Agglomerative Hierarchical Model Architecture . . . . .	50
Bayesian Gaussian Mixture Model Architecture . . . . .	51
<b>Appendix III - Detailed Results Analysis</b>	<b>53</b>
Thorough Prediction Evaluation . . . . .	53
Complete Cross-Model Comparison . . . . .	55

# List of Figures

1	Illustration of magnetic switchback in interplanetary magnetic field . . . . .	5
2	Trajectory orientation diagram of Parker Solar Probe . . . . .	8
3	Raw magnetic field vector components . . . . .	13
4	Statistical modal analysis plot . . . . .	16
5	K-Means clustering functionality on specimen data sample . . . . .	18
6	DBSCAN clustering functionality on specimen data sample . . . . .	19
7	Dendrogram illustrating agglomerative hierarchical clustering method . .	21
8	Gaussian contour plot illustrating probabilistic Bayesian mixture clustering	22
9	Unsupervised model prediction overlay for third encounter . . . . .	25
10	Model evaluation over varying statistical mode cadences . . . . .	28
11	Machine Learning prediction orbital trajectory overlay plots . . . . .	30
12	Orbital overlay plot with added solar wind feature . . . . .	32
13	Raw magnetic field data for second encounter . . . . .	42
14	Raw magnetic field data for fourth encounter . . . . .	43
15	Statistical modal analysis of second and fourth encounters . . . . .	45
16	Algorithm showing the initialisation of the K-Means model . . . . .	48
17	Algorithm showing the initialisation of the DBSCAN model . . . . .	49
18	Algorithm showing the initialisation of the Hierarchical model . . . . .	50
19	Algorithm showing the initialisation of the Gaussian Mixture model . .	51
20	Model prediction overlays on second and fourth encounters . . . . .	54
21	Cross-model comparison results for second encounter prediction data . .	55
22	Cross-model comparison results for fourth encounter prediction data . .	56

# List of Tables

1	Characteristics of all evaluated Parker Solar Probe orbital encounters . .	12
2	Initial features and boundary metrics used in interim model evaluation .	15
3	Cross-model classification statistics for differing mode cadences and en- counters . . . . .	27

# Glossary

*Agglomerative Hierarchical Model*

- An unsupervised analysis method that develops clusters based on a bottom-up approach of instance merging.

*Alfvén Wave*

- A magnetohydrodynamic oscillation that originates from the restorative tensile forces of curved magnetic fields.

*Bayesian Gaussian Mixture*

- A probabilistic machine learning algorithm that uses variational Bayesian inference to cluster data into prior assumed groups of Gaussian distributions.

*Density-Based Spatial Clustering of Applications with Noise*

- A non-parametric unsupervised learning algorithm that develops clusters based on the relative densities of instances within a dataset.

*Electrostatic Analyser*

- An instrument which detects the specific energy properties of charged particles by using electrostatic forces to regulate the passage of ions and electrons.

*Expedition-Maximisation*

- An iterative method which aims to maximise a prior estimate of a set of statistical parameters.

*Faraday Cup*

- A simple electrostatic instrument designed to measure the flow rate of charged particles in a vacuum.

*Fluxgate Magnetometer*

- An active measurement device that records the absolute background magnetic field using principles of hysteresis and magnetic saturation.

*Hyper-Parameter*

- Parameters that can be modified to change the learning outcome of a machine learning algorithm.

*K-Means*

- An unsupervised learning algorithm that develops its clusters based on an instance's specific proximity relative to a mean cluster centroid.

*Magnetic Switchback*

- A sudden inversion and rectification in the polarity of the interplanetary magnetic field.

- Magnetohydrodynamic Fluid*
- A highly conducting fluid that possess electrical properties which are governed by Lorenz forces.
- Parker Solar Probe*
- A NASA space probe launched in 2018 with the primary intent of studying the plasma environment surrounding the Sun.
- Search Coil Magnetometer*
- A magnetometer that uses inductive sensing to measures the relative change in magnetic field density.
- Solar Wind*
- A constant stream of charged particles flowing radially outward from the Sun to the edge of the solar system.
- Unsupervised Machine Learning*
- A class of machine learning algorithms that generate predictions based on internal data characteristics within the provided feature set.

# Nomenclature

## Greek Characters

$\alpha\Pi$  - Positive pulse duration

$\beta\Pi$  - Negative pulse duration

$\Delta\mathcal{E}_s$  - Peak-to-peak amplitude

$\Delta H$  - Measurable external field contribution

$\Delta I$  - Current for particles with velocity greater than cut-off threshold

$\mathcal{E}_0$  - Electromotive force in drive coil

$\mathcal{E}_o$  - Offset voltage

$\mathcal{E}_s$  - Induced EMF in sense coil

$\mu_i$  - Mean of cluster sum-of-squares

$\Pi$  - Pulse period

$\Sigma_i$  - Covariance matrix of  $i^{th}$  distribution

$\phi_i$  - Class weights of  $i^{th}$  distribution

## Roman Characters

$A$	- Core cross-sectional area
$\vec{B}$	- Magnetic field vector
$B_R$	- Radial magnetic field
$B_N$	- Normal magnetic field
$B_T$	- Tangential magnetic field
$\vec{E}$	- Electric field vector
$f_j$	- Particle distribution function for all plasma species $j$
$H_0$	- Auxiliary saturation magnetic field strength
$H_c$	- Hysteresis field strength offset
$I$	- Clustering inertia
$m_j$	- Plasma particle mass
$\mathcal{N}$	- Number of coil windings
$\mathcal{N}(\mu, \sigma)$	- Normal distribution function with mean $\mu$ and standard deviation $\sigma$
$q_j$	- Plasma particle charge
$r$	- Heliocentric distance
$R_\odot$	- Solar radius
$S$	- Cluster sum-of-squares
$t$	- Time
$\vec{v}$	- Plasma particle velocity
$\vec{x}$	- Plasma particle spatial position
$\mathbf{x}$	- Feature space

# List of Abbreviations

<i>DBSCAN</i>	- Density-Based Spatial Clustering of Applications with Noise
<i>EM</i>	- Expectation-Maximisation
<i>EMF</i>	- Electromotive Force
<i>FIELDS</i>	- Electromagnetic Fields Investigation <i>FIELDS</i> is an instrument suite onboard the Parker Solar Probe. It is primarily designed to measure the magnetic field environment around the solar corona.
<i>GM</i>	- Gaussian Mixture
<i>MHD</i>	- Magnetohydrodynamic
<i>ML</i>	- Machine Learning
<i>OOS</i>	- Out of Sample
<i>PSP</i>	- Parker Solar Probe
<i>RTN</i>	- Radial-Tangential-Normal
<i>SCM</i>	- Search Coil Magnetometer
<i>SPC</i>	- Solar Probe Cup
<i>SWEAP</i>	- Solar Wind Electrons Alphas and Protons <i>SWEAP</i> is an instrument suite onboard the Parker Solar Probe. It is primarily designed to measure the plasma environment that constitutes the solar wind.

# 1 Introduction

Since its theorisation in 1957 and its first in-situ detection in 1959, the solar wind has been the subject of study for numerous heliophysical processes (Obridko & Vaisberg 2017). Due to the importance of this phenomenon, the solar wind has been studied in great detail through various means of in-situ and remote sensing techniques. This dominant physical structure is comprised of various plasma species, and consequential magnetic fields which flow radially outward; propelled by the Sun's outer corona (Parker 1965). The solar wind is an essential medium through which solar activity is transferred to interplanetary space. Most notably this wind is the primary means through which Alfvén waves propagate (Alfvén 1942). Alfvén waves and the associated magnetic fields are an important energy transfer mechanism within the solar wind. These turbulent interactions are significant drivers of the magnetohydrodynamic processes observed at close heliocentric distances, and thus it is essential to understand the causes and consequences of these phenomena.

Upon its first solar encounter in 2018 the in-situ instrumentation onboard the Parker Solar Probe (*PSP*) spacecraft detailed the occurrence of rapid magnetic field reversals in the immediate environment surrounding the Sun. These frequent field inversions occurred most prominently beyond the Alfvén critical surface (Bourouaine & Perez 2019, Tsinganos et al. 1996). These events are referred to as magnetic switchbacks and have profound effects on the energy characteristics of the solar wind within the immediate surrounding of these magnetic structures (Kasper et al. 2019). During the occurrence of a switchback the polarity of the magnetic fields switches half a rotation before rapidly realigning. It has also been noted that the thermal and kinetic energy characteristics of the solar wind increases significantly because of these switchbacks. The causes relating to the formation of these events are currently unknown. Thus, there is a growing need to further understand the physical factors pertaining to these switchback events, and their effects on the dynamic evolution of the heliophysical environment.

Originally named Solar Probe Plus, PSP has advanced ever closer to the Sun since its launch in 2018 (Whittlesey et al. 2020). The primary goal of Parker Solar Probe is to study the plasma and in-situ environment around the Sun and solar corona. This probe is designed to complete 24 orbital encounters before concluding the mission in the year 2025, where it will aid in developing our insight and understanding of the complex processes that define solar physics. The significance of Parker Solar Probe's close heliocentric position enables the spacecraft to capture heliophysical data at an unprecedented level. This level of scientific detail, along with the high precision instrumentation onboard has allowed PSP to make momentous discoveries and findings (most notable that of magnetic switchbacks). To achieve such close proximity to the Sun, PSP utilises a series of gravitational assists from Venus to place it in an elliptical orbit with an eventual 88 day period (Guo 2008). This highly eccentric orbit provides the spacecraft with a sufficiently close encounter enabling various components of the solar wind and solar magnetic field to be studied.

## 1 INTRODUCTION

Due to the turbulent nature of the solar wind and frozen-in magnetic fields, it can often be very difficult to characterise these events, on localised scales, using theoretical mathematical modelling. In such cases seemingly non-algorithmic problems are too complex to characterise through conventional means. To overcome these obstacles Machine Learning (*ML*) practices can be adopted. Machine learning has the ability to identify statistical trends and features in complex and high dimensionality datasets. This characteristic makes *ML* a useful tool in heliophysics. Thus in the context of classifying magnetic switchbacks it may be beneficial to apply machine learning techniques, as a means of resolving interesting properties from the stochastic magnetic field data.

Due to the recent discovery and data acquisition of magnetic switchbacks, very little is known and researched regarding the subject. As a result of this lack of understanding, there is no definitive method of identification and characterisation (Froment et al. 2021). This gap in our scientific understanding is significant, as in order to properly understand the inner functionality of the solar wind and associated heliophysical environment these solar-magnetic phenomena must be properly characterised and understood. Thus, we chose to undertake this project in order to address some of these shortfalls and difficulties.

In this report we investigate the possibility, and extent to which magnetic switchbacks can be identified, classified, and statistically evaluated using the principles and practices of machine learning. This scientific analysis was performed using data sourced from instruments onboard PSP. Throughout the course of this project we outlined the method, tools, and findings used to reproduce our results. This research will uncover new information pertaining to the anomalistic events of magnetic switchbacks, as well as develop a statistically proven method of switchback classification.

## 2 Literature Review

### 2.1 The Solar Wind

The solar wind is a large scale, and complex phenomenon that spans the length of the solar system. By studying this wind and its constituent traits, an understanding can be developed as to how the heliospheric environment and solar activity is coupled/projected to interplanetary space. The solar wind is a continuous outward expanding plasma that is largely propelled by the heat of the Sun's corona (Parker 1965). This plasma stream consists of a variety of ion species which originate from the Sun and flow radially, towards the edge of the solar system. As the wind expands outward the bulk energy properties, most notably the acceleration, density, and speed of Alfvén waves<sup>1</sup> decreases (Bavassano & Bruno 1992). Concurrently, these wind properties show significant non-uniformity and transient features which play crucial roles in the development of the immediate heliospheric environment surrounding the Sun (Kasper et al. 2021). The solar wind can broadly be classified under three bulk categories (namely fast, slow, and eruptive)<sup>2</sup>, which originate from different sources and solar regions (Verscharen et al. 2019). These differing energy properties lead to the complex and extensive plasma which we observe.

This constantly evolving plasma flow behaves predictably on a macroscopic level, and have shown to that of a Magnetohydrodynamic (*MHD*) system (Roberts et al. 1991). Due to these observed properties, this turbulent plasma can be modelled using known MHD equations. Based on the assumption that the motion of the plasma in the solar wind is governed purely by electromagnetic forces, the particle interactions of this wind can be defined by the Vlasov equation (Verscharen et al. 2019):

$$\frac{\partial f_j}{\partial t} + \vec{v} \cdot \frac{\partial f_j}{\partial \vec{x}} + \frac{q_j}{m_j} (\vec{E} + \vec{v} \times \vec{B}) \cdot \frac{\partial f_j}{\partial \vec{v}} = 0 \quad (1)$$

Where:

$f_j$  = Particle distribution function for all plasma species  $j$

$t$  = Time [s]

$\vec{v}$  = Plasma particle velocity [m/s]

$\vec{x}$  = Plasma particle spatial position [m]

$q_j$  = Plasma particle charge [C]

$m_j$  = Plasma particle mass [kg]

$\vec{E}$  = Electric field vector [N/C]

$\vec{B}$  = Magnetic field vector [T]

---

<sup>1</sup>Alfvén waves are magnetohydrodynamic oscillations that originate from the restorative tensile forces of curved magnetic fields (Keiling 2021).

<sup>2</sup>Although these names pertain to the velocities of the plasma streams; it must be noted that the flow rates alone do not exclusively determine wind type.

By treating this plasma in the framework of MHD, the motion of the solar wind can be defined by a set of conservation laws. As outlined by equation (1), energy is primarily transferred between the different macro and microscopic scales of the solar wind through interactions with the electromagnetic field, which is often turbulent (Bavassano & Bruno 1992). This turbulent energy transfer is brought on through the process of eddy cascading, where large turbulent flow structures transfers energy to faster decaying micro eddy currents (Bavassano & Bruno 1992). On a macroscopic scale this mechanism of energy transfer acts to uniformly distribute the energy and flow characteristics (e.g. density, temperature, and velocity) of the plasma throughout the solar wind as the flow develops (Roberts et al. 1991). Corollary, at relatively close heliocentric distances the transient flow of the solar wind is significantly more prominent and dominates the physical characteristics that govern the behaviour of the solar wind (Kasper et al. 2021).

For an ideal MHD fluid the magnetic field exhibits a behaviour known as frozen-in-flow, whereby the plasma remains associated with the same magnetic field lines (i.e. the magnetic field lines within a plasma are convected with the flow)<sup>3</sup> (Alfvén 1942). For large scale, well-established MHD plasmas that possess gradually changing magnetic field topologies; frozen-in-flow can be discerned. This frozen-in-flow behaviour can be observed in the solar wind, whereby the interplanetary magnetic field (established by the Sun) is carried along with the wind. The motion of this plasma results in the frozen-in magnetic field forming a distinct shape (commonly referred to as the Parker spiral), and follows principles characterised by Alfvéns' theorem (Alfvén 1942). As mentioned, at close heliocentric distances the characteristics of the interplanetary magnetic field are greatly influenced by the more turbulent plasma which constitute solar wind, and thus breakdown the frozen-in-flow approximation (leading to the frozen-out-flow condition) (Bavassano & Bruno 1992). Likewise these fields can have considerable effects on the properties of the surrounding plasma environment. Thus it is of high importance to understand the relational interactions between the solar wind and interplanetary magnetic field.

## 2.2 Magnetic Switchbacks

From the Vlasov equation (1) it can be seen that the magnetic field plays a vital role in governing the motion of plasma particles. The complex magnetic field generated by the Sun, extends towards the outer edges of the solar system, where it acts to direct the flow of particles subtended by the solar wind (Garner 2016). Throughout the solar wind this interplanetary magnetic field is the main driver of energy and momentum transfer, and is hence a crucial component in heliophysics (Solanki et al. 2006).

From recent near-Sun observations, a significant presence of Alfvén waves has been identified around coronal holes primarily near the Sun's equatorial regions (Bale et al. 2019). Due to the relational characteristics of the solar wind and interplanetary magnetic field, coronal holes are typically associated with fast solar wind velocities (Verscharen et al. 2019). This observation is significant as typically the equatorial region of the Sun is dominated by relatively slow wind speeds. Some theories speculate that these Alfvén waves originate as a result of fast solar wind overtaking relatively slower plasma flows (Hatfield 2021). Thus these theories align with observations relating to the phenomenon, and seem to justify their occurrence.

---

<sup>3</sup>This phenomenon is known as Alfvéns' theorem and breaks down for small-scale turbulent plasma environments, resulting in magnetic reconnection and the subsequent mixing of plasma species.

In its extreme case, these Alfvénic structures have been observed to kink and fully reverse in magnetic polarity (McManus et al. 2020). These sporadic events have been noted most prominently at relatively close solar proximities ( $\lesssim 60 R_{\odot}$ ), and tend to align themselves radially with respect to the reference of the Sun (Bale et al. 2019, Bourouaine & Perez 2019). These sporadic events have been referred to as magnetic switchbacks, and the cause/consequences of such structures are unknown (Bandyopadhyay et al. 2020). However, from recent observations and studies into magnetic switchbacks, several interesting properties have been discovered. Figure 1 elucidates the observed characteristics of these magnetic structures and their prominence around the Sun.

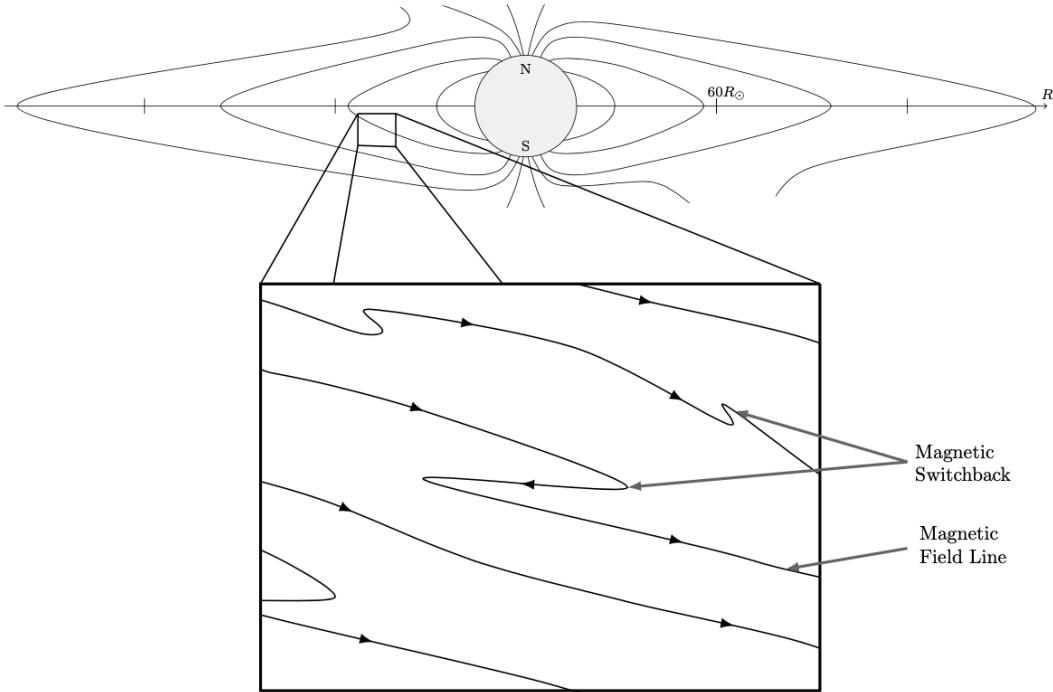


Figure 1: Illustration detailing the scale and structural variation of magnetic switchbacks. In the image notable features of the switchbacks can be identified such as the equatorial location, and heliocentric distances (indicated as being within  $60 R_{\odot}$ ) where such events are most commonly associated. It is also evident that these structures form complete polarity reversals with varying size, scale, and duration.

From figure 1, it is apparent that these switchbacks have variable magnitudes and scales. From past observations, these switchback events have been identified to last between a few second to several minutes, the magnitude of which dissipates as the Alfvén waves<sup>4</sup> propagate into interplanetary space (Bale et al. 2016). These magnetic switchbacks occur more frequently, and with greater extent closer towards the Sun (McManus et al. 2020). The scarp magnetic gradients associated with these events leads to frozen-out-flow, resulting in the rapid transfer of kinetic energy to the surrounding solar wind (Kasper et al. 2019). This residual energy is imparted to the plasma environment when the polarity of these switchbacks re-invert, resulting in localised wind velocity spikes (Kasper et al. 2019). These events are potentially a major contributing source of energy for the driving of the solar wind, and heating of the Sun's corona, and thus it is imperative to characterise and understand these heliomagnetic occurrences (Farglette et al. 2021).

