Characterizing Ultraviolet Bursts in a Solar Coronal Hole using Machine Learning Techniques

Jonas Thoen Faber

Thesis submitted for the degree of Master of Science in Astronomy

Institute of Theoretical Astrophysics
University of Oslo

16th May 2022

Abstract

The Interface Region Imaging Spectrograph (IRIS) provides simultaneous imaging and spectral data of the chromosphere and transition region (TR) with great spatial, spectral and temporal resolution. A Solar coronal hole (CH) spectral observation by IRIS is studied for non-Gaussian Si IV profiles that fit the characteristics of UV burst events. Machine learning techniques such as k-means clustering and principal component analysis were applied on the data-set for classifying observed profiles into distinct groups. The least frequent representative profiles (RP), in terms of cluster size, managed to identify profiles that is characteristic for UV bursts events, i.e. complex, non-Gaussian shaped Si IV 1394 Å and Si 1403 Å profiles.

It was found that UV burst like events are ubiquitous in a CH atmosphere but with less probability than in an Active Region (AR) atmosphere. The UV burst events were found to be associated with small-scale jet structures and compact brightenings. The simultaneous study of spectral profiles of Si IV 1394 Å, Si IV 1403 Å, C II 1335 Å, C II 1336 Å, Mg II k 2797 Å, Mg II h 2804 Å, and Mg II triplet 2798.7 Å and 2798.8 Å showed signatures of mass flows and intermittent heating in different layers of the Solar atmosphere that were associated with UV bursts.

The obtained results suggest that UV bursts-like events in a CH condition as observed by IRIS are similar to the features of the TR explosive events observed by previous generation instruments. The IRIS UV bursts in CH are very much alike those observed in the AR, i.e the presence of complex Si IV spectral profiles that deviates strongly from a smooth Gaussian-line. The main difference is that the Ni II λ 1393 Å blend on the blue wing of Si IV 1394 Å are not as frequent and evident in CH spectra due to the low abundance of the cool chromospheric plasma in the CH environment as compared to the AR. Apart from the blend feature, these events are very much alike which conclude a close association between these categories of events.

Acknowledgments

First, I want to thank Nancy Narang for being a fantastic supervisor. You have always been there to help me achieve my scientific goals and provided me with great knowledge in Solar physics during the last two years. You set me out on a journey that have been stellar. I would also like to give a special thanks to Jayant Joshi for introducing great inputs to my thesis and making the results very interesting. Luc Rouppe van der Voort and Tiago Pereira have invested a great amount of their time in my work so of course you have my extraterrestrial gratitude. Thank you RoCS and ITA for making this journey possible and for the vast amount of coffee. I would not have made it without it.

Finally, I would like to thank my girlfriend, friends and family for their great support during my studies. And it would not have been possible to endure the academic journey without my great fellow students. We have been banging our heads on the wall countless number of times, but we finally made it in the end. Thank you all!

Contents

Chapter 1

Introduction

The Universe consists of a countless number of stars and they have existed for billions of years, long before life on Earth began to emerge. Even with the great amount of stars, it is difficult to study them in detail as almost all are too far away from Earth. There is only one star that stands out among all, and that is the Sun. The distance between Earth and Sun is approximately eight light minutes. The second closest star system Alpha Centauri lies approximately four light years away which is a huge difference compared to the Sun. To make an analogy, if you are standing in front of the Institute of Theoretical Astrophysics at Blindern and the Sun is one meter in front of you, Alpha Centauri would be near Kristiansand. Therefore the Sun is soon to be a proxy for studying the complex dynamics happening at the surface of any star.

The Sun has existed for as long as humans have been around and caught its attention to many throughout the history of mankind. Many observational instruments have been used to understand how our neighbouring star brights up and emits large amount of energy. In later years, advanced imaging and spectroscopic instruments have been developed and launched into space to produce observations with unprecedented results. Thanks to the invention of the spectrograph, it is now possible to separate radiation from the Sun into individual components, namely separate wavelength intensities. Each unique wavelength point is measured with its own intensity value. Using a number of wavelength points within some interval, it is possible to observe the dynamics that is prevalent at the Solar atmosphere.

The Sun emits electromagnetic (EM) radiation as a result of nuclear reaction at the core of Sun. Particles in the core are constantly colliding and eventually a fusion process is happening between the collisions. The great abundance of hydrogen is the main source of energy as hydrogen fuse into heavier elements. These reactions occur due to intense heating and high density compression among the particles. The nuclear reaction inflict the hot gas to rise towards the photosphere through the process of energy transport. When the gas reaches the atmosphere, the particles interacts with the gas in the Solar atmosphere and form spectral line signatures throughout the EM spectrum. These spectral lines are unique everywhere depending on the composition of the atmosphere but generally form a signature-shaped profile in the ideal case. As the

Solar atmosphere is highly dynamic everywhere, there are no spectral lines that exactly fits the ideal spectral line. The level of atmospheric dynamics varies but the strongest spectral lines occur due to magnetic field evolution in the Solar atmosphere at different spatial and temporal scales. The magnetic field dynamics is coupled throughout the solar atmosphere, which results in interrelated complex phenomena that is evident at different layers in the atmosphere. The association between such phenomena should be further studied and that will potentially help to increase our knowledge about the mass and energy cycle of the Solar atmosphere.