<sup>4</sup>These magnetic switchbacks are likely related to Alfvén waves due to the observed similarities in flow velocity and the magnetic field correlations.

Due to the relatively sporadic and random nature of magnetic switchbacks it can often be very unintuitive/difficult to classify a distinct event. Thus, various techniques of statistical categorisation have been implemented. One such method involves the use of statistical modes highlighted by Bale et al. (2019), whereby the modes of the magnetic field data are calculated from a series of container subsets (*see Materials and Methods*). This form of modal analysis has the advantage of isolating/mitigating the momentary spikes associated with switchbacks, and subsequently identifying the ‘true’ trend of the interplanetary magnetic field. The application of statistical modes is a proven method of switchback analysis and can be a useful tool in the classification of these events.

## 2.3 Parker Solar Probe

The Parker Solar Probe is a heliospheric research satellite designed to take in-situ measurements of the plasma environment immediately surrounding the Sun’s corona (Jannet et al. 2021). PSP has a highly eccentric orbit with an eventual perigee equivalent to 8.85 solar radii and has collected data closer to the Sun than any foregoingspace mission since its launch in 2018 (Jannet et al. 2021). The Parker Solar Probe is fitted with four instrument suites (primarily in-situ) adapted to measuring various properties of the solar wind at close heliocentric distances. These various instruments allow the probe to measure an assortment of different physical factors relating to the Sun; such as interplanetary plasma and magnetic fields, electromagnetic fluctuations, and solar transients (Bale et al. 2016).

In order to accurately measure the magnetic field PSP makes use of several magnetometers which form part of the FIELDS suite. Namely one triaxial Search Coil Magnetometer (*SCM*) to detect fast temporal changes in the magnetic field, and two fluxgate magnetometers which provide a measurement of the absolute magnitude of the background magnetic environment. In the context of magnetic switchback detection, the majority of these background signals are derived from measurements made using the fluxgate magnetometer array. These magnetometers provide independent absolute field measurements that allow for the interpretation of solar phenomena such as shocks, reconnection, and turbulence (Bale et al. 2016).

Fluxgate magnetometers, unlike SCMs, are active measurement devices that work by utilising the physical properties associated with magnetic saturation and hysteresis. These sensors are comprised of three main components; namely a drive coil, magnetically permeable core, and sense coil. A typical fluxgate sensor operates by applying an oscillating waveform to the primary drive coil, this intern establishes a corresponding magnetic field which permeates the core material to induce an electric current in the secondary sense coil. When the magnetometer is exposed to an external magnetic field, the resulting interference created by the interaction between the internal and external magnetic field is reflected in the output of the sense coil (Verscharen et al. 2019). The signal recorded in the sensecoil can subsequently be used to infer the magnitude of the externally applied field. This feature makes these fluxgate sensors useful for detecting events that are characterised by changes in the absolute signature of the external magnetic field (such as switchbacks). From equation (2), the working principle of a basic fluxgate magnetometer can be further quantified:

$$\mathcal{E}_s(t) = \mathcal{E}_0 \sum_{k=1}^{\infty} (1 - e^{-2\pi\beta k i}) \frac{\sin(\pi\alpha k)}{\pi k} \cos\left(\frac{2\pi k t}{\Pi}\right) \quad (2)$$

Where:

$\mathcal{E}_s$  = Induced EMF in sense coil [V]

$\mathcal{E}_0$  = EMF of drive coil  $[-\frac{2N_s A B_s}{c\alpha\Pi}]$

$\Pi$  = Pulse period [s]

$\alpha\Pi$  = Positive pulse duration  $[\frac{H_c}{4H_0}\Pi]$

$\beta\Pi$  = Negative pulse duration  $[\frac{1}{2}\left(1 - \frac{\Delta H}{H_0}\right)\Pi]$

From the formula above it is evident that the signal recorded in the sense coil is a superposition of the magnetic field produced by the drive coil and external field environment. Using the known properties of the primary, the field characteristics of the external field can be determined. PSP's fluxgates are positioned with opposing orientation, in order to provide the spacecraft with a full spatial resolution (Bale et al. 2016).

Another instrument that is often utilised conjunction to the magnetometers is the Solar Probe Cup (*SPC*). The SPC is a part of the Solar Wind Electrons Alphas and Protons (*SWEAP*) instrument suite and is responsible for the measurement of the solar wind. This SPC is a form of integrating electrostatic analyser more generally known as a Faraday Cup (Case et al. 2020). This cup derives its measurements by sweeping/integrating through a voltage range in the mesh grid separating the collector plate from the instrument aperture (Verscharen et al. 2019). Charged particles with energies lower than that of the dictated voltage will experience an overwhelming electrostatic force of repulsion that prevents the particle from being recorded at the end collector (thereby only resolving species with energies greater than the cut-off threshold). This cut-off threshold can be modifying by changing the offset and peak-to-peak voltage across the gridded mesh.

Using this working principle of electrostatic selection, the SPC is capable of measuring the specific flow rate of both electrons and ions in the solar wind. Equation (3) shows how the relational current measurement is derived between the recorded current on the collector plate and the applied offset voltage:

$$\Delta I(\mathcal{E}_o, \Delta\mathcal{E}) = I\left(\mathcal{E}_o - \frac{\Delta\mathcal{E}}{2}\right) - I\left(\mathcal{E}_o + \frac{\Delta\mathcal{E}}{2}\right) \quad (3)$$

Where:

$\Delta I$  = Current for particles with velocity greater than cut-off threshold [A]

$\mathcal{E}_o$  = Offset voltage [V]

$\Delta\mathcal{E}_s$  = Peak-to-peak amplitude [V]

As mentioned, the solar wind is tightly coupled to the interplanetary magnetic field and thus the SPC can be used alongside the magnetometers to infer/undercover relational characteristics between the two entities.

By combining the properties and measurements of these magnetometers and electrostatic analyser, PSP is capable of resolving an in-situ time series record of the surrounding magnetic field and plasma environment in all three dimensions around the spacecraft (radial, tangential, normal) (Froment et al. 2021). Due to the restricted trajectory of PSP, the measured magnetic field components are fixed relative to the spacecraft. This is illustrated in figure 2:

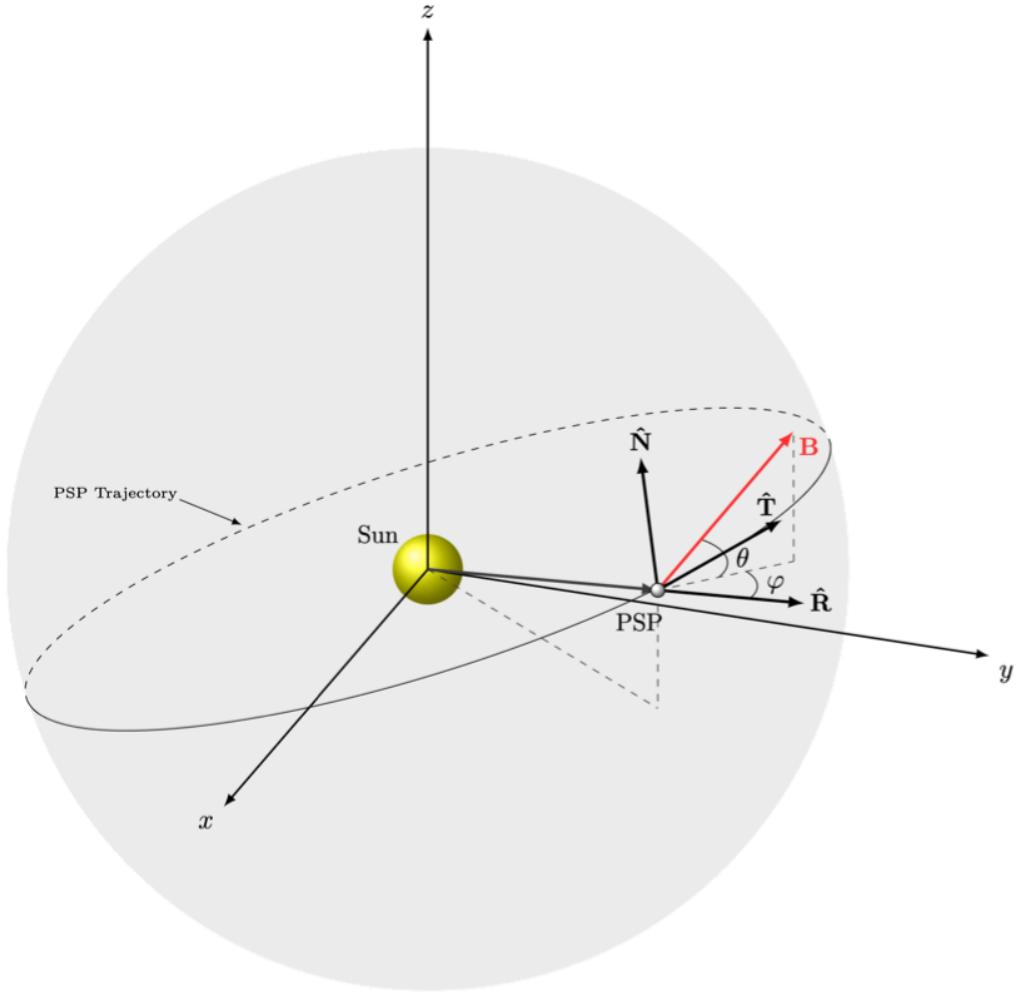


Figure 2: Diagram outlining spacecraft relative reference axis and associated vector components. This Radial-Tangential-Normal (*RTN*) coordinate reference frame is widely used for interplanetary missions as a means of expressing the measured local velocity and magnetic field vector components (Harvard 2020). This system is particularly useful when measuring on the partially inclined orbit of  $3.4^\circ$  as the recorded vector components do not correspond to that of the solar orientated reference frame (Guo 2008).

Based on the vector components highlighted in figure 2, the various magnetic field components can be resolved relative to the Sun-spacecraft line<sup>5</sup>. Knowing that magnetic switchbacks occur most prominently in the radial direction, this data can be used to accurately visualise and identify the field characteristics, including structures such as switchbacks (Froment et al. 2021).

---

<sup>5</sup>This feature is to be considered when evaluating the vectorised component of the magnetic field data.

## 2.4 Machine Learning in The Space Environment

By nature, heliophysics is a very data driven field. Due to the need to compute and understand the complex environment, many applications require the use of statistical algorithms that aid this high computational demand. Machine learning is a powerful tool when modelling and classifying data as it can generalise and uncover underlying trends in large datasets. As mentioned, this characteristic is especially notable in the heliophysical environment (Amaya et al. 2020a). Hence, the application of such statistical algorithms is becoming increasingly popular in modern heliophysics. This interest arises due to ML's ability to characterise non-algorithmic, and complex high dimensionality data problems (Sadykov 2020).

More particularly unsupervised learning<sup>6</sup> has become widely adopted in the field due to its inherent functional utility and implementation simplicity (Roberts et al. 2020). These machine learning practices can often be more desirable/beneficial as they are insensitive to time dependent data correlations, they do not require time consuming data labelling practices, and are applicable to indistinct data trends (Roberts et al. 2020). Thus based on these reviewed factors, unsupervised machine learning is the most ideal approach to be adopted for the context of classifying and interpreting magnetic switchbacks (Amaya et al. 2020b).

Due to the highly flexible and efficient modelling ability of unsupervised learning, the inherent issues associated with studying switchbacks can potentially be overcome. By understanding the advantages and limitations of these algorithms, various aspects and qualities (regarding the underlying scientific properties of the data) can be deduced (Géron 2019). This includes developing insight into the internal relationships of the feature data to better identify the relational attributes contributing to the occurrence of switchbacks and the source data, as well as providing a means of classifying these stochastic events.

In order to evaluate the performance of an unsupervised learning algorithm, many evaluation techniques can be used. In the context of using unsupervised learning to classify an event, the prediction performance can be evaluated against a known/proven basis (*see Materials and Methods*).

In a binary classification scenario<sup>7</sup> the most common and useful metrics of evaluation include the model accuracy, precision, recall, and F1-Score (Virpioja et al. 2011). These complementary aspects of classification evaluation overcome biasing factors such as class imbalance and provide a more encompassing statistic by which a classification prediction can be quantified. By comparing different aspects of the prediction results with the ideal/benchmark data, these listed performance metrics can be derived. From the sub-equation (4) the respective binary classification evaluation metrics can be quantified as follows:

---

<sup>6</sup>Unsupervised machine learning is a type of statistical learning algorithm that extracts underlying patterns in high dimensionality unlabelled data groupings (Education 2020).

<sup>7</sup>The classification of magnetic switchbacks is inherently a problem in the order of binary classification, as the model is expected to categorise the occurrence or non-occurrence of a switchback event.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4a)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4b)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4c)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4d)$$

Where:

$TP$  = True Positive Predictions

$TN$  = True Negative Predictions

$FP$  = False Positive Predictions

$FN$  = False Negative Predictions

It is evident from above that each metric evaluates a differing quality of the prediction comparison. This evaluation diversity allows each model to be assessed on a variety of performance monitoring factors, in order to identify the respective strengths and weaknesses of each algorithm.

## 2.5 Machine Learning Methods

As discussed, unsupervised learning is the ideal class for our purpose of ML evaluation due to its ability to learn from underlying trends within the model features (Roberts et al. 2020). Of the various unsupervised learning techniques, one of the most adopted methods is the application of clustering algorithms. These algorithms cluster data groups based on probabilistic parameters in the underlying feature set (Andreassen et al. 2019). This form of learning can be used to identify intrinsic characteristics associated with the inputs to the model, making them very desirable tools for scientific analysis (such as data associated with magnetic switchbacks). By forming these clusters based on the respective likelihood, these models are able to formulate internal data relationships in higher order data groupings. Some of the limitations that are to be considered/accounted for when using unsupervised learning techniques include the inherent computational complexity and exhaustive training times, high risk of model divergence leading to inaccuracies, and lack of output validation and model transparency (Education 2020).

In the context of magnetic switchbacks, various unsupervised clustering models can be fitted to in-situ spacecraft data. These models can then form cluster predictions, on the unlabelled datasets, from which the groupings can be analysed. For the analysis, we chose to evaluate the performance of four different clustering techniques. Namely:

- K-means clustering
- Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*)
- Agglomerative Hierarchical clustering
- Bayesian Gaussian Mixture models

Each selected model has distinct properties that enable clusters to be formed. Due to this variation, the relative performance is expected to be highly inconsistent (Bousquet et al. 2011). As a result of these differing objective approaches, a cross-evaluation between these four models will reveal intrinsic properties associated with magnetic switchbacks (i.e. more idyllic prediction results can infer relational features of switchbacks depending on the fundamental functionality from which the models are based).

## 3 Materials and Methods

### 3.1 Data Acquisition

In order to perform the machine learning analysis, we first obtain the magnetic field data pertaining to the respective RTN vector components. This level two data is acquired from open-source repositories<sup>8</sup>. The raw data contains the relevant switchback structures, as well as the respective source data from which the model features are to be constructed. From initial observations of the raw data, we see that the files are stored in batches of one day recordings, each datum within the file is sampled with a one-minute cadence<sup>9</sup>. To develop the data map of the magnetic field distribution across the entire encounter of PSP, we sequentially load the magnetic field data for each day. We perform this evaluation across three separate encounters (namely the second, third, and fourth encounter of the spacecraft), the dates of the corresponding encounter periods are displayed in table 1:

Table 1: Characteristics of all evaluated Parker Solar Probe orbital encounters

Orbital Epoch	Encounter Period	Heliocentric Distance (au)
Second encounter	13/03/2019 – 29/04/2019	0.17 – 0.60
Third encounter	08/08/2019 – 26/09/2019	0.17 – 0.61
Fourth encounter	08/01/2020 – 27/02/2020	0.13 – 0.67

Noting that magnetic switchbacks occur predominantly at heliocentric distances less than  $60 R_{\odot} \approx 0.28 \text{ au}$ , we aim to conduct our evaluation around this distance threshold. This will ensure that all magnetic switchbacks recorded over the given encounter will provide us with sufficiently diverse statistical data.

Once the sequential data had been acquired, we represent the raw data in a visual manner in order to evaluate and assess the quality of the recorded results. This is achieved by visualising the time series variations in magnetic field density at differing heliocentric distances for all RTN coordinates. Figure 3 shows the relevant magnetic field measurements taken by PSP over a 45-day transit period during the third orbital epoch<sup>10</sup>:

---

<sup>8</sup>We acquired the magnetic field data from the following source:  
[http://research.ssl.berkeley.edu/data/psp/data/sci/fields/l2/mag\\_RTN\\_1min/](http://research.ssl.berkeley.edu/data/psp/data/sci/fields/l2/mag	RTN_1min/)

<sup>9</sup>Due to this observation we note that all recorded switchbacks with cadences greater than this one-minute threshold are encompassed by the data. Likewise, all lesser switchback events are disregarded.

<sup>10</sup>For the purposes of brevity, we purposefully chose not to include the diagrams and illustrations pertaining to the other evaluated encounters. Henceforth, all displays will refer to the third orbital encounter although these observations are replicable for all other encounters (*see Appendix I*).

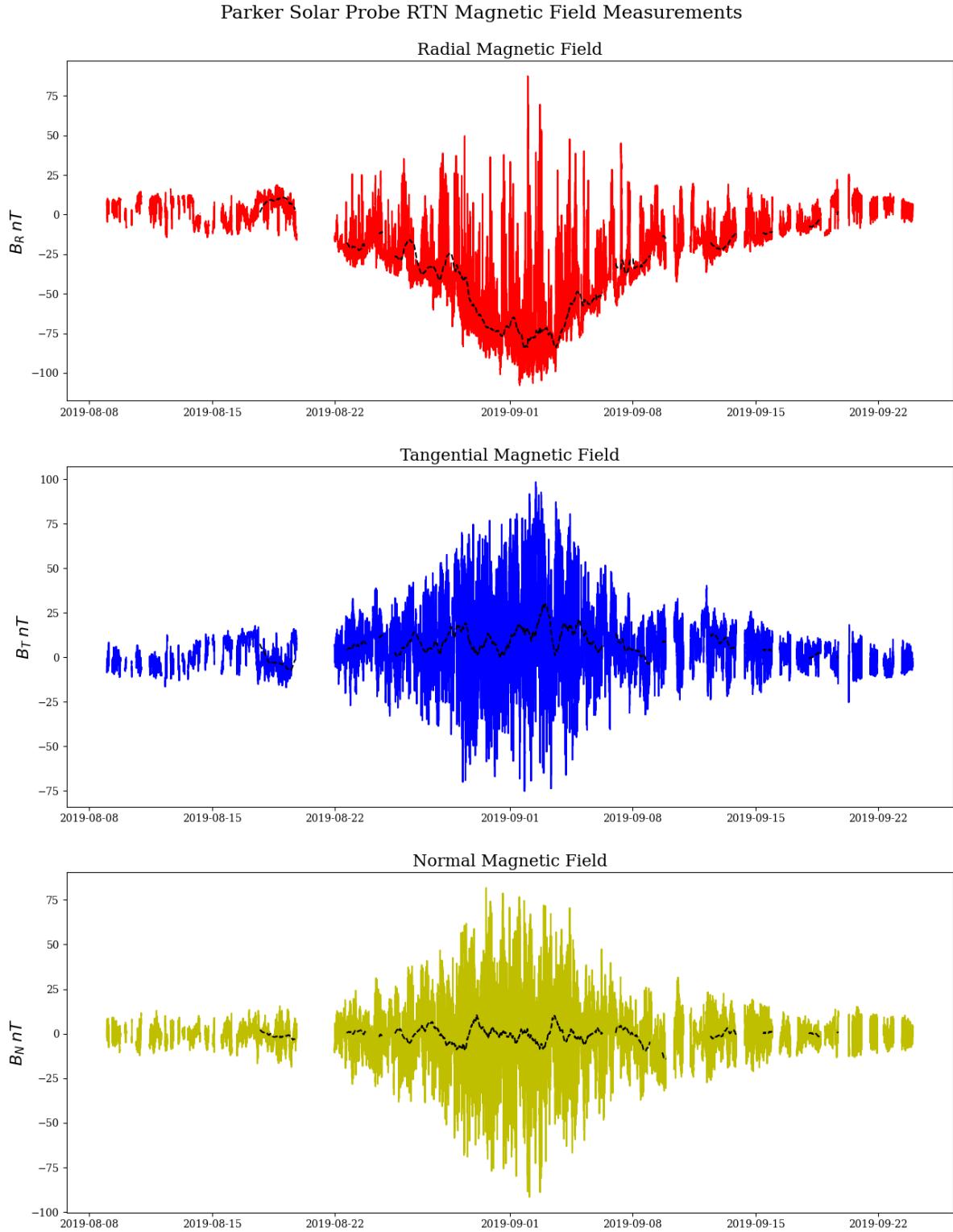


Figure 3: Raw data plots containing magnetic field vector components acquired from Parker Solar Probe magnetometers. This plot illustrates the comparison of observed radial (red), tangential (blue), and normal (yellow) magnetic field measurements, relative to the mean expected background field profile (black). This one-minute cadence, time series data is displayed for the third solar encounter between August and September 2019.

Following the observations made by Kasper et al. (2019), it is evident that the occurrence of magnetic switchbacks is greatest at close proximities to the Sun. We also note that these field fluctuations occur most prominently in the radial direction and hence are most apparent in the radial subplot depicted in figure 3. These observations are consistent with what we see in literature and validates the characteristics of our data.

We also note, from figure 3 that the magnetic field densities are intrinsically linked to the absolute distance of PSP from the Sun. This relationship follows a proportionality trend that can be described by  $|B_R| \propto 1/r^2$ . As a result, this distance parameter must therefore also be considered when developing the classification algorithms. This is because the information pertaining to the spacecraft distance provides an additional dimensionality from which the model can be referenced<sup>11</sup>. In correspondence with this observation, we source and add the PSP heliocentric proximity data to our magnetic field data. The proximity data is not derived from any PSP instrumentation, and therefore we must acquire the data independently (*see Appendix I*).

From these raw magnetic field datasets, we decide to allocate the radial magnetic field component as the primary basis of comparison when evaluating the models. Although the radial component of the magnetic field contains the most switchbacks, we will not discard the other vector components. This is because these fields possess underlying data characteristics that relate to the occurrence of switchbacks, and thus shall be used as features which are to be provided to the unsupervised models.

## 3.2 Feature Engineering

Before the features can be constructed, we must first remove the discontinuities and inconsistencies within the datasets. This is achieved by applying the appropriate methods of imputation and mapping, as well as removing any null and anomalistic instances. Concomitantly, we interpolate and map the spacecraft's heliocentric distance data to the magnetic field data depending on the varying epoch scales. We act to apply these cleaning techniques in order to create an appropriate feature space from which the models can be developed.

To ensure the simplicity and reproducibility of our research, we aim to minimise and simplify the features which are passed to the models. Thus, we select the respective features based on their physical relation to the source data. Following this consideration, we conclude that raw parameters such as the absolute magnetic field vector angles (altitude and azimuth), percentage change of radial magnetic field relative to previous epoch, and the interpolated distance of PSP from the Sun would be used in the development of the unsupervised models. Table 2 depicts the relevant features used in the development of the interim models:

---

<sup>11</sup>Due to computational and resources limitations, this distance metric is only sampled with a cadence of one day. This sample rate downcasts the available distance information relative to the minutely sampled magnetic field data.

Table 2: Initial features and boundary metrics used in interim model evaluation

Feature Parameter	Value Bound	Metric
Altitudinal Magnetic Field Vector Angle ( $\theta$ )	$[-\pi, \pi]$	rad
Azimuthal Magnetic Field Vector Angle ( $\varphi$ )	$[-\pi, \pi]$	rad
Radial Magnetic Field Percentage Change	$(-\infty, \infty)$	%
PSP Radial Distance from Sun	[0.05, 1.01]	au

It is apparent from the parameters highlighted in table 2 that the selected features all share some underlying relationship to switchbacks. This fundamental characteristic is essential as these considerations will aid the performance of the ML models. Another point that we observe is that the variabilities in the upper and lower bounds of these features are starkly different<sup>12</sup>. This contrasting variability can potentially hinder the performance of the unsupervised models, which are sensitive to feature scaling, and thus we needed to make considerations regarding how this data is to be processed.

### 3.3 Feature Scaling

Due to the learning behaviours of different ML architectures pre-scaling is an essential step as it aids the training and fitting process of these algorithms (Géron 2019). Generally, the two most common data scaling functions include MinMax scaling (scaling variables between zero and one), and standard/Gaussian normalisation (Géron 2019). The application of each respective function is dependent on the characteristics of the data. Due to the linear nature of Min-Max scaling, this function can distort and skew the resolution for nonlinear distributions (particularly data that has an infinite domain). Noting that our data possesses a more statistical distribution, we chose to implement the standard scaler algorithm. This normalisation function can be expressed as:

$$\mathcal{N}(0, 1) = \frac{x - \mu}{\sigma} \quad (5)$$

Where:

$\mathcal{N}$  = The standard normal distribution

$x$  = Data instance

$\mu$  = Distribution Mean  $\left[ \frac{1}{N} \sum_{i=1}^N x_i \right]$

$\sigma$  = Distribution standard deviation  $\left[ \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \right]$

It is evident from this function that highly variable datasets can be bound within a zero centred distribution possessing a unitary standard deviation, allowing the models to make more generalised interpretations of the prediction results. Due to the clustering behaviours of the different unsupervised models and the high variability of the feature space, we modify the data depending on the input requirements of each model. For the Euclidian based models (namely: K-means, and Hierarchical clustering), this data is normalised to facilitate the formation of cluster boundaries<sup>13</sup>.

<sup>12</sup>Particularly the bounds of the magnetic field percentage change, as this infinite range is due to the occurrence of a division by zero. Although not an ideal characteristic, we left the data in this boundless format as it will be dealt with during the pre-processing stage.

<sup>13</sup>We do not apply this same normalisation technique on the density-based models, as this would

## 3.4 Statistical Modal Analysis

One issue that arises when using unsupervised ML is the lack of evaluation criteria. This is a direct result of the fact that there is no definitive method of identifying the occurrence of magnetic switchbacks, and thus no basis through which the clusters of the unsupervised ML predictions can be evaluated. To overcome this observation, we choose to use a proven statistical model in order to determine the occurrence of a switchback. Following the method outlined by Bale et al. (2019), we decide to utilise statistical modal analysis as a means of background field identification (although this is not the only method of evaluation). By representing statistical switchbacks in binary format, we can utilise classification metrics to make comparisons with the trained model predictions.

Following the procedure highlighted by Bale et al. (2019), these modes can be computed. We begin this process by sampling the full series of the radial magnetic field. This data is then partitioned into subsets of fixed and equal cadence<sup>14</sup>. These subsamples are then compiled into histograms with a bin density proportional to one-tenth of the subset size. From these histograms the modes can be extracted. By sequentially recording these modes, we are able to produce the plots over the specified encounter.

Succeeding the conduction of the modal analysis we are able to generate results for varying mode cadences. By modifying the histogram bin size in relation to the mode sample size, we can produce various mode statistics and compare the results. Figure 4 shows the overlayed results of the sampled modes with respect to the radial magnetic field vector component:

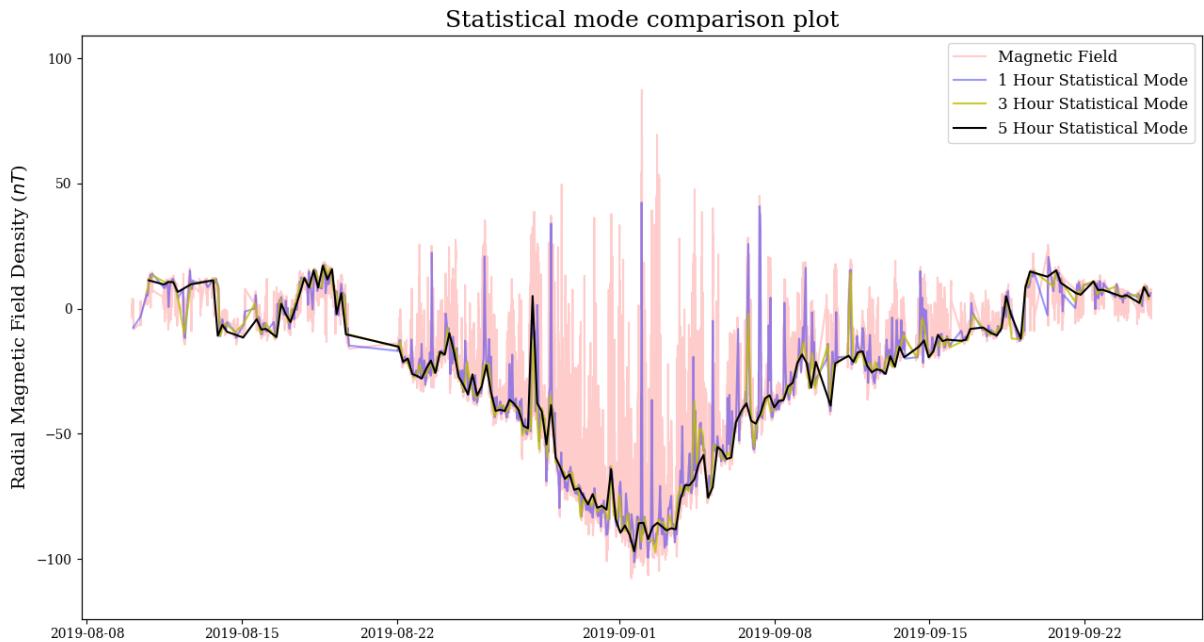


Figure 4: Diagram depicting results of statistical modal analysis on third encounter radial magnetic field data (*red*). This figure illustrates the recorded mode magnetic field signature for the one-hour (*blue*), three-hour (*yellow*) and five-hour mode cadences (*black*).

modify the relative densities of the data distribution and lead to a more restrictive model

<sup>14</sup>These cadences are variable so that we could conduct our analysis over a dynamic range.

From the plots depicted in figure 4, it is evident that by varying the mode cadence we are able to selectively isolate switchbacks of differing timescale. As the sampling period of the modes increases, less switchbacks are included, and subsequently more of the background signature is isolated. At a five-hour cadence, we are able to isolate the majority of the switchback structures (indicating that the dataset does not contain switchbacks with longer time periods). From this analysis we are now able to deduce the presence of switchbacks by recording all points where the sign of the field differ from that of the modal field.

## 3.5 Unsupervised Machine Learning Models

### 3.5.1 K-Means

K-Means is one of the most common clustering algorithms due to its implementational convenience and its relative efficiency (Krishna & Narasimha Murty 1999). This algorithm is a form of Expectation-Maximisation (*EM*) model, meaning that the algorithm derives its predictions by assigning an expectation value which it aims to optimise (Likas et al. 2003). In the context of K-Means modelling, this expectation value is defined by the number of desired clusters which is to be specified beforehand. These clusters are then optimised by minimising an inertial metric.

Clusters are formed by first randomly initialising  $K$  number of centroids ( $K$  corresponding to the specified number of classes), to which each instance is assigned depending on the relative proximity to a given centroid (Géron 2019). The locations of the respective instances are then calculated based on the mean distance of each instance in the corresponding class. This process is iterated until the relative inertia of each centroid converges. This centrical inertia can be defined as follows:

$$I = \arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2 \quad (6)$$

Where:

$I$  = Clustering inertia

$S$  = Cluster sum-of-squares  $[S_1, S_2, \dots, S_k]$

$k$  = Number of cluster subsets

$\mathbf{x}$  = Feature space  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$

$\mu_i$  = Mean of cluster sum-of-squares

From (6) it is evident that by optimising the total inertia of the system, the model acts to reduce the total variance of instances within a cluster. Once fully converged any instances which lie within the sphere of influence of a particular centroid is considered to fall under that respective cluster. The end result of this K-Means algorithm is illustrated in figure 5.

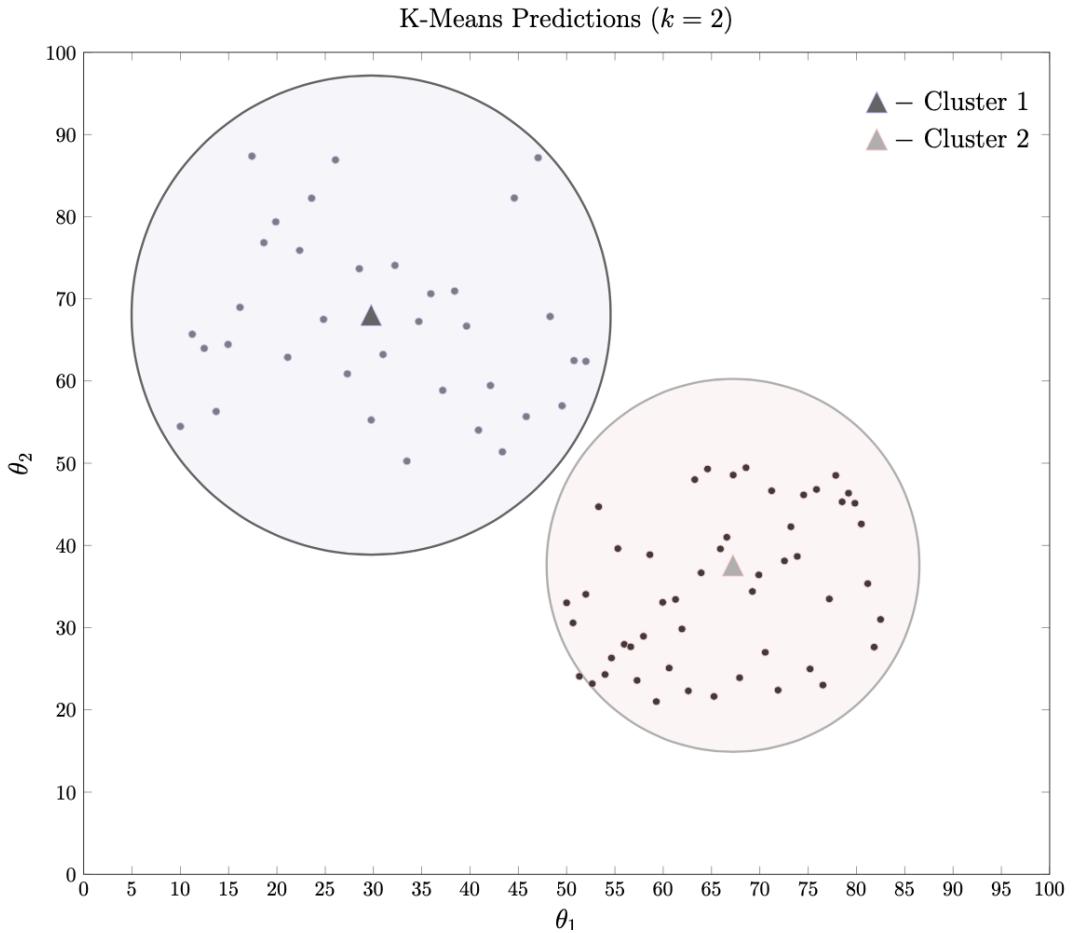


Figure 5: Diagram depicting the clustering behaviour of a two cluster K-Means architecture, given a two-dimensional feature space (where:  $\theta_1$ ,  $\theta_2$  represent hypothetical features). From this illustration it is evident that the hyper-spheres are ideal for clustering uniformly distributed datasets with minimal overlap. The importance of feature scaling is also evident from the figure, as non-uniformly scaled datasets do not form idealised cluster groupings.

In higher dimensions these K-Means clusters form hyper-spheres by which the features are clustered. However, due to this mechanism of cluster formation, K-Means algorithms assume the provided data is spherically clustered and that all features are scaled accordingly. Thus, these models typically tend to be highly sensitive to pre-scaling conditions and considerations must be put in place to ensure the model is optimised correctly. Along with these issues, because of the randomised nature of the initialisation process, K-Means algorithms can lack repeatability. As a result, we must perform several iterations in order to ensure the outputs are consistent.

In some cases, the feature sets from which the K-Means models are evaluated can often be too large. To overcome this a smaller subset of the overall data space can be sampled. This approach is called mini-batch K-Means, and is a useful adaptation when clustering datasets that may overwhelm computational resources or act to overfit the predictions of the unsupervised models (Géron 2019). We use the latter reasoning in our implementation as the large and highly variable dataset can cause the K-Means algorithm to improperly converge. We also need to consider the potential consequences that sampling may have, as this can also potentially lead to the formation of inadequate clusters (as too few data instances can overly restrict the model’s performance).

Along with down sampling 20 % of the data, we also ensure our K-Means implementation is appropriately optimised by scaling the feature data using the standard normalisation technique (outlined beforehand). We ensure that the model is initialised with two clusters as each grouping will either correspond to the occurrence of a switchback or nominal magnetic field behaviour respectively.

### 3.5.2 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*), is a non-parametric algorithm that groups clusters based on the relative densities of instances within the dataset (Ester et al. 1996). These densities are determined based on a simple algorithm that measures the continuity of regions of similar density (*see Appendix II*).

Clusters are formed by measuring the number of instances that fall within a specified distance denoted by a specified noise variable  $\varepsilon$ . If a minimum threshold of instances are found to lie within these specified density regions, the central instance is regarded as a core instance (Géron 2019). Sequences of these core instances are considered part of the same clustering and all outliers are deemed to be anomalies. This cluster formation process (with varying noise variables) is evident in figure 6.

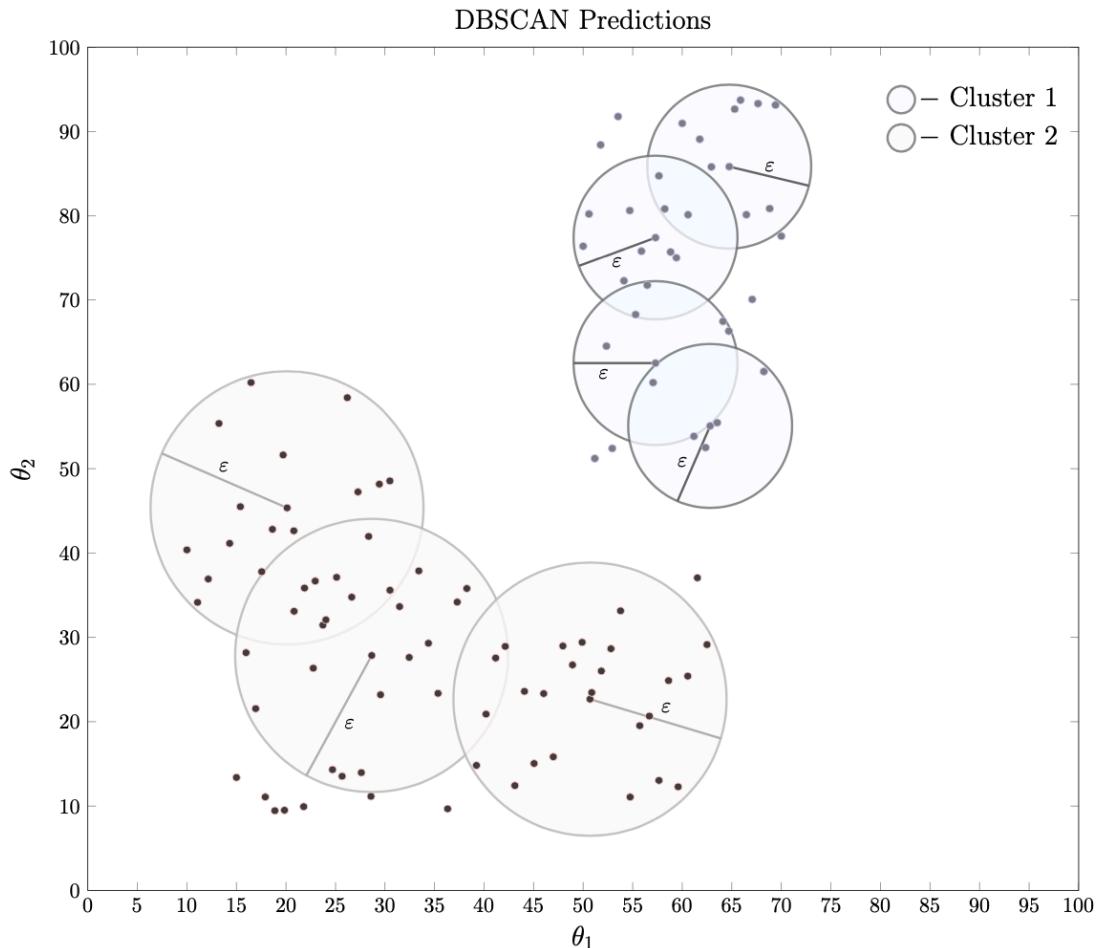


Figure 6: Diagram showing how density-based clusters are formed using the DBSCAN algorithm. By specifying the sampling noise ( $\varepsilon$ ) cluster groups of overlapping density can be formed on the provided feature space ( $\theta_1, \theta_2$ ). We can also note, from the plot, that outlier data points are neglected by the DBSCAN model.