1.1 Aim & Motivation

Obtaining detailed data is only getting better with the ever advancing technology. With help from the National Aeronautics and Space Administration's (NASA) *Interface Region Imaging Spectrograph* (IRIS) telescope (De Pontieu et al. 2014, 2021), it is possible to study the photosphere, chromosphere, transition region (TR) and corona in great detail compared to its predecessors. IRIS enables observation of small-scale events in far ultraviolet (FUV) and near ultraviolet (NUV) passbands on resolutions that was not possible before. The spatial resolution is 0.33 and 0.4 arcsec in FUV and NUV respectively. The spectral resolution of the spectrograph is 0.026 Å and 0.053 Å in FUV and NUV respectively. The FUV and NUV passbands cover bright spectral lines such as Mg II k 2797 Å and Mg II h 2804 Å in the chrompsphere, C II 1335 Å and 1336 Å in upper chromopshere and Si IV 1394 Å and 1403 Å in TR.

The high resolution spectroscopy provided by IRIS enables observation of small-scale features at the chromosphere and TR of the Solar atmosphere. Features such as explosive events and/or ultraviolet (UV) bursts fall under such categories (Young et al. 2018). These are typically short-lived phenomena that form non-Gaussian spectral lines in the TR with intense brightenings and are regularly observed in active regions (AR) (Huang et al. 2017). As many studies have focused mainly on AR, it is of great interest to study non-AR where the conditions of the atmosphere are not as dynamic, such as quiet Sun (QS) or coronal hole (CH) regions and look for the frequency of similar events. The aim in this thesis is to study spectral observations of a CH environment and identify spectral profiles that are characteristic for events such as explosive events and UV bursts. The findings of such events are further studied by linking the correlation with other phenomena at different layers of the atmosphere. The spectral lines that are of interest are found by machine learning techniques as each observation contains a vast number of data points.

The numerical implementation of algorithms and results are publicly available from `https://github.com/JonasTFab/Master/tree/main/iris/iris_code`

1.1.1 Outline

Chapter 1 gives a brief introduction on the layers of the Solar atmosphere and introduce energetic events of interest in this thesis at different heights. Basic theory on radiative transfer is explained in chapter 2. In chapter 3, the space-born observatory IRIS and

its specification is presented. The potential diagnostics of observation by IRIS are discussed and also the data-set used in the thesis is introduced. Chapter 4 presents numerical solutions by implementing machine learning techniques to classify the entire data-sets. The preparation of the data-set is explained in detail. In chapter 5 the results of applying machine learning techniques on the CH data-set are presented and discussed. The main focus of the results was locating and investigating phenomena related to UV bursts-like events. A brief summary of the findings is found in chapter 6.

1.2 Solar Atmosphere

Physical processes of the Sun vary across its radial distance from its centre. The innermost section being the dense, hot core where nuclear reactions constantly balance out the contraction of the gravitational force. The reactions cause ionised hydrogen to fuse into heavier elements and produce photons that escape the core reaching the radiative zone. Here, photons are absorbed and emitted countless number of times. After some period, the radiation enters the convection zone where the energy is transported outwards through convection. The energy is transported as heat by blobs of plasma into the Solar atmosphere. Due to the vast number of collisions in the Solar interior, the estimated travel time for photons to migrate from the core towards the atmosphere is approximately 170 000 years (Priest 2014).

The photons eventually enter the atmosphere where they are potentially free to traverse towards Earth and can be observed by telescopes sensitive to parts in the EM spectrum. The Solar atmosphere is mainly divided into three categories. The lowest region is called the *photosphere*, followed by the *chromosphere* and finally the *corona*. In between the chromosphere and corona comes the TR where temperature is drastically increased from a few thousand degrees to millions of degrees of kelvin. Different parts of the EM originate from different heights in the atmosphere, where height in Solar physics community is often referred to the layers of the atmosphere. The exact physical heights of these layers are not precisely determined as the Solar atmosphere is highly dynamic and heterogeneous. Assuming the photosphere as the Solar surface, the surface height will not be uniform everywhere as the plasma are in constant motion. As the plasma is heated and cooled by atomic and magnetic processes, it will rise and fall periodically. It is therefore more physically meaningful to refer to the layers of the atmosphere as different processes are evident at different layers.