It is evident from figure 6 that by lowering the  $\varepsilon$  condition, clusters of higher density can be isolated. Due to the non-parametric nature of this algorithm, when implementing this model on the magnetic field data we are required to iteratively find the ideal hyper-parameters necessary to develop the two cluster sets (*see Appendix II*). From the clustered results, we consider the anomalistic/outlier prediction set to correspond with the presence of magnetic switchbacks.

### 3.5.3 Agglomerative Hierarchical Clustering

Hierarchical Clustering is a simple unsupervised algorithm that derives its predictions by pairing instances in a manner that minimises the relative proximity of all points in the dataset (Johnson 1967). There are two conventional methods of hierarchical clustering approaches; bottom-up (Agglomerative) and top-down (Divisive). Both models make the associated predictions based on the hierarchical inheritance of properties within the feature space. These data properties typically pertain to the relative proximities of the associated instances, and are determined using various methods (Murtagh & Contreras 2012). However, in the context of our research we choose to evaluate our data with the Euclidean distance parameter. We choose this training metric as it is the most popular form of proximity evaluation and shows relatively good performance when compared to the alternative methods. This Euclidean distance is defined as follows:

$$\|a - b\| = \sqrt{\sum_i (a_i - b_i)^2} \quad (7)$$

Where:

$a, b$  = Respective positions of subset instances

By iteratively grouping each instance based on the minimum Euclidean distance, the respective cluster mapping can be produced. From this mapping we can set the desired ‘cut-off’ point based on the intended number of clusters. This division of clusters is visualised in the following dendrogram:

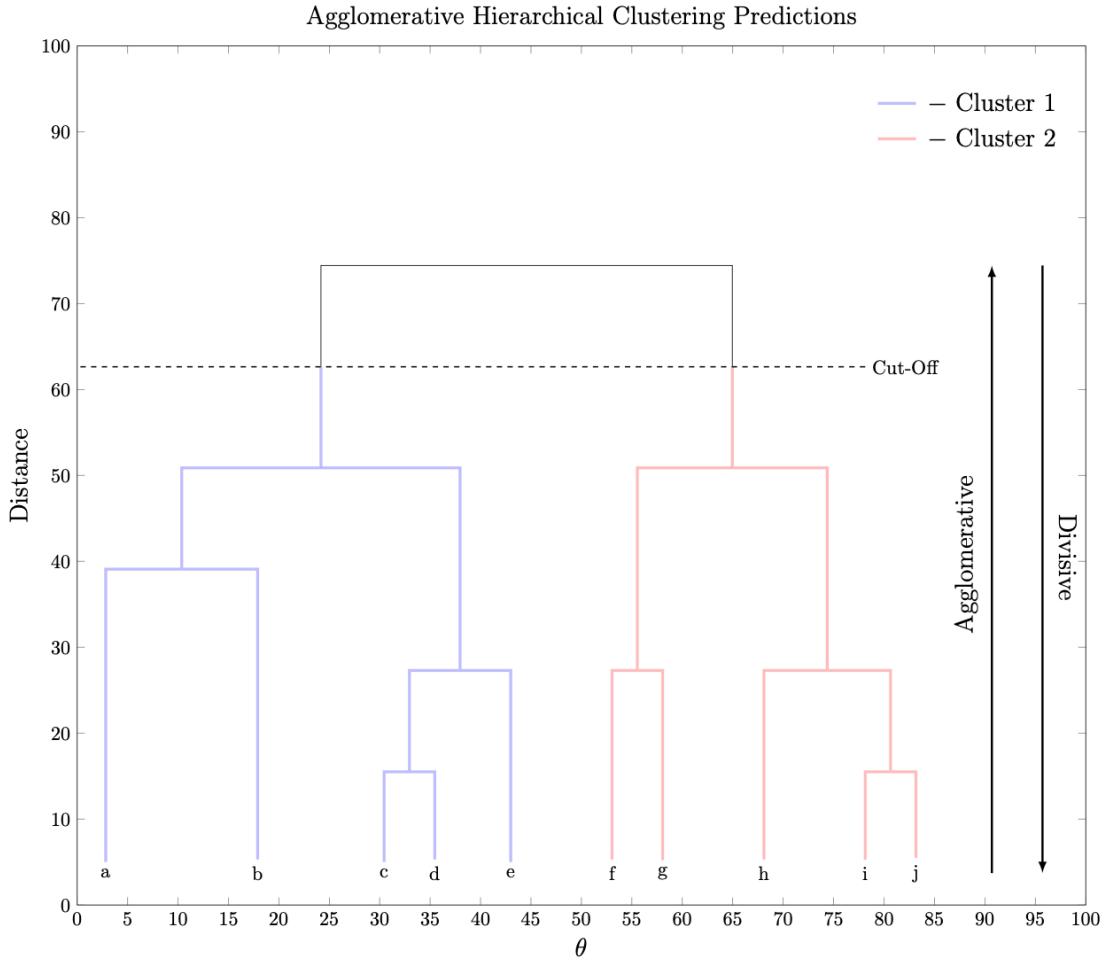


Figure 7: Dendrogram illustrating how classes are derived using both the agglomerative and divisive techniques. This figure shows how outlier data points may be difficult to characterise as the Euclidean model architecture will bias predictions based on these outliers. Thus, feature scaling is essential for the development of accurate Agglomerative models.

From figure 7, the working principal of both the Divisive and Agglomerative Hierarchical algorithms can be illustrated. For the purposes of our research, we choose to use the Agglomerative Hierarchical method as this algorithm is the most commonly adopted form of clustering and groups variables based on their relative proximity.

It is evident from (7) that the proximity of each instance is determined by the effective distance between each feature vector. This consequential factor is what makes Hierarchical clustering very sensitive to the relative scaling of the instances within the data. Thus, when implementing these models we have to critically consider the properties of the data being provided to the model.

When applying the Agglomerative clustering algorithm in the context of our project we ensure each factor is considered when implementing the model. Along with specifying the number of desired cluster groupings, we also scale and normalised the feature data using the standard normalisation technique. This normalisation ensured any outliers and data inconsistencies are eliminated, and that the hierarchical algorithm will not bias its clusters on outlying anomalies. Following the model initialisation, we separately fit each encounter dataset to a hierarchical model and record the associated predictions.

### 3.5.4 Bayesian Gaussian Mixtures

The Bayesian Gaussian Mixture Model is a subclass of unsupervised ML models that falls under the general Gaussian Mixture (*GM*) category. These GM models are expectation-maximisation models similar to that of the K-Means algorithm, with the addition of probabilistic hyper-parameters (McLachlan & Peel 2004). GM models function by assuming the provided feature set was generated by sampling instances from a mixture of unknown Gaussian distributions. As a result of this sampling assumption, these models attempt to identify the likelihood a respective instance belongs to a particular class/mixture component (Roberts et al. 1998). Similarly to that of K-Means, when developing a GM model it is necessary to indicate the number of mixture components before a data set is sampled. This can potentially limit the performance of the model if the number of desired classes is not known prior to implementation (Lee et al. 2003). Figure 8 illustrates how these Gaussian Mixture models cluster instances of a particular feature set:

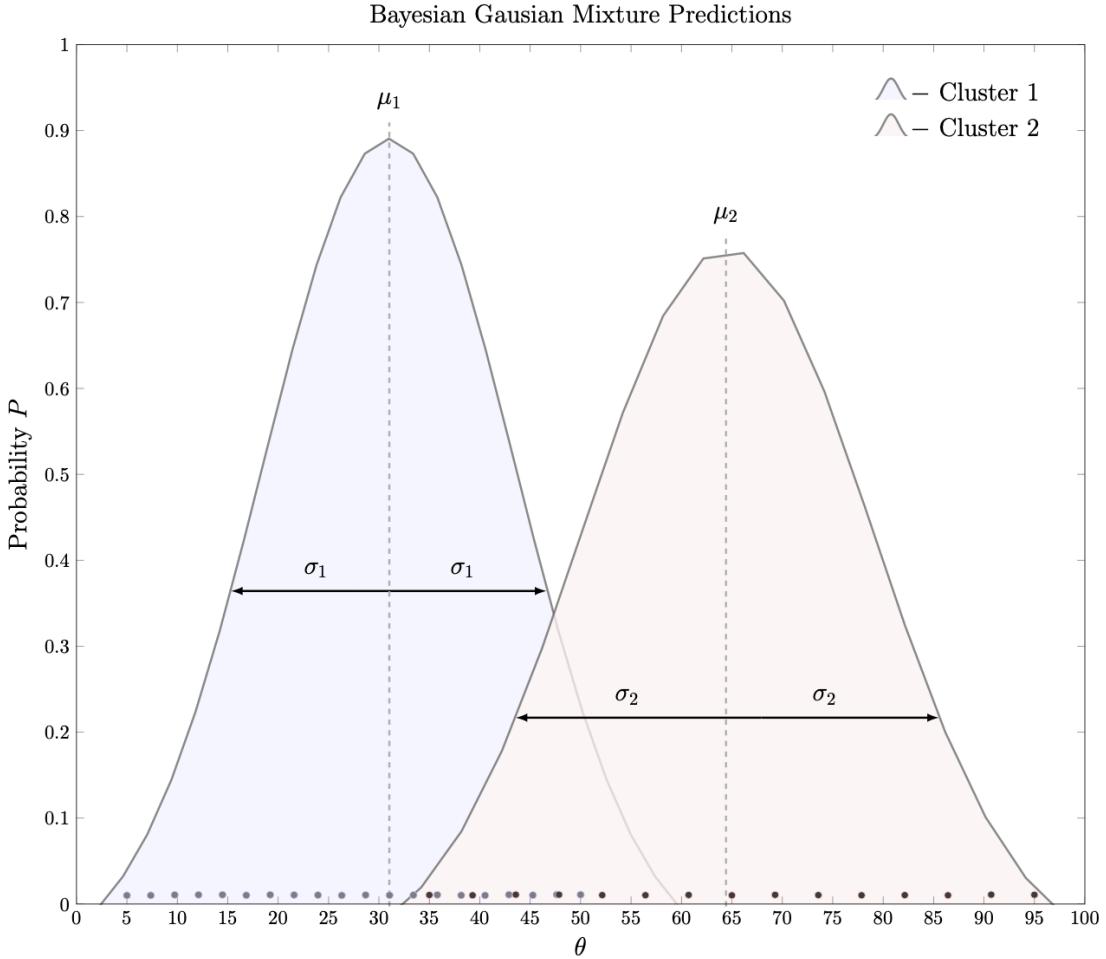


Figure 8: Diagram of probability distribution ( $P$ ) for one-dimensional feature space ( $\theta$ ) used by GM model. These one-dimensional plots extend to form hyper-ellipsoids in higher dimensionality feature spaces.

From figure 8 it is evident that these models are able to classify instances with overlapping decision boundaries. This is referred to as soft classification and is what differentiates these GM models from conventional EM algorithms such as K-Means (Pedregosa et al. 2011). In higher feature spaces, these models form hyper-ellipsoidal decision boundaries which possess additional statistical parameters which are to be determined. As a result of the additional parameters, GM models typically do not scale well with large datasets<sup>15</sup> (which is often a typical characteristic of solar-physical data).

Variational Bayesian estimation is a probabilistic optimisation algorithm that aims to minimise the respective cost probability between a prior value and posterior expected value (Constantinopoulos et al. 2006). This optimisation technique arises from Bayesian Inference and is used to estimate the analytical solution to the posterior probability of an unobserved variable (Pedregosa et al. 2011). By using Variational Bayesian estimation to infer the probability of a particular instance belonging to a given distribution, this statistical inference tool can be used in conjunction with the conventional GM principal framework to give rise to the Bayesian Gaussian Mixture model. This Bayesian component provides the GM model with an expectation value through which it can achieve maximisation. The probability by which a Bayesian Gaussian Mixture model assigns a particular data instance to a mixture component  $K$  can be defined as follows:

$$p(\theta|x) = \sum_{i=1}^K \phi_i \mathcal{N}(\mu_i, \Sigma_i) \quad (8)$$

Where:

$\phi_i$  = Class weights of  $i^{th}$  distribution

$\mathcal{N}$  = Normal distribution function

$\mu_i$  = Mean of  $i^{th}$  distribution

$\Sigma_i$  = Covariance matrix of  $i^{th}$  distribution

These Bayesian Gaussian Mixture Models are well suited for probabilistic applications where the data series may be highly randomised and contain statistical outliers. The attributes associated with GM models enable them to be integrated with real world problems where the datasets can often be defined by probabilistic decision boundaries. Due to these features the GM model can act as a generative model and is capable of being used to make Out of Sample (*OOS*) predictions<sup>16</sup>, as well as generate new instances based on random probabilistic sampling.

---

<sup>15</sup>Gaussian Models scale depending on the covariance type. In general GM models fit a function described by  $O(Kn^2)$  (with  $K$  indicating the number of clusters and  $n$  indicating the relative dataset size).

<sup>16</sup>This characteristic is similar to supervised machine learning models and can be very useful in the context of switchback analysis when predicting future sampled events.

In order to implement this model in our research we first initialise the algorithm to identify the two statistical distributions associated with the occurrence of a switchback and non-switchback. To overcome the data scaling issue (outlined previously), we randomly sample a 20% subset of the instance data. This sampling technique allows the models to be trained on statistically similar and more algorithmically efficient data, without significantly compromising the integrity of the final model<sup>17</sup>. However, because this subset data is not fully representative of the original dataset, we need to make considerations regarding the performance reliability of the trained model. Following this data manipulation stage, we implement the GM algorithmic sequence (*see Appendix II*) on the provided data, and save the resulting model parameters.

---

<sup>17</sup>This random sampling technique only works for significantly large datasets (typically: > 50 000 instances), however seems to break down when smaller subsets are analysed. This is due to an insufficient amount of data being provided to the model.

# 4 Results and Discussion

## 4.1 Prediction Evaluation

Following the methodological procedure outlined previously (*see Materials and Methods*), we are able to appropriately train all four unsupervised models, as well as develop the statistical mode data for the purposes of evaluation. The predictions of the models are based on the provided feature set with each model architecture producing unique interpretations of the results. We conduct the model assessment over three separate encounters (as mentioned previously), with each encounter's cluster being recorded in order to perform the separate assessments independently. Each encounter is evaluated by following the same assessment protocol<sup>18</sup>. Figure 9 displays the different switchback classification labels (indicated by the colour bars) generated by the four interim models overlayed with the coronal magnetic field distribution for the third encounter:

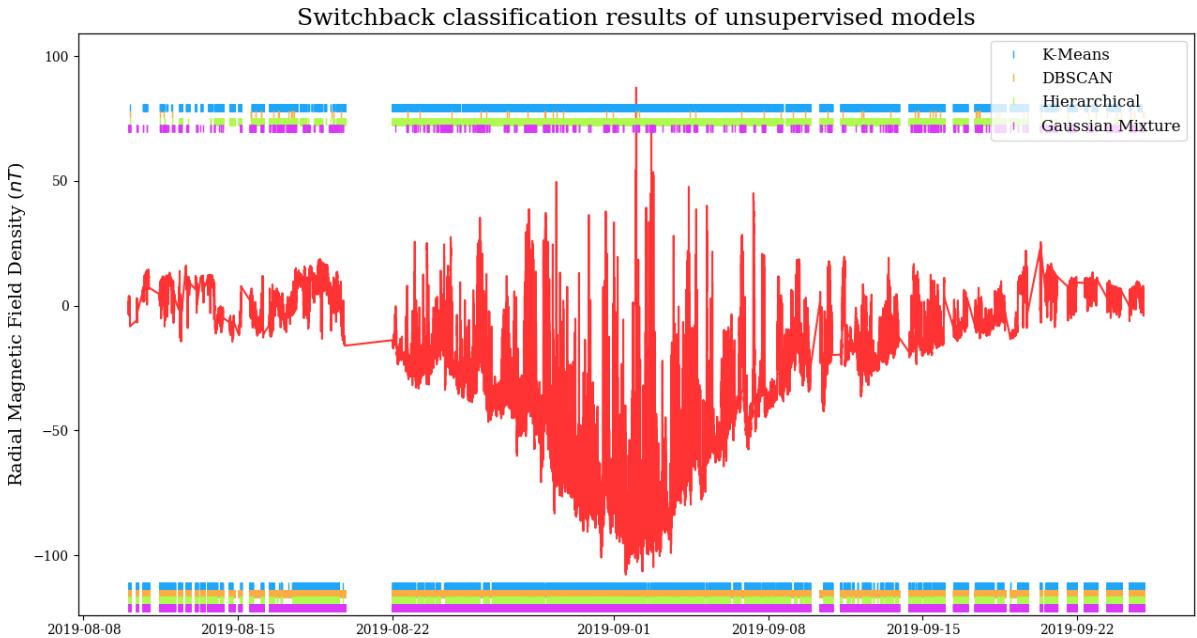


Figure 9: Overlay plot highlighting the regions where each unsupervised model predicted either switchback or nominal field activity. The background plot (*red*) indicates the radial magnetic field component for the third encounter. The respective marks at the bottom of the overlay indicate the zero predictions (i.e. no switchback activity). The markings on the top of the graph indicate the predicted switchback signatures.

<sup>18</sup>As mentioned previously, for the purposes of conciseness, we choose to only display the results pertaining to the third encounter. Although these results are shown to possess similar data characteristics (*see Appendix II*)

Upon observation of these visualised predictions in figure 9, we note several features relating to the data. Namely, the K-means, and Hierarchical model do not produce very well generalised predictions. We also note that these models showed their best performance across the beginning and ends of the orbital encounter periods (beyond the  $60 R_{\odot}$  range of switchback formation). This behaviour seems to indicate that these models break down when the density and frequency of the switchbacks increases. The observations regarding these two models are consistent across all evaluated encounters and indicate that these models are incapable of deciphering switchbacks from the provided dataset.

Conversely to these aforementioned observations, the density based and probabilistic architectures of the DBSCAN and GM algorithms show very promising performance. More particularly, we can see that the DBSCAN model produce relatively conservative prediction results and seemed to identify the major and most significant/prominent magnetic field reversals, whereas the Bayesian model is capable of flagging a wider variety of fluctuations resembling that of a switchback. These very generalised clusters seem to detect more fluctuations and indicate that the GM models are more sensitive to variations in the magnetic field and feature data. We also note upon review of the GM algorithm, that the model possesses a more evenly balanced class weighting. The remarks we have made regarding the illustrated performance of the DBSCAN and GM model are consistent across all evaluated encounters. This shows some indication that these models are capable of very accurately predicting the occurrence of magnetic switchbacks, as well as being a reliable source of switchback classification for new encounters.

## 4.2 Cross-Model Comparison

In order to thoroughly evaluate and study these ML algorithms, we must apply independent and alternative methods of examination to the prediction series of the trained models. By incorporating the proven method of statistical modal analysis for switchback identification (*see Materials and Methods*), we quantify the relationships between the various model architectures. This form of analysis helps us develop an appropriate understanding of the underlying operation for each mode, as well as validate or contradict what observations we have made thus-far.

Using the binary targets produced during the statistical modal analysis, as well as the classification evaluation metrics outlined earlier (*see Literature Review*), we examine and compare the classification performance of each algorithm. This process is conducted by treating the benchmark mode dataset as a basis of truth through which we can comparatively assess the classification algorithms. We evaluate each model's classification potential by recording the respective number of true and false predictions for each of the four algorithms made with respect to this basis model.

By sampling the predictions across differing mode cadences and orbital encounters, we generate a sufficient dataset relating to the results of each model's classification performance. These statistics are then averaged and summarised in order to encompass the overall functionality of the trained ML models. Table 3 depicts these averaged cross-model evaluation results for each algorithm with respect to the different statistical mode cadences (averaged over encounters) and encounters (averaged over mode cadences):

Table 3: Cross-model classification statistics for differing mode cadences and encounters

Metric	Algorithm	Mode Cadence Avg. (%)			Encounter Avg. (%)		
		1 Hour	3 Hour	5 Hour	Enc 2	Enc 3	Enc 4
Accuracy	K-Means	28.1	28.7	28.5	20.9	40.6	23.8
	DBSCAN	93.4	92.4	92.1	94.6	92.4	90.9
	Hierarchical	71.6	71.5	71.3	80.5	43.0	90.9
	Gaussian Mixture	91.7	90.8	90.5	91.3	91.6	90.2
Precision	K-Means	6.52	7.68	7.73	5.47	8.27	8.19
	DBSCAN	36.5	36.5	35.4	30.5	37.1	40.9
	Hierarchical	14.9	16.2	16.2	5.02	8.84	33.3
	Gaussian Mixture	35.3	35.6	36.3	29.0	39.0	39.2
Recall	K-Means	75.1	76.2	74.2	84.8	68.2	72.4
	DBSCAN	0.84	0.74	0.68	0.69	0.83	0.74
	Hierarchical	26.0	30.8	28.5	14.8	70.5	0.02
	Gaussian Mixture	28.8	26.1	25.5	43.5	21.8	15.1
F1-Score	K-Means	12.0	13.9	13.9	10.3	14.7	14.7
	DBSCAN	1.65	1.44	1.34	1.36	1.62	1.46
	Hierarchical	6.42	8.42	8.39	7.49	15.7	0.04
	Gaussian Mixture	29.5	27.6	27.4	34.8	27.9	21.8

From this table we discern the various levels of stability and reliability between each model. These averaged results provide us with an intuition into the relative deviation and consistency of each algorithm, as well as to how sensitive or insensitive the models are with respect to changes in the evaluation benchmark. These metrics also provide us with an understanding into the areas of specific performance and each model's area of strength.

The specific details represented in table 3 can be reflected and visualised in the cross-model comparison results (for the third orbital encounter) depicted in figure 10:

#### 4.2 CROSS-MODEL COMPARISON

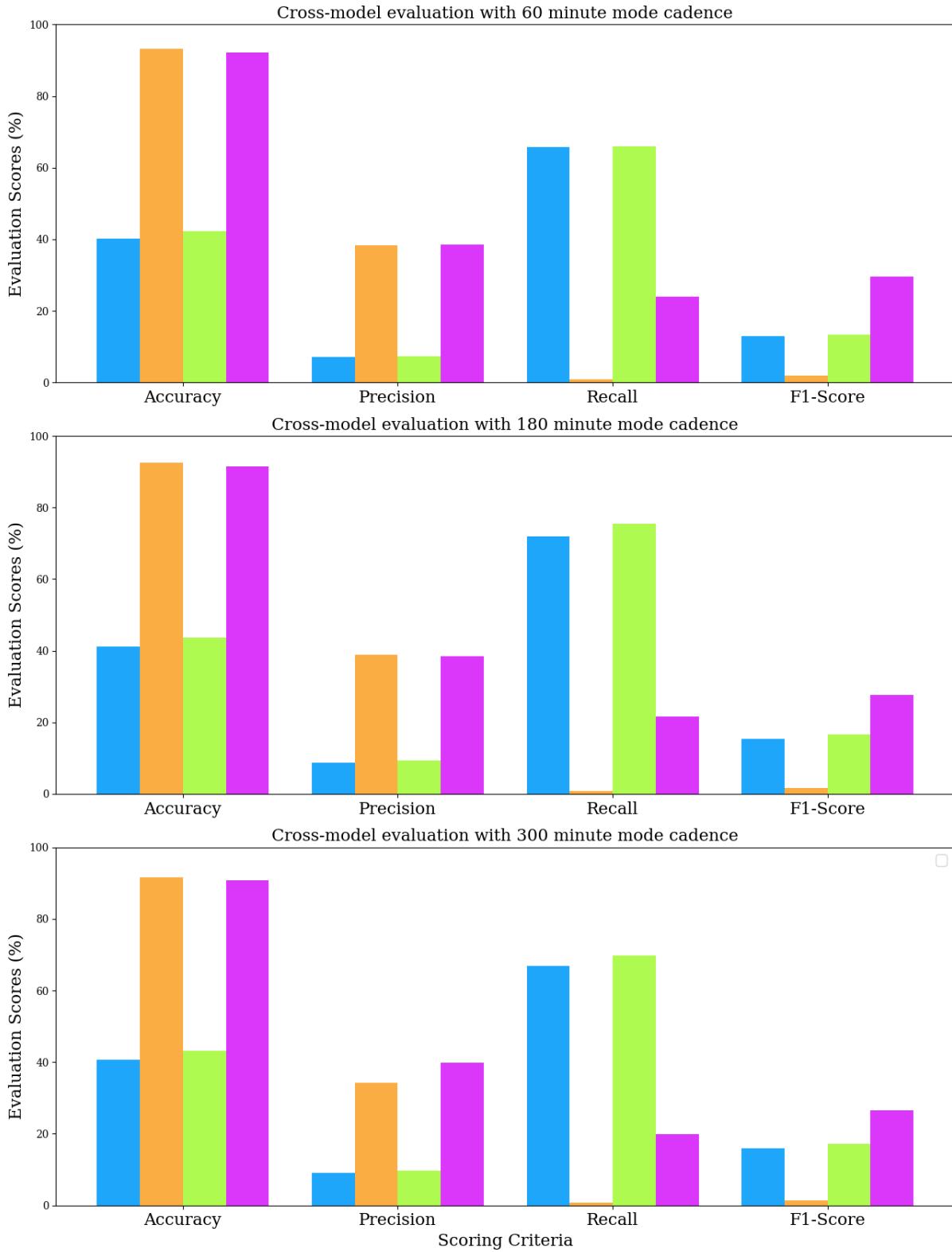


Figure 10: Diagram of the specific performance of each unsupervised model with respect to various statistical mode cadences. This diagram illustrates the performance of the K-Means (blue), DBSCAN (orange), Hierarchical (green), and Gaussian Mixture (purple) in each respective category.

Upon initial observation of table 3 and figure 10, we note that the K-means model not only produced highly variable results, but also shows a generally poor performance (with the exception of recall which is an irrelevant metric when all other aspects of accuracy and precision are low). These unreliable predictions are a direct result of the clustering nature of these algorithms. Due to the spherical assumption which K-means makes regarding the feature set, it is more appropriate to initially scale the feature data using Min-Max normalisation. However, for linearity and data preservation reasons outlined formerly (*see Materials and Methods*), this type of scaling will have an adverse impact on the quality of the dataset and thus impede the model's overall ability to develop clusters. Thus, from this observation, we can reason that the K-means model is not suited to forming accurate and precise clusters from this type of statistical data.

From these results we also identify similar attributes in the Hierarchical model as with the K-means algorithm. We note that this agglomerative model is highly sensitive to the provided feature space (regardless of any pre-scaling and normalisation) and is most likely contributing to the high variability which we are seeing. These undesirable characterise and unreliable performance under different encounter conditions are most likely a result of outlier biasing. Due to the greedy nature of Hierarchical clustering algorithms, most agglomerative models fail to accurately represent statistical distributions with outlier data points (as with what we are observing in the context of switchback classification).

Following the analysis of the DBSCAN algorithm it is evident that the low recall (observed in table 3 and figure 10) is indicative of the conservative nature of these models. Although this low recall consequently hinders the model's F1-score, this algorithm constantly demonstrates the highest accuracy and precision. This constancy and impressive performance makes DBSCAN highly suited for classifying magnetic switchbacks.

Correspondingly, this same level of performance is identified in the GM model. This Bayesian based algorithm produced the best overall performance and most generalised prediction results. The GM model shows accuracy and precision scores with only a few percentage deviation relative to DBSCAN. This model also seems to perform optimally when compared to data sampled with a high mode cadence. The result of this generalisability and high prediction accuracy are reflected in the GM model's relatively high F1-score and accuracy scores across all evaluated encounters and cadences.

In general, the models showed higher levels of responsiveness and variation when comparing different orbital encounters, as opposed to different cadence samples. This is expected as the feature sensitivity of each model is a larger contributing factor to the outcome of the model predictions as opposed to the varying cadence samples which was sourced from the same features. The level of relative variation among each observation is indicative of the reliability of each model, particularly over different encounter periods. From analysis of the cross-model comparison we note that the relative performance of each model with respect to the statistical mode data seems to support what is observed in the initial model comparison (relative to the raw radial field data).

## 4.3 Orbital Analysis

Although quantifying the model performance is a valid and appropriate method of identifying the best overall models, we aim to further understand how these models compare relative to PSP's orbital data. By examining the prediction results with respect to the scientific data we build an intuition into the relative performance of the best models, as well as to how and where the poor performing models diverge from the reality.

#### 4.3 ORBITAL ANALYSIS

We choose to perform this scientific analysis by using a novel method of evaluation. This is achieved by averaging and overlaying the prediction results<sup>19</sup> of each model with respect to the PSP orbital trajectory. By conducting the analysis in this manner, we are not only able to understand how these models perform on orbital data, but also study the regional and idiosyncratic features of each ML tool. This overlay plot visualises the relative regions where the algorithms detected the presence of switchback activity, as well as allow us to identify where each model either complemented or contradicted the consensus of the majority. The plots depicted in figure 11 illustrates the effectiveness of these new proposed methods of analyses for all three evaluated encounters:

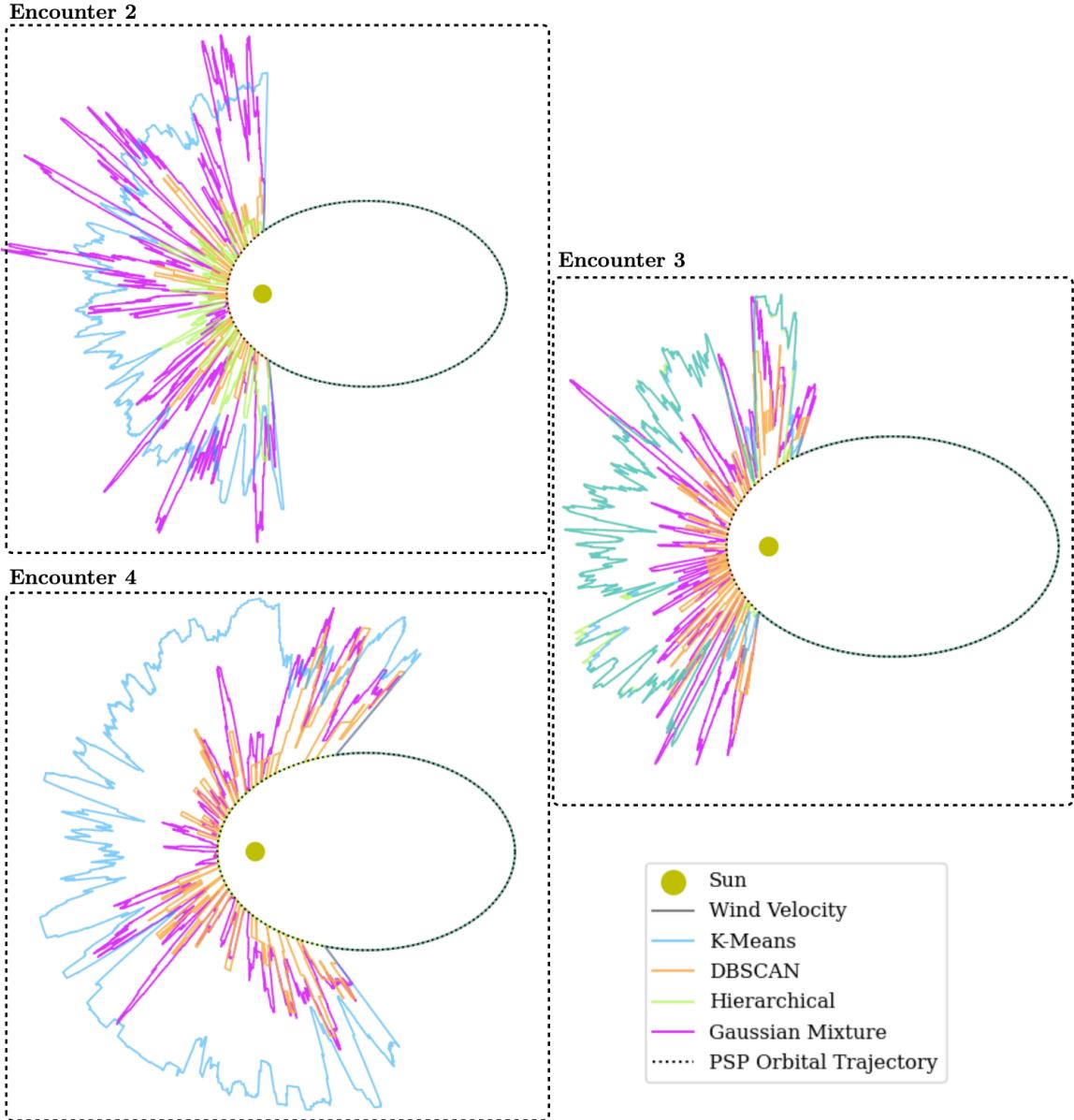


Figure 11: Plots depicting the averaged unsupervised ML prediction series overlaid with the orbital trajectories for second, third, and fourth encounters. These plots are generated using a rolling average window of 12-hours to collimate and smoothen the regional predictions of each model.

<sup>19</sup>By Averaging the model outputs we are able to agglomerate prediction instances to develop a detailed prediction distribution and subsequently conduct our study over extended switchback periods.

From figure 11 we observe several interesting features. Firstly, the relative amplitude of each signal indicates the level of localised recall. This is apparent as models with high recall as recorded in the cross-model comparison seem to possess very high amplitudes on the orbital overlay plot. These plots indicate the level and regions of attention for those models with relatively low recall. This attention (particularly evident in encounter four where peaks seem to be concentrated at the bottom left of the plot) provides us insight into the respective areas of high magnetic switchback activity. Over certain regions of the graph, all models show some level of correlation regarding the areas of high and low prediction densities. These peaks provide us with insight into the relative likelihood of switchbacks occurring within different regions of the orbital trajectory. We are also able to note several areas of similarity between all models. These interpretations seem to be related to the relative size, scale, occurrence of switchbacks.

Due to the nature of these plots and their relationship to the likelihood of switchback occurrence, we decide to evaluate these models with respect to the solar wind velocity. We chose to perform this separate means of validation, as the orbital relationship between the likelihood of switchbacks and the relative correlation with solar wind may provide insight into the applicability of the respective models (*see Literature Review*). These solar wind plots not only act to validate the model predictions, but also infer new scientific data regarding the relationship between the solar wind speed and magnetic switchbacks.

By incorporating the recorded SPC solar wind velocity for the fourth encounter<sup>20</sup> with the overlay plot, we identify regions of model correlation, and better understand the relationship between the equatorial solar wind velocity and the presence of magnetic switchbacks<sup>21</sup>. Figure 12 illustrates the resulting orbital plot for this fourth encounter:

---

<sup>20</sup>It came to our attention that there was no available/usable SPC data pertaining to the second and third encounter. This issue arose due to several instrumental issues (including: erroneous settings, non-ideal energy ranges, and general operational issues) compromising the quality of the acquired data during the majority of the perihelion periods of these two encounters (M. Stevens, private communication, 2022).

<sup>21</sup>We acquired the SPC solar wind velocity data from the following source:  
<http://sweap.cfa.harvard.edu/pub/data/sci/sweap/spc/L3/>

#### 4.3 ORBITAL ANALYSIS

Switchback prediction comparison across forth encounter orbital trajectory

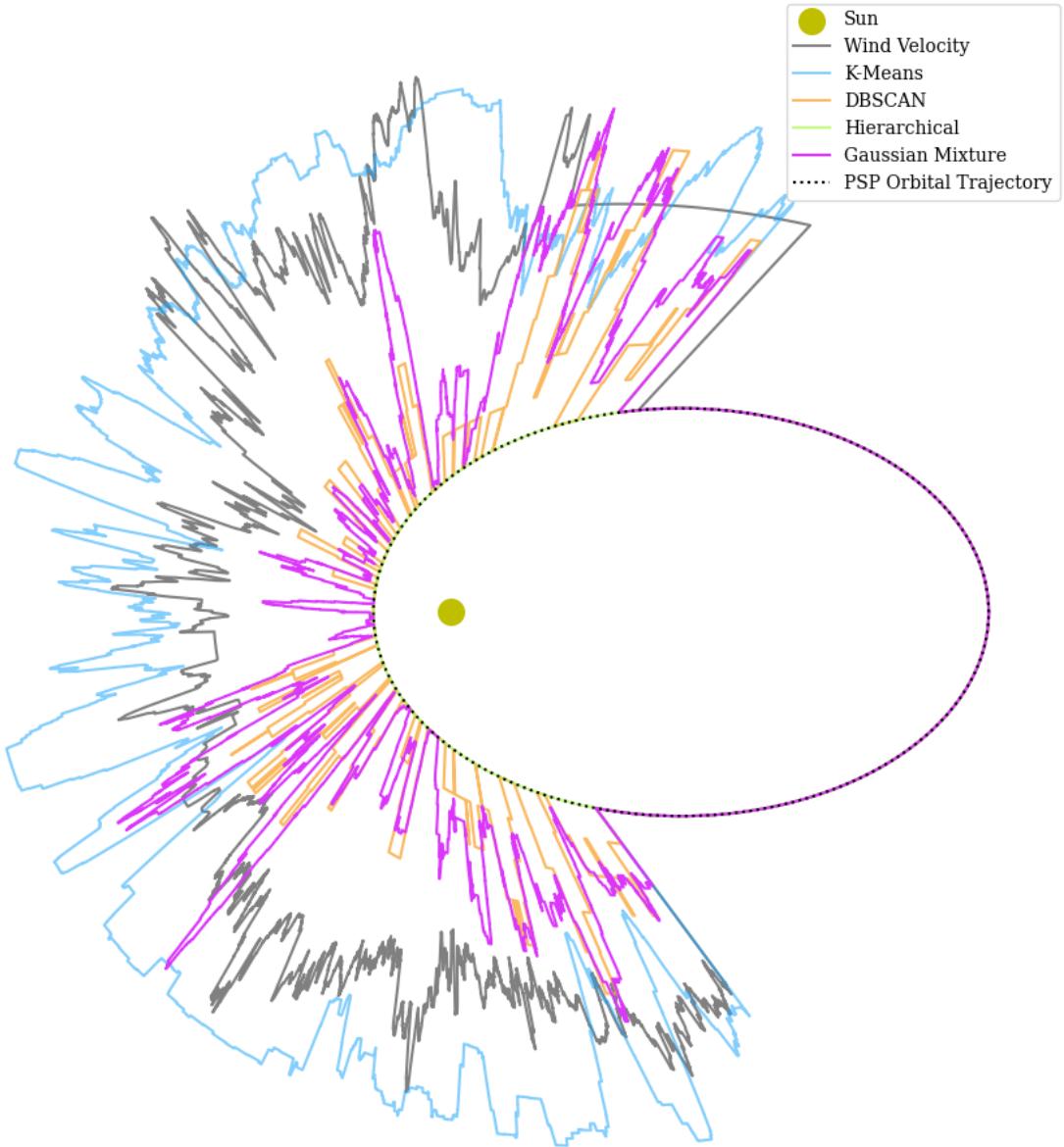


Figure 12: Orbital overlay diagram representing the averaged prediction components of each ML model, with the addition of SPC solar wind velocity data. We highlight that the amplitudes of the wind velocity signal corresponding to the respective velocity magnitude. The incorporation of the solar wind data (*dark grey*) exposes correlations within the prediction set, and further validates each model’s implementability.

From observation of figure 12 we identify several features that are not apparent in the other evaluations<sup>22</sup>. Firstly, it is evident that some level of correlation is apparent in the recorded plot, with most of this correlation occurring at the closer solar encounter distances. This correlation is not necessarily proportional to the amplitude (i.e. localised recall), rather these correlations seem to align with areas where multiple models simultaneously detect switchbacks. We also noted some algorithms that were thought to produce poor results (i.e. K-Means) show some degree of correlation to the data. Although the level of sensitivity is related to models with high accuracy.

---

<sup>22</sup>The remarks made regarding the solar wind analysis have not been supported by other encounter evaluations. Thus, any conclusions we draw are only specific to the encounter in question.

This plot provides us with insight into the relative relationship each model shares with that of the correlation data, as well as to how this wind velocity data relates to the occurrence of switchbacks. This method of evaluation can help us infer the level of model representability and how correlated these switchback predictions are relative to that of the velocity.

## 4.4 Discussion

Upon observation of these various performance aspects and following the processes outlined by the methodology and literary review, we collimate a conclusion as to which model is best suited for the classification of magnetic switchbacks. Throughout the analysis the algorithm predictions were highly variant relative to each-other, and are largely dictated by the model's underlying architecture and interpretation of the feature data.

From the various performance evaluations, we note that both the K-Means and Hierarchical models produce prediction results that are unsuitable for use in the classification of magnetic switchbacks. This poor performance quality arises as a result of these models' inability to produce accurate, consistent, and reliable predictions. These observations are a direct result of each models underlying architecture and learning behaviours on the provided feature set. One way to address this issue is to tailor features and scaling methods in a manner that makes them better adapted for these learning algorithms (although, the performance capacity these models may also be physically limited by the nature of the problem).

Conversely to our observation regarding the K-Means and Hierarchical models, the DBSCAN and GM mixture models show exceptional classification ability. This ideal performance is observable across all evaluations including the raw data comparison, cross-model comparison using statistical modes, and the orbital evaluation relative to the solar wind velocity. Not only do we see these desirable results in each case, but these algorithms produce invariant and reliable predictions across all encounters and parameter variations. These qualities are indicative of robust models and guarantees the results provided on unseen encounters will be satisfactory to a similar degree as that observed in this evaluation.

Thus, from all our observations conducted throughout this project, we conclude that the best unsupervised ML algorithms, for the application of identifying and classifying magnetic switchbacks is the DBSCAN and Bayesian Gaussian Mixture model<sup>23</sup>. Furthermore, depending on the condition of application, particular models may be more desirable. For example, if high accuracy and low model error is desired, then the DBSCAN model is preferable. However, if a more generalised algorithm is required, with a well-balanced and highly reliable prediction series, then the GM model is more appropriate for that particular application.

---

<sup>23</sup>This conclusion applies to all cases with conditions similar to that of the second, third, and fourth encounters. As our analysis did not assess the use of these models in switchback intervals with significantly different properties.

Along with the aforementioned utility of these models, the differing architectures provide unique traits and characteristics that can be exploited for further science and application. The DBSCAN model produces a unique cluster set on the provided features. This lightweight architecture ensures that each new encounter cluster is independent of the previous. Moreover, this model scales well with regards to the size of the provided feature set. This characteristic means that these models can be implemented very quickly and provide consistent results regardless of the initial criteria and sample size of the dataset (given a sample size that is statistically representative of the overall feature set).

Due to the semi-supervised and probabilistic nature of the Bayesian Gaussian Mixture model this algorithm is capable of an assortment of applications beyond the clustering of the provided feature set. The semi-supervised behaviour of GM algorithms is advantageous as they can be implemented as a supervised models on OOS datasets, making these current trained models ready for implementation in other datasets. This flexibility can allow these models to be implemented relatively seamlessly on unseen data allowing for comparison to be drawn between the differences and similarities of separate encounter observations<sup>24</sup>. The generative nature of these models is also beneficial as it can be utilised for inference and predictive modelling. This can allow these models to be implemented alongside magnetic field modelling algorithms to determine their validity and performance on realistic data.

Following the methodology outlined throughout this document, we identify and uncovered new fields of scientific potential that have arisen as a result of our findings. Some such future considerations include developing and experimenting the algorithmic performance using different architectures (beyond the four unsupervised models developed in this project). It is also worth experimenting with new and more tailored model features that aid the performance of the ML algorithms. These models can also be evaluated over more encounters (or potentially other Sun orbiting spacecraft with similar instrumentation), to provide a more in-depth insight into the underlying ability to classify magnetic switchbacks using machine learning and statistical methods. Lastly these current models can be analysed on more solar wind data, as the single encounter observation investigated in our research may not be sufficient to draw appropriate conclusions regarding the definitive relationship between the solar wind and occurrence of magnetic switchbacks.

---

<sup>24</sup>This can be done by training a separate model on the new data and comparing the associated parameters, allowing for comparisons between the different statistical distributions to be made possible.

## 5 Conclusions

The solar wind and associated physical phenomena are essential to study and understand as this space plasma is the primary medium through which solar activity is transferred/interacts with interplanetary space. The MHD behaviour of the solar wind results in the dynamic complexity which we observe within this plasma. One of the main driving forces of this complex motion is the interaction between the interplanetary magnetic field and this flowing substrate. In recent near-Sun observations, this magnetic field has been observed to kink and fully reverse in polarity particularly at close solar distances. These events are referred to as magnetic switchbacks and occur most prominently in wind originating from coronal holes. The causes of such events are to a large degree unknown, and the implications are only now being discovered. These switchbacks can have significant energy implications on the surrounding solar wind and thus it is of high importance to identify and characterise these events.

To address this complex issue, we utilise statistical and machine learning methods. We develop, train, and evaluate four unsupervised machine learning models on feature data extracted from the Parker Solar Probs magnetometers in order to appropriately characterise these events. By using statistical tools such as those outlined by Bale et al. (2019) we develop and evaluate these machine learning models to identify which algorithm was best suited for the application of switchback classification.

In order to properly gauge the capability and applicability of the four unsupervised models at identifying and classifying the presence of magnetic switchbacks, we conduct a series of independent and complementary analyses. These assessments allow us to identify the various performance factors such the visual comparison of each instance's prediction relative to the raw magnetic field data, the classification scoring and classification efficacy of each model relative to the statistical mode basis, as well as each algorithm's implementability on scientific data including the model comparison to solar wind speeds as well as prediction accuracy relative to the Parker Solar Prob's orbital position. When conducting these separate analyses, we identify characteristics and features in the prediction assessments that are indicative of an idealistic model for switchback classification.

Throughout this project, we focus on evaluating the models based on the properties which we deem to be important. Some such considerations included the relative accuracy at which a model is capable of making predictions. This involved the predicting of the occurrence and absence of magnetic switchbacks, both visually and analytically. Along with this high degree of desired accuracy we also look to select models which constantly demonstrate this level of performance across the entire evaluated data range. This consistency ensures the models are capable of maintaining high accuracy across differing switchback formations. Analogously to these evaluation criteria, we also aim to identify models that showed strict repeatability across all encounters. This repeatability acts to insure the utility/applicability of these algorithms on unseen or unencountered data.

## 5 CONCLUSIONS

From the model evaluation process, we note that the K-Means and Hierarchical models were not suited for the application of switchback classification. This poor performance is a result of each algorithm's inability to accurately model the data at hand (as a result of their underlying architectures). Conversely, we observed that the DBSCAN and Bayesian Gaussian Mixture models are very well suited at predicting the more probabilistic switchback data.

These algorithms showed satisfactory results over across our evaluation criteria, including the level of consistency and reliability we aspired for. Both these algorithms demonstrate statistical performance scores with accuracies all greater than 90 %. Furthermore, when these predictions are averaged and compared to orbital scientific data, we notice that each algorithm maps the solar wind with a relatively high degree of localised correlation (particularly at close heliocentric distances where the presence of switchbacks are most notable).

Thus, due to the associated advantages of these models and the additional utility which was provided by these algorithmic architectures, we arrive at the conclusion that these models are the most appropriate tools for the application of switchback classification. More particularly the DBSCAN model is best used in applications that require high accuracy and throughput with the tolerance of a low recall rate, whereas the GM model is better suited for more generalised applications that require more robust and applicable data.

Following the conducted method, we identify several areas of further research that can be conducted to advance and build on what we have achieved. This includes considerations such as improving the quality of model features and the types of evaluated ML architectures, improving the statistical reliability of the current results by conducting the scientific analysis over larger encounter ranges, and by studying the models using more representative solar wind velocity data a deeper intuition can be developed into the relational characteristics these models share with that of both the magnetic field and solar wind.

# References

- Alfvén, H. (1942), ‘Existence of electromagnetic-hydrodynamic waves’, *Nature* **150**(3805), 405–406.
- Amaya, J., Dupuis, R., Innocenti, M. E. & Lapenta, G. (2020a), ‘Visualizing and interpreting unsupervised solar wind classifications’, *Frontiers in Astronomy and Space Sciences* **7**, 66.
- Amaya, J., Dupuis, R., Innocenti, M. E. & Lapenta, G. (2020b), ‘Visualizing and interpreting unsupervised solar wind classifications’, *Frontiers in Astronomy and Space Sciences* **7**.
- URL:** <https://www.frontiersin.org/article/10.3389/fspas.2020.553207>
- Andreassen, A., Feige, I., Frye, C. & Schwartz, M. D. (2019), ‘Junipr: a framework for unsupervised machine learning in particle physics’, *The European Physical Journal C* **79**(2), 1–24.
- Bale, S., Badman, S., Bonnell, J., Bowen, T., Burgess, D., Case, A., Cattell, C., Chandran, B., Chaston, C., Chen, C. et al. (2019), ‘Highly structured slow solar wind emerging from an equatorial coronal hole’, *Nature* **576**(7786), 237–242.
- Bale, S., Goetz, K., Harvey, P., Turin, P., Bonnell, J., De Wit, T. D., Ergun, R., MacDowall, R., Pulupa, M., André, M. et al. (2016), ‘The fields instrument suite for solar probe plus’, *Space science reviews* **204**(1), 49–82.
- Bandyopadhyay, R., Goldstein, M., Maruca, B., Matthaeus, W., Parashar, T., Ruffolo, D., Chhiber, R., Usmanov, A., Chasapis, A., Qudsi, R. et al. (2020), ‘Enhanced energy transfer rate in solar wind turbulence observed near the sun from parker solar probe’, *The Astrophysical Journal Supplement Series* **246**(2), 48.
- Bavassano, B. & Bruno, R. (1992), ‘On the role of interplanetary sources in the evolution of low-frequency alfvénic turbulence in the solar wind’, *Journal of Geophysical Research: Space Physics* **97**(A12), 19129–19137.
- Bourouaine, S. & Perez, J. C. (2019), ‘On the interpretation of parker solar probe turbulent signals’, *The Astrophysical Journal Letters* **879**(1), L16.
- Bousquet, O., von Luxburg, U. & Rätsch, G. (2011), *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, Vol. 3176, Springer.
- Case, A. W., Kasper, J. C., Stevens, M. L., Korreck, K. E., Paulson, K., Daigneau, P., Caldwell, D., Freeman, M., Henry, T., Klingensmith, B. et al. (2020), ‘The solar

## REFERENCES

- probe cup on the parker solar probe', *The Astrophysical Journal Supplement Series* **246**(2), 43.
- Constantinopoulos, C., Titsias, M. & Likas, A. (2006), 'Bayesian feature and model selection for gaussian mixture models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(6), 1013–1018.
- Education, I. (2020), 'What is unsupervised learning?'.  
**URL:** <https://www.ibm.com/cloud/learn/unsupervised-learning>
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise., in 'kdd', Vol. 96, pp. 226–231.
- Farglette, N., Lavraud, B., Rouillard, A. P., Réville, V., De Wit, T. D., Froment, C., Halekas, J. S., Phan, T. D., Malaspina, D. M., Bale, S. D. et al. (2021), 'Characteristic scales of magnetic switchback patches near the sun and their possible association with solar supergranulation and granulation', *The Astrophysical Journal* **919**(2), 96.
- Ferrini, D. (2021), *Individual Project Initial Report*.  
**URL:** [https://ev.turnitinuk.com/app/carta/en\\_us/?u=16113048&o=166965069&lang=en\\_us&student\\_user=1](https://ev.turnitinuk.com/app/carta/en_us/?u=16113048&o=166965069&lang=en_us&student_user=1)
- Ferrini, D. (2022), *Individual Project Interim Report*.  
**URL:** [https://ev.turnitinuk.com/app/carta/en\\_us/?student\\_user=1&u=16113048&lang=en\\_us&o=174522516](https://ev.turnitinuk.com/app/carta/en_us/?student_user=1&u=16113048&lang=en_us&o=174522516)
- Froment, C., Krasnoselskikh, V., de Wit, T. D., Agapitov, O., Farglette, N., Lavraud, B., Larosa, A., Kretzschmar, M., Jagarlamudi, V., Velli, M. et al. (2021), 'Direct evidence for magnetic reconnection at the boundaries of magnetic switchbacks with parker solar probe', *arXiv preprint arXiv:2101.06279*.
- Garner, R. (2016), 'Understanding the magnetic sun', <https://www.nasa.gov/feature/goddard/2016/understanding-the-magnetic-sun>.
- Géron, A. (2019), *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, " O'Reilly Media, Inc." .
- Guo, Y. (2008), Trajectory design of solar probe+ using multiple venus gravity assists, in 'AIAA/AAS Astrodynamics Specialist Conference and Exhibit', p. 7365.
- Harvard (2020), *user guide for parker solar probe sweap investigation data products 2020*, 1.3 edn.  
**URL:** [http://sweap.cfa.harvard.edu/sweap\\_data\\_user\\_guide.pdf](http://sweap.cfa.harvard.edu/sweap_data_user_guide.pdf)
- Hatfield, M. (2021), 'Switchbacks science: Explaining parker solar probes magnetic puzzle', <https://www.nasa.gov/feature/goddard/2021/switchbacks-science-explaining-parker-solar-probe-s-magnetic-puzzle>.
- Jannet, G., Dudok de Wit, T., Krasnoselskikh, V., Kretzschmar, M., Fergeau, P., Bergerard-Timofeeva, M., Agrapart, C., Brochot, J.-Y., Chalumeau, G., Martin, P. et al. (2021), 'Measurement of magnetic field fluctuations in the parker solar probe and solar orbiter missions', *Journal of Geophysical Research: Space Physics* **126**(2), e2020JA028543.

- Johnson, S. C. (1967), ‘Hierarchical clustering schemes’, *Psychometrika* **32**(3), 241–254.
- Kasper, J. C., Bale, S. D., Belcher, J. W., Berthomier, M., Case, A. W., Chandran, B. D., Curtis, D., Gallagher, D., Gary, S., Golub, L. et al. (2019), ‘Alfvénic velocity spikes and rotational flows in the near-sun solar wind’, *Nature* **576**(7786), 228–231.
- Kasper, J., Klein, K., Lichko, E., Huang, J., Chen, C., Badman, S., Bonnell, J., Whittlesey, P., Livi, R., Larson, D. et al. (2021), ‘Parker solar probe enters the magnetically dominated solar corona’, *Physical review letters* **127**(25), 255101.
- Keiling, A. (2021), ‘The dynamics of the alfvénic oval’, *Journal of Atmospheric and Solar-Terrestrial Physics* **219**, 105616.
- URL:** <https://www.sciencedirect.com/science/article/pii/S1364682621000778>
- Krishna, K. & Narasimha Murty, M. (1999), ‘Genetic k-means algorithm’, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **29**(3), 433–439.
- Lee, D.-S., Hull, J. & Erol, B. (2003), A bayesian framework for gaussian mixture background modeling, in ‘Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)’, Vol. 3, pp. III–973.
- Likas, A., Vlassis, N. & J. Verbeek, J. (2003), ‘The global k-means clustering algorithm’, *Pattern Recognition* **36**(2), 451–461. Biometrics.
- URL:** <https://www.sciencedirect.com/science/article/pii/S0031320302000602>
- McLachlan, G. J. & Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.
- McManus, M. D., Bowen, T. A., Mallet, A., Chen, C. H., Chandran, B. D., Bale, S. D., Larson, D. E., de Wit, T. D., Kasper, J. C., Stevens, M. et al. (2020), ‘Cross helicity reversals in magnetic switchbacks’, *The Astrophysical Journal Supplement Series* **246**(2), 67.
- Murtagh, F. & Contreras, P. (2012), ‘Algorithms for hierarchical clustering: an overview’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(1), 86–97.
- Obridko, V. & Vaisberg, O. (2017), ‘On the history of the solar wind discovery’, *Solar System Research* **51**(2), 165–169.
- Parker, E. (1965), ‘Dynamical theory of the solar wind’, *Space Science Reviews* **4**(5–6), 666–708.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research* **12**, 2825–2830.
- Roberts, D. A., Ghosh, S., Goldstein, M. L. & Mattheaus, W. H. (1991), ‘Magneto-hydrodynamic simulation of the radial evolution and stream structure of solar-wind turbulence’, *Physical review letters* **67**(27), 3741.
- Roberts, D. A., Karimabadi, H., Sipes, T., Ko, Y.-K. & Lepri, S. (2020), ‘Objectively determining states of the solar wind using machine learning’, *The Astrophysical Journal* **889**(2), 153.

- Roberts, S., Husmeier, D., Rezek, I. & Penny, W. (1998), ‘Bayesian approaches to gaussian mixture modelling’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11), 1133–1142.
- Sadykov, V. (2020), On clustering of machine learning attempts in heliophysics: Examples and general picture, in ‘EarthCube RCN Workshop: Machine Learning in Heliophysics and Space Weather Forecasting: Advances, Perspectives and Synergies’, number ARC-E-DAA-TN77358.
- Solanki, S. K., Inhester, B. & Schüssler, M. (2006), ‘The solar magnetic field’, *Reports on Progress in Physics* **69**(3), 563.
- Tsinganos, K., Sauty, C., Surlantzis, G., Trussoni, E. & Contopoulos, J. (1996), ‘On the relation of limiting characteristics to critical surfaces in magnetohydrodynamic flows’, *Monthly Notices of the Royal Astronomical Society* **283**(3), 811–820.
- Verscharen, D., Klein, K. G. & Maruca, B. A. (2019), ‘The multi-scale nature of the solar wind’, *Living Reviews in Solar Physics* **16**(1), 1–136.
- Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O. & Kurimo, M. (2011), ‘Empirical comparison of evaluation methods for unsupervised learning of morphology.’, *Trait. Autom. des Langues* **52**(2), 45–90.
- Whittlesey, P. L., Larson, D. E., Kasper, J. C., Halekas, J., Abatcha, M., Abiad, R., Berthomier, M., Case, A., Chen, J., Curtis, D. W. et al. (2020), ‘The solar probe analyzers—electrons on the parker solar probe’, *The Astrophysical Journal Supplement Series* **246**(2), 74.

# Appendix I - Comprehensive Data Capture

For all raw data products used in this project we ensure to apply appropriate data cleaning practices. This includes filtering the data ranges and anomalistic values, removing any missing instances, and interpolating data with differing cadences.

## Magnetic Field Data Acquisition Process

To capture and frame the data in a usable format, we followed several procedures and steps. In order to acquire the data for use as raw inputs in the feature engineering process, we utilised two primary sources. Namely the open source [magnetic field repository](#), and a separate .dat file containing the fixed [solar proximity data](#) of PSP.

To acquire the magnetic field data, we executed the respective procedure as outlined:

- We source the data from individual .cdf file formats, with each file representing one days worth of magnetometer data with a constant sampling cadence of one-minute.
- This data is requested from the source using a Python web API which we have developed.
- In order to preserve computational resources we discard all unnecessary or meta data from these .cdf files
- We save the respective epoch and RTN magnetometer data in DataFrame format in order to ensure the data is indexable and correctly formatted.
- We sequentially load each file across all desired encounter windows and expand this data to the common DataFrame containing all the raw encounter specific data.

Following this outlined process, we are able to store the raw magnetic field data for all RTN coordinates across the evaluated encounters. This raw data is depicted in figures 13, 3, and 14 respectively.

## Second and Fourth Encounter

As mentioned earlier (*see Materials and Methods*) the raw magnetic field data for all evaluated encounters was recorded, however for monotonous reasons we did not include the second and fourth encounters in the main segment of the report. This exclusion is a result of the consistency and similarities between the data across all encounter periods. Figures 13, and 14 depict the raw field plots for the second and fourth encounters respectively.

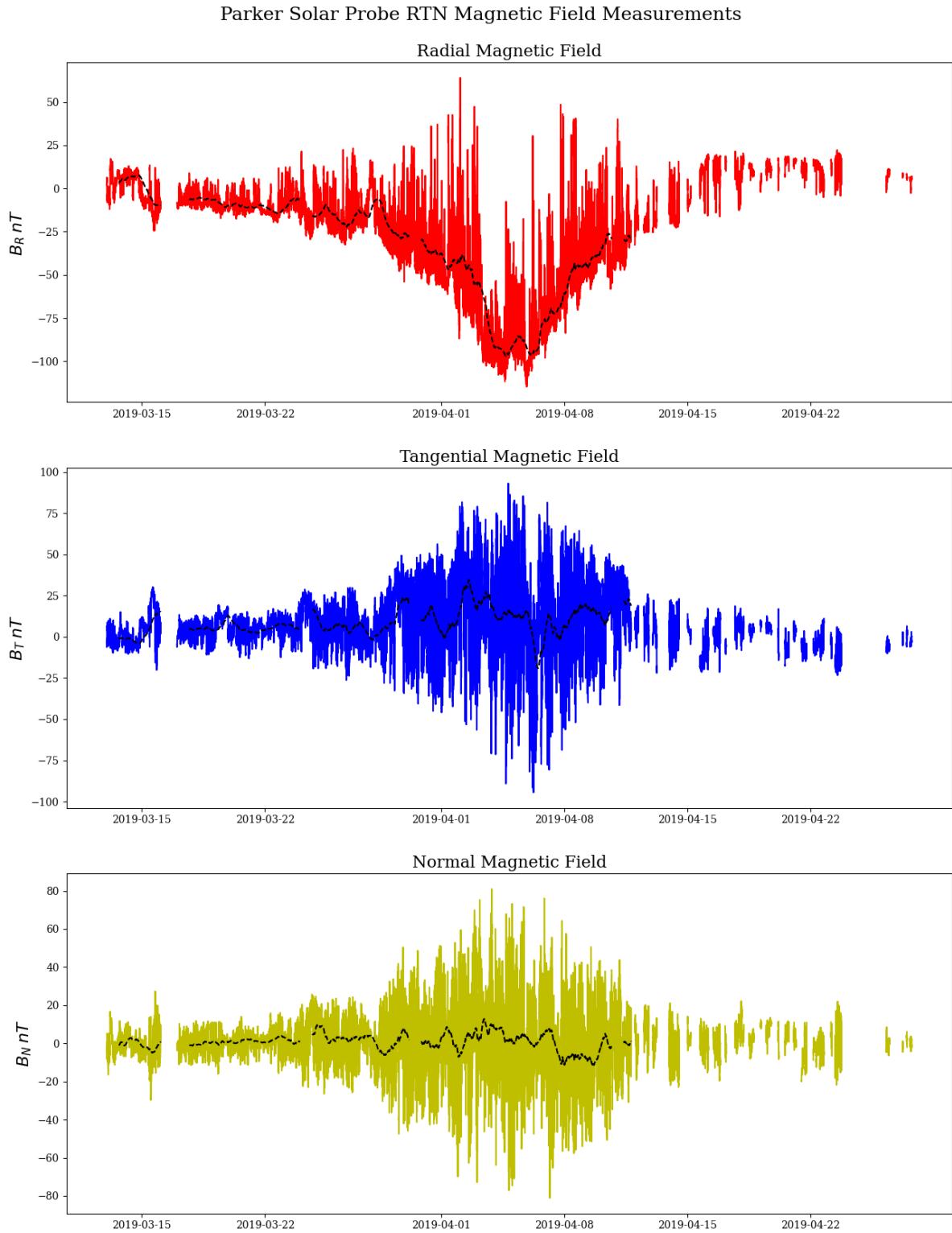


Figure 13: Raw data plots containing Parker Solar Probe magnetometer vector components for second encounter (13/03/2019 – 29/04/2019). This plot illustrates the comparison of observed radial (red), tangential (blue), and normal (yellow) magnetic field measurements, relative to the mean expected background field profile (black).

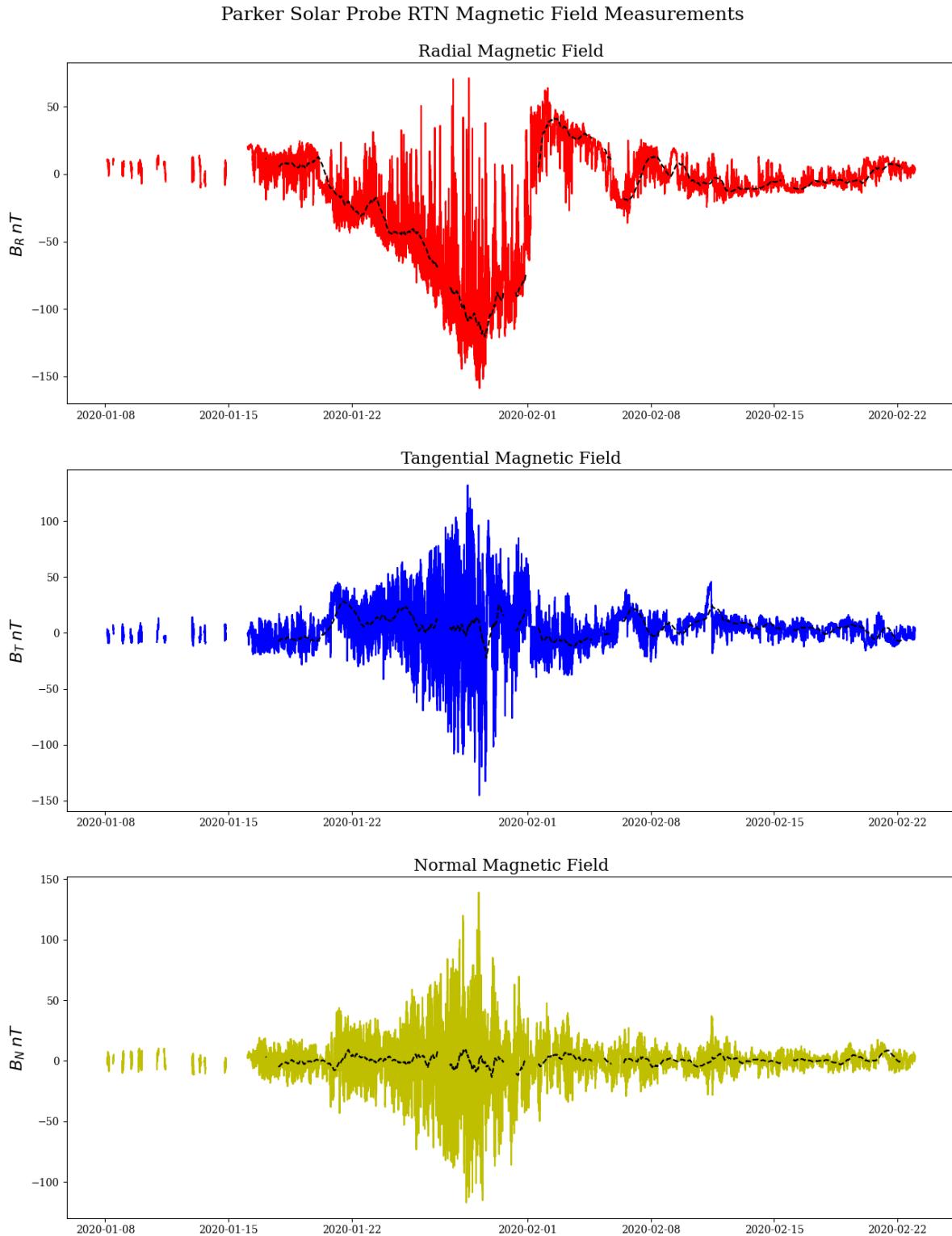


Figure 14: Raw data plots containing Parker Solar Probe magnetometer vector components for fourth encounter (08/01/2020 – 27/02/2020). This plot illustrates the comparison of observed radial (red), tangential (blue), and normal (yellow) magnetic field measurements, relative to the mean expected background field profile (black).

## Proximity Data Acquisition Process

In order to source the data pertaining to the Parker Solar Probe's heliocentric distance, we follow a slightly different procedural method (as opposed to the magnetometer data acquisition) to arrive at the same format. This process is outlined as follows:

- We load the data from its online source to a local repository for easier access and implementation. This data contains the respective heliocentric distance for every day of operation (represented in astronomical units), and possesses a sample cadence of 24-hours across the entire mission life.
- From this data we sample the desired encounter windows by specifying the required date range.
- This data is interoperated using linear projection mechanisms<sup>25</sup>, to ensure this data matches the cadence of the raw magnetic field data.
- We merge and align the proximity data for each encounter with each respective magnetic field **DataFrame**, in order to ensure coherency among the separate data formats.

The raw product of this data acquisition process will form the basis from which the feature data is to be developed. Thus, it is essential that we ensure the time indices and formatting is consistent and uniform throughout the data acquisition process. As this will allow the subsequent ML algorithms to be developed appropriately, and that the methodology of the project is structured in a reproducible manner (for further evaluation on later encounters or different data structures).

## Solar Wind Data Acquisition

Due to the similarities in repository structure of the SPC data and magnetometer data, we sample this solar wind velocity data using the same process as outlined earlier. We also note that this solar wind velocity data is in the processing order of level three. Due to the nature of how the SPC derives its measurements, these velocity data are inferred through the application of a nonlinear least-squares model. The velocity predictions are derived by fitting the SPC data to a set Maxwellian ion populations. As a result of this estimation, various errors and considerations are to be accounted for when representing the data. For our analysis of the fourth encounter (*see Results and Discussion*), we choose to use the  $V_p$  velocity model. The  $V_p$  protocol signifies the estimated velocity of the centre of mass of proton field. This bulk velocity parameter directly corresponds to the true solar wind speed and is thus useful when analysing our machine learning models.

---

<sup>25</sup>We are required to perform this interpolation, as the 24-hour cadence of the raw data is insufficient to encompass the typical life span of a magnetic switchback. This consideration ensures that we can better sample our predictions, and improved the overall model accuracy

## Extended Interpretation of Statistical Modes

Using the cleaned and imputed data from the PSP magnetometers we are able to apply and visualise the respective statistical modes for these separate encounter periods. From this analysis we observe several characteristics including how these data plots differ from the third encounter. For example, one such anomaly that is notable in the second and fourth encounter periods includes how the radial field polarity seems to be predominantly positive at the exit of the these encounters. These features differ from the more symmetric distribution noted in the third plot (illustrated in figure 4) and can potentially affect how the ML models perform on this data. Figure 15 highlights this difference as well as other differences which results from the separate modal analyses:

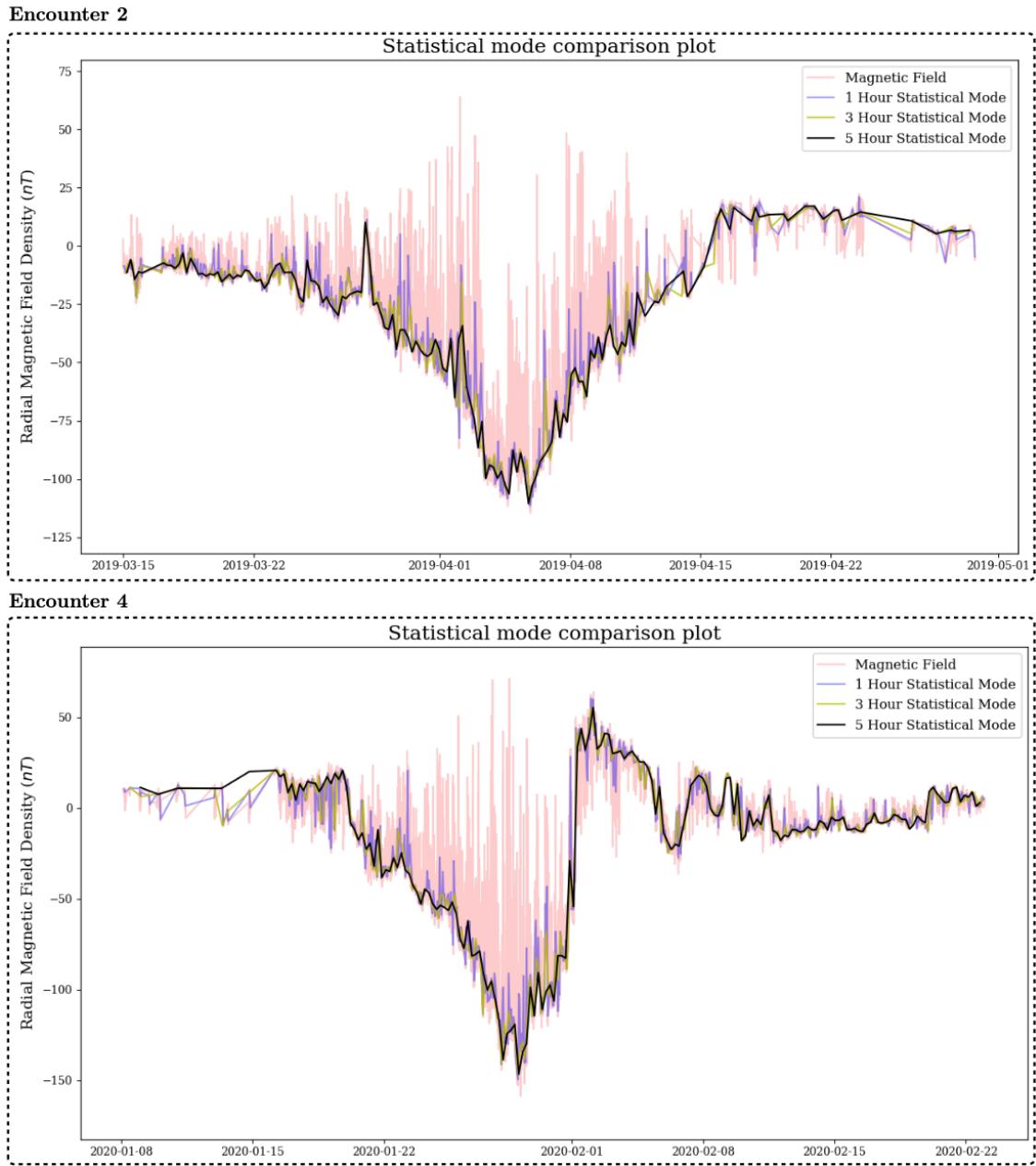


Figure 15: Diagram showing results of statistical modal analysis on the radial magnetic field data for the second and fourth solar encounter of PSP. These plots show areas of differences and similarities when compared to that of the third encounter analysis results.

From figure 15 we note that the statistical modal analysis is capable of identifying the background trend of the magnetic field data. Furthermore, we also see that the modes of the fourth encounter show significantly less variability over differing mode cadences. This seems to indicate that the switchbacks recorded over the fourth encounter had on average shorter lifespans (as most of the switchbacks were discarded from the magnetic field profile with the lowest sample cadence of one-hour).

Following this extended statistical modal analysis, we conclude that although the observations in each encounter period depict differing degrees of variability, these plots seem to share large scale similarities. These similarities show that the statistical modes are consistent enough to be used as a benchmark for switchback classification and comparison. These small scale differences can also potentially compliment the ML analysis, and provide some degree of variability in our analysis and comparison.

# Appendix II - Model Architecture & Theory

As a result of time considerations, human error, and computational resources, we deem it infeasible to manually identify each switchback event over all the evaluated encounters. This lack of target/labeled data implies that our selection of machine learning algorithm is limited. As outlined previously (*see Literature Review*), we are able to narrow down our selection to models that fall under the class of unsupervised machine learning.

These unsupervised models are ideally suited for the classification of magnetic switchbacks due to its associated benefits and proven history of feasible implementation in solar-physical applications. Moreover, in order to ensure the simplicity and repeatability of our models, we act to minimise the number of features used in this classification problem as well as the amount of data pre-processing and computation that was required.

By outlining the architectural method employed by each algorithm, we illustrate the different learning architectures and model functionalities. Allowing us to perform a comparative analysis into the various algorithms and their suitability for identifying and classifying magnetic switchbacks on the provided feature data.

## K-Means Architecture

We can detail the algorithmic process of the K-Means architecture using pseudocode. This format enables us to visualise the initialisation process of the algorithms, including how the features were preprocessed and fitted to the model as well as to how each model formulated and converged on its clusters. The pseudo-algorithm depicted in figure 16 overleaf, illustrates how the K-Means model was implemented on the magnetic switchback feature data:

---

**Algorithm 1:** K-Means Initialisation

---

**Data:** Encounter specific magnetic field and prob proximity data

- 1 **Preprocessing (Raw Input Data):**
  - ⇒ create feature components from raw input
- 2       magnetic field and proximity data
- 3       ⇒ apply Scikit-Learn standard scaling to all feature components
- 4       ⇒ generate and return random subset batch containing 20 % of feature data
- 5
- 6
- 7 preprocess feature data (1)  $\leftarrow$
- 8 initialise `sklearn.cluster.kmeans` model with  $K = 2$  centroids
- 9 fit model with preprocessed feature set
- 10 **for Epoch in K-Means Enumeration:**
- 11     assign each preprocessed instance to nearest centroid
- 12     calculate mean coordinate of each clustered instance
- 13     assign mean coordinate as the updated centroid location
- 14 **end**
- 15 save positions of the converged centroids
- 16 **with Fully Sampled Features & Cluster Centroids:**
- 17     assign labels to all feature instances, depending on respective centroid proximity
- 18     convert cluster predictions to `DataFrames`
- 19     save cluster predictions for given encounter
- 20 **end**

**Result:** Export cluster prediction labels, and trained K-Means centroid cluster model

---

Figure 16: Methodology outlining the algorithmic functionality of the K-Means unsupervised model. This pseudo-algorithm shows the various steps and considerations that we made in order to ensure an efficient and appropriately fit model.

We can see from figure 16 that the K-Means architecture is based off of an iterative model which achieves centroidal convergence in relatively few steps. Due to the random initialisation of these centroids, the repeatability and convergence of these algorithms are not guaranteed. Thus we must consider multiple simulations before we can ensure the validity of the prediction results.

Additionally, we see how this iterative algorithm and its means of computation can scale disproportionately with an increased feature sample size. Thus from this pseudo-algorithm we identify the importance of implementing mini-batch K-Means, and how this additional preprocessing step can aid in the resources preservation and generalisation of the model.

## DBSCAN Architecture

The density based approach of the DBSCAN model is vastly different when compared to other commonly used unsupervised algorithms. This unique learning approach is independent of the number of predefined clusters, and can make these nonparametric models ideal for classifying unknown datasets with an undefined number of ideal clusterings. The implementational simplicity of these models can be identified in the pseudo-algorithm shown in figure 17 below:

---

**Algorithm 2: DBSCAN Initialisation**


---

**Data:** Encounter specific magnetic field and prob proximity data

- 1 **Preprocessing (Raw Input Data):**
  - ⇒ create feature components from raw input
- 2               magnetic field and proximity data
- 3               ⇒return raw feature data
- 4 initialise `sklearn.cluster.DBSCAN` model with sample noise hyper-parameter  $\varepsilon \approx 4$
- 5 **with Feature data (1)↔:**
  - 6     **for Instance in Feature Set:**
    - 7         count the number of samples that occur within the  $\varepsilon$  threshold
    - 8         all neighbouring instances are clustered together
    - 9         **if N-samples < min sample hyper-parameter:**
      - 10             define *Instance* as core instance
    - 11         **else :**
      - 13             define *Instance* as anomaly
    - 14     **end**
    - 16     cluster all sequences of core instances
    - 17     assign anomalistic labels to unique/outlier cluster
  - 18 **end**
  - 19 convert cluster predictions to `DataFrames`
  - 20 save predicted cluster labels

**Result:** Export cluster prediction labels, and trained DBSCAN model

---

Figure 17: Methodology outlining the functionality of the DBSCAN model. This diagram demonstrates the instance relative learning process adopted by these algorithms.

From this figure 17 we observe how the relative spacing of each instance is crucial for the formulation of the respective clusters. It is also evident how these minimalistic data preprocessing requirements can lead to the robustness of these algorithms. Enabling us to implement these models on larger unseen datasets with relatively predictable/expected results.

## Agglomerative Hierarchical Model Architecture

The clustering approach adopted by the Agglomerative Hierarchical model, is considered ‘greedy’. This is because the model primarily aims to maximise the immediate euclidean space of the feature set, rather than reducing cluster variation in latter agglomerations. Thus, due to this learning behaviour, these models are particularly susceptible to initial conditions of the feature space and therefore require careful preprocessing considerations. These preliminary factors can be detailed in the algorithm depicted in figure 18:

---

### **Algorithm 3:** Hierarchical Initialisation

---

**Data:** Encounter specific magnetic field and prob proximity data

**1 Preprocessing(Raw Input Data):**

- ⇒ create feature components from raw input

**2**       magnetic field and proximity data

**3**       ⇒ apply Scikit-Learn standard scaling to all

**4**       feature components

**5**       ⇒ return preprocessed features

**6** preprocess feature data (1)←

**7** initialise `sklearn.cluster.AgglomerativeClustering` model  
with  $n\_clusters = 2$

**8 while** Feature Space is not  $n\_clusters = 2$ :

**9**     calculate euclidean distance between each other  
feature instance in *Feature Space*

**10**    sample instance pair with smallest distance

**11**    assign mean euclidean distance, between this  
minimum instance pair, as new instance in  
*Feature Space*

**12 break**

**13** assign cluster labels based on agglomeration pattern

**14** save decision boundary of agglomerative model

**15** convert cluster predictions to `DataFrames`

**16** save predicted cluster labels

**Result:** Export cluster prediction labels, and trained  
hierarchical model decision boundary  
parameters

---

Figure 18: Methodology outlining the algorithmic functionality of the Agglomerative Hierarchical unsupervised clustering model. This figure shows how this model formulates the various levels of groupings and how the idealised number of clusters are identified.

From this pseudo-algorithm it is evident how this architecture develops its Euclidean based agglomerations, as well as to how the desired cut-off/cluster size is determined. This process also highlights how outliers and data anomalies must be mitigated to ensure an evenly distributed feature space for appropriate model development.

## Bayesian Gaussian Mixture Model Architecture

Similarly to the K-Means model, the Bayesian Gaussian Mixture model is an iterative algorithm. However, this model differs from that of K-Means as it requires more steps in order to achieve convergence. This probabilistic model is also required to update more parameters to achieve the high degree of freedom that is usually associated with these GM algorithms. Figure 19 displays the various steps and considerations that must be made in order for these algorithms to be initialised and fitted to the switchback feature data:

---

### Algorithm 4: Bayesian GM Initialisation

---

**Data:** Encounter specific magnetic field and prob proximity data

```

1 Preprocessing(Raw Input Data):
    ⇒ create feature components from raw input
2           magnetic field and proximity data
3           ⇒ return preprocessed features

4 preprocess feature data (1)←
5 initialise sklearn.mixture.BayesianGaussianMixture
    model with  $n\_components = 2$ 
6 initialise mixture components
7 for Epoch in GM Enumeration:
8     calculate the probability of each instance belonging
        to a given mixture component
9     assign each preprocessed instance to most likely
        mixture
10    update cluster variables to minimise cluster mean
        and standard deviation
11 end

12 save positions of the converged mixture components
13 with Fully Sampled Features & Mixture Variables:
14     assign labels to all feature instances, depending on
        respective centroid proximity
15     convert cluster predictions to DataFrames
16     save cluster predictions for given encounter
17 end

Result: Export cluster prediction labels, and trained
GM cluster model

```

---

Figure 19: Methodology outlining the algorithmic functionality of the Bayesian Gaussian Mixture model.

From figure 19 we note how the GM model defines instance probabilities and how Bayesian inference is incorporated into the foundational Gaussian mixture model to derive the respective clusterings.

By implementing these features on differing architectures we are able to evaluate the relative performance of each algorithm, and subsequently identify which architecture is most ideal for classifying switchbacks on the provided dataset.

# Appendix III - Detailed Results Analysis

The results and information acquired during our scientific evaluation of each unsupervised algorithm extends beyond the findings summarised formerly (*see Results and Discussion*). We can further explore these results by studying the relationships between the performance of the unsupervised models on all other evaluated encounters (namely the second and fourth solar encounters) not detailed previously.

## Thorough Prediction Evaluation

From the visualisation of the model predictions for each respective encounter, we conclude several observations. Firstly, we note that the prediction quality and consistency of most models remain relatively invariant across all evaluated encounters. This is evident as both the DBSCAN and Bayesian Gaussian Mixture showed ideal and repeatable performance with similar representation of switchbacks over regions with differing density in the case of all encounters. On a similar note, the K-Means model seems to show the same level of unappreciable prediction quality across the second and fourth encounters as with the previously evaluated third encounter. This algorithm also showed the same level of model breakdown over regions of close encounter where the density and consistency of switchbacks was highest.

All these observations are consistent with the exception of the Agglomerative Hierarchical model, whereby the resulting predictions of the second and fourth encounter show almost no degree of data interpretation. This exceptionally variable recall rate indicates that the observations made regarding the third encounter are anomalistic as the majority of predictions from this Hierarchical model tend to be biased towards outliers.

Another point of observation is the predominantly positive magnetic field polarity at the exit of the second and fourth encounter. These differing features seem to have no direct effect on the prediction performance of the K-Means, DBSCAN and GM models. However, these feature inconsistencies impact the quality of the Hierarchical model's results more so than other architectures. This high feature sensitivity, of the Hierarchical algorithm, is most likely what results in this inconsistent and poor prediction quality. The points and observations we raise can be better illustrated by overlaying each model predictions with the respective radial magnetic field plot for each encounter. These results for the second and fourth orbital encounters can be seen in the subplots depicted in figure 20:

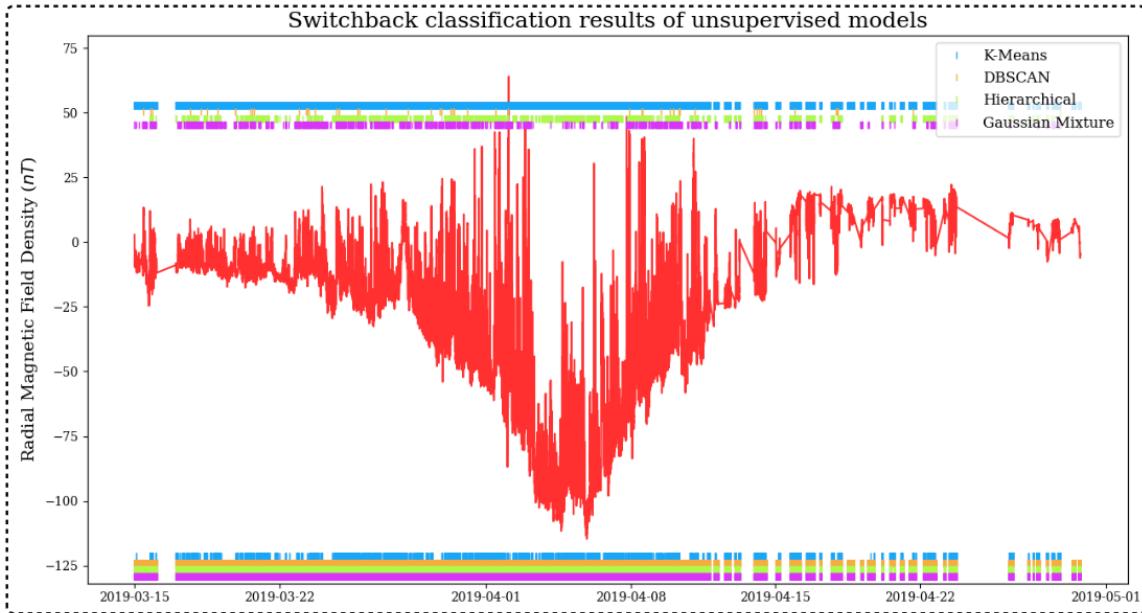
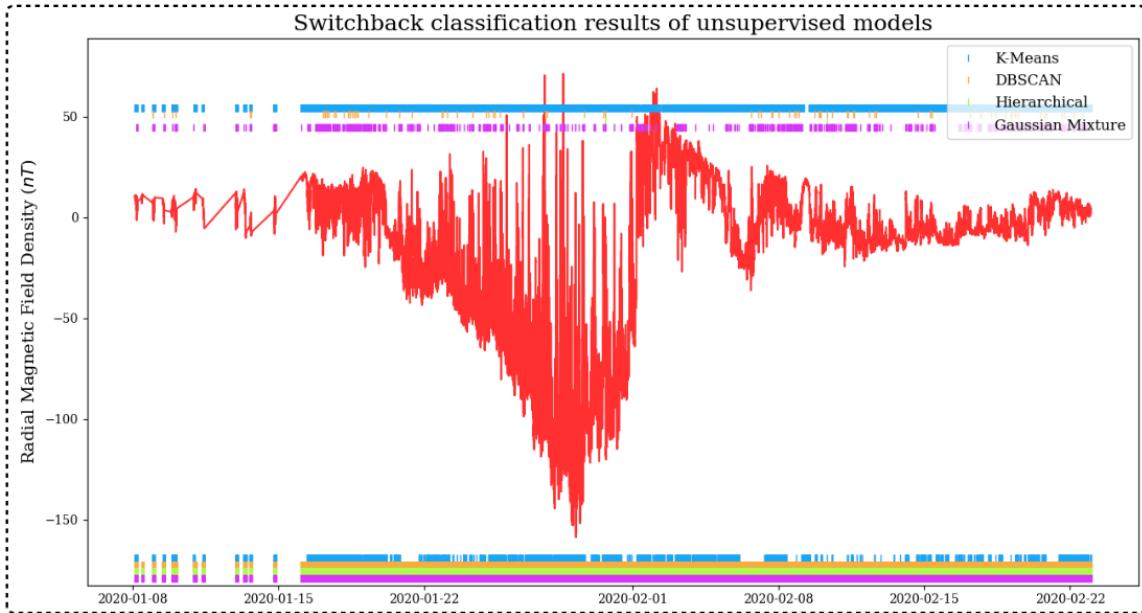
**Encounter 2****Encounter 4**

Figure 20: Plots depicting model predictions overlaid with the radial magnetic field plots (red) for second and fourth encounters respectively. The respective marks at the bottom of the overlay indicate the zero predictions (i.e. no switchback activity). The markings on the top of the graph indicate the predicted switchback signatures.

From the above plots we clearly identify which models show the most reliability and consistency over the separate encounters. Another noteworthy observation that we identify from these plots is how the DBSCAN and GM models seem to adjust to regions of differing switchback density. This dynamic trait indicates that the models are extracting the relevant relational trends from the underlying features. Furthermore, figure 20 also clearly highlights the unreliable performance of the Hierarchical model. These observations support and tie in with what was stated and observed in previous analyses.

## Complete Cross-Model Comparison

Following the summarised results presented in table 3, the specific classification results for encounters two and four could be visualised in figures 21, and 22 respectively:

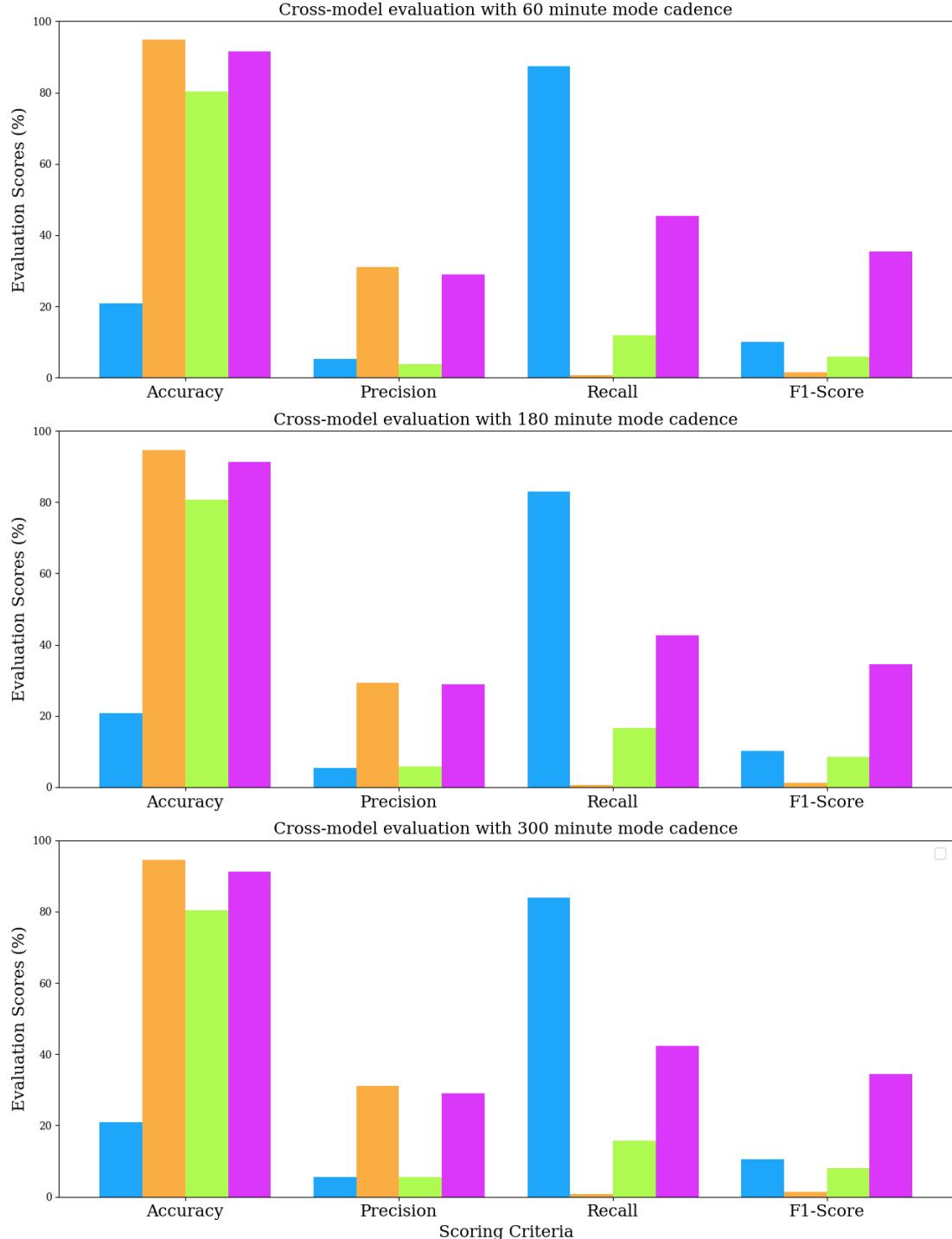


Figure 21: Bar graphs illustrating the cross-model comparison performance of the K-Means (blue), DBSCAN (orange), Hierarchical (green), and Gaussian Mixture (purple) in each classification category for the second orbital encounter of PSP. This evaluation is conducted using three separate statistical mode cadence windows, namely: one-hour (*top*), three-hour (*middle*), and five-hour (*bottom*).

*COMPLETE CROSS-MODEL COMPARISON*

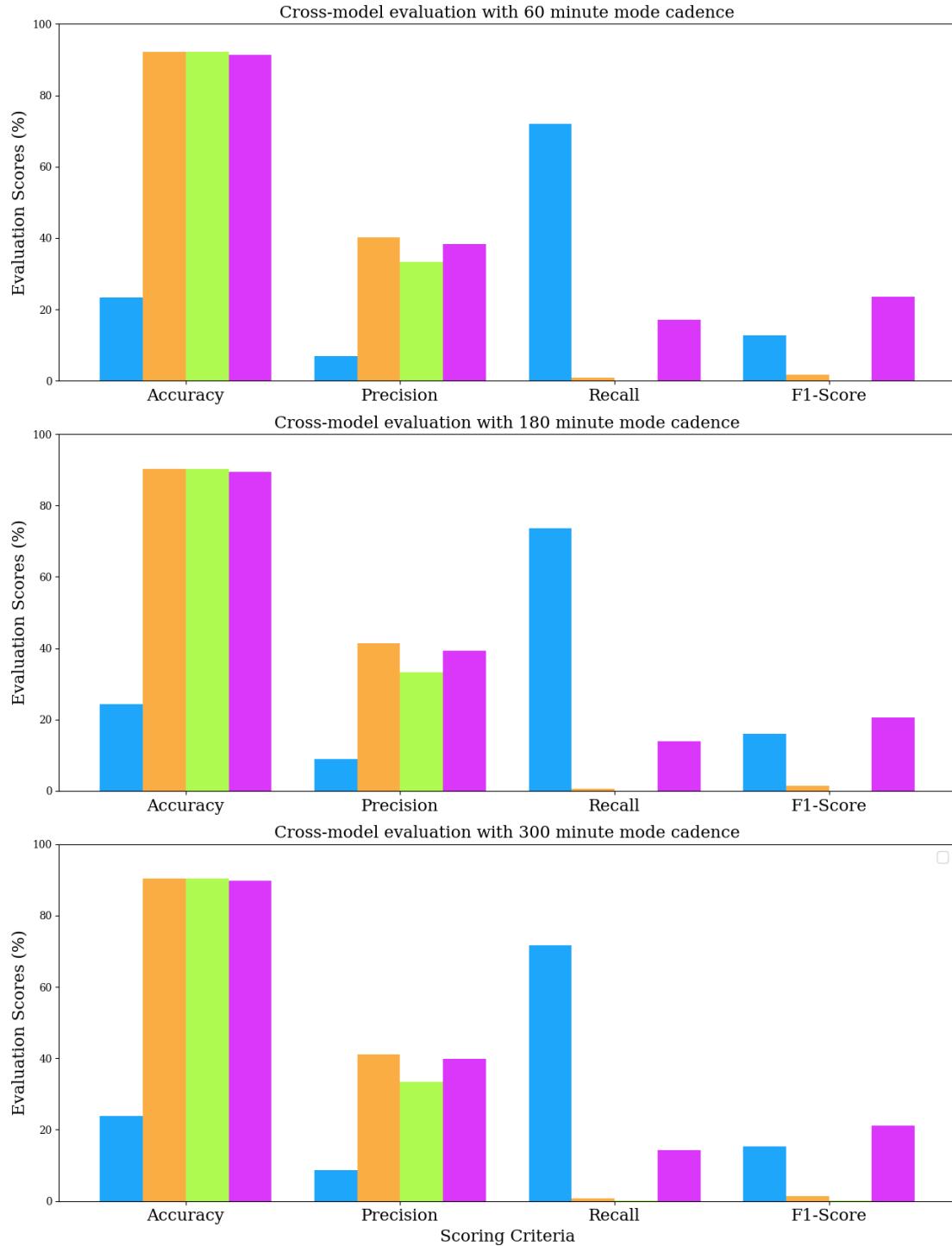


Figure 22: Bar graphs illustrating the cross-model comparison performance of the K-Means (*blue*), DBSCAN (*orange*), Hierarchical (*green*), and Gaussian Mixture (*purple*) in each classification category for the fourth orbital encounter of PSP. This evaluation is conducted using three separate statistical mode windows, namely: one-hour (*top*), three-hour (*middle*), and five-hour (*bottom*).

These graphs represent the individual classification scores for each model over the respective cross-model comparison for each encounter. By comparing these encounters, we gauge the susceptibility of the various models with respect to both the feature data and the changes in encounter parameters. The graphs represented in figures 21, and 22 show significant levels of similarity with that of the results quoted in figure 10. We note that the DBSCAN and GM model possess very consistent and reliable results over all encounters. This consistency is regardless of any anomalies or unexpected observation, indicating that these models produce highly generalised predictions with little feature susceptibility.

Conversely, the trends and statistics recorded across all cross-model comparisons for the K-Means and Hierarchical models support what was observed in other analyses and comparisons. These low prediction statistics and variant results are indicative of poor-quality models with unreliable performance metrics. Furthermore, upon observation of the hierarchical model we see that the same prediction variability that we noted previously is also present in the results of the cross-model comparison. More particularly we identify that the hierarchical model shows variable and contradictory accuracy, precision, and recall statistics. These observations are significant as they point to a model that has no degree of reliability and is unable to appropriately represent the feature space in a manner that is reproducible to the same level of tolerance.

From the critical analysis and review of all our sources of evidence, it is apparent that the measured performance and expectations comply with what was observed over all evaluated encounters. These findings reveal that the observations and conclusions discussed earlier tend to complement (with very little signs of contradiction) what was noted in this detailed analysis. Therefore following the remarks made regarding these detailed results and because these findings further support the concluding statements made regarding the appropriateness of each ML model, we can more firmly guarantee the reliability of our analytical conclusion.