On a large scale structure, the Solar surface is often distinguished by three different types of regions based on observation. Those events are formally named *active region* (AR), *quiet Sun* (QS) and *coronal hole* (CH). These regions are mainly distinguished by their magnetic field. The strongest magnetic field are evident in an AR environment in sunspots and pores. In a QS, magnetic field are spatially small-scaled with weak mixed polarity. A CH environment are large Solar regions where the magnetic field lines extends to the heliosphere. CH are recognized in coronal images as large dark patches by e.g. the Atmospheric Imaging Assembly (AIA) onboard the Solar Dynamics Observatory (SDO) (Lemen et al. 2012) e.g. at the 193 Å channel. AIA gather imaging

data of the Solar full-disk in the extreme ultraviolet (EUV) passband. QS are typically referred to the static case where the Sun is approximated to a spherical symmetric ball of plasma, even thought this statement is far from reality. The properties of the Sun then mainly depends on the radial component and therefore magnetic interactions are comparatively negligible. AR are superimposed on the QS atmosphere where magnetic fields contributes in a significant manner. They are recognized from Solar-disk images as large, bright loops that origin from the photosphere and extend into the corona. They appear due to magnetic flux from below the photosphere emerge into the corona where plasma follow the magnetic reconnected fields. The large dark patches in a CH are cooler with a less dense corona where the magnetic field is open. In the Solar atmosphere, CHs are one of the longest-lived features and CH in the polar regions can even last for years. The AR, QS and CH therefore draw attention to analysis and comparing due to the present of differences in Solar activities.

1.2.1 Photosphere, Chromosphere, Transition Region & Corona

The photosphere is the first layer the photons enters in the atmosphere. It is a thin layer compared to the entire atmosphere and extends only a few hundreds of kilometers. It is sometimes considered as the surface of the Sun as the atmosphere is opaque below this layer due to the dense composition of materials. Therefore, spectral lines formed in the photosphere are absorption lines. The temperature is relatively low and typically range from 4 000 to 6 000 K.

In the Solar interior where convective cells are originated, plasma reach out to the surface and form a pattern of granules and supergranules in the photosphere. Granules are irregular in shape and have a range of sizes but they all are bright at the centre where hot plasma is flowing upwards. The plasma eventually extends horizontally towards the edge of the granule where cool material falls back towards the surface and form dark boundaries, formally known as *intergranular lanes*. The granules, supergranules and intergranular emits radiation that is observed from photospheric lines.

The chromosphere lies on top of the photosphere and extends a few thousands of kilometers in contrast with the thinner photosphere. The chromosphere is very dynamically complex compared to the other regions and is less dense than the photosphere. Here, only a few very strong spectral lines are observed in the UV domain, e.g. Mg II k 2797 Å and Mg II h 2804 Å. Only the core of Mg II h&k are observed as an emission with a central reversal and the formation of such spectral line occur due to complex interactions between elements in the Solar atmosphere.

Mg II k 2797 Å and Mg II h 2804 Å are strong chromospheric lines that are popular among many observers to recover details of this particular atmospheric region. The Mg II lines is located at NUV domain that covers partially the photosphere and chromosphere. The NUV range from $\approx$ 2000-3900 Å. The wings of the Mg II h&k line mostly contain absorption lines that can be soured from the photosphere region. Other strong emission lines that originates from the upper chromosphere are the C II 1335 Å and C II 1336 Å lines. These are observed in the FUV domain and range from $\approx$ 1200-2000 Å. Using observations in both NUV and FUV can result in information of features that

potentially extends over a range of height.

The TR is a thin region of a few hundred kilometers that lies between the chromosphere and the corona. Here the temperature drastically rises from a few thousands kelvin to millions of kelvin. The extreme increase of heat are typically a result from the complex atmospheric dynamics and interactions of magnetic fields that takes place in the Solar atmosphere. It is not yet fully understood what is the source of the extreme heating, but it draw a lot of attention among the astrophysical community as energy and mass are transported through this layer from lower to higher atmosphere. Studies favour magnetic reconnection as a major part of the heating. Si IV 1394 Å and Si IV 1403 Å are two lines observed from the TR.

The corona is the outermost layer of the Solar atmosphere above the TR and extends millions of kilometres into the vacuum of space. It is where the temperature has increased to millions of degrees of kelvin as a result of the complex atomic processes from regions below. Spectral profiles from the corona are mostly emission lines in the EUV domain due to the low atomic density. Radiation from the EUV that originates from the Solar TR and corona is strongly absorbed by the Earth's atmosphere. Therefore, in order to study this part of the EM spectrum, space-born telescopes that operate outside Earth's atmosphere is required. IRIS will not be used to study spectral lines formed in the corona in this study, though it is able to observe coronal spectras in extreme flare conditions.

1.3 Transition Region Ultraviolet Bursts

Observations of high resolution small-scale events with spatial scales of a few arcseconds in the TR were not possible until 2013 when IRIS was launched into space. IRIS provides simultaneous UV imaging and spectroscopy so that features in both observational form was comparable with great spatial accuracy. The increase of spatial and spectral resolution of IRIS is a factor three and two respectively compared to its predecessors (De Pontieu et al. 2014, 2021; Young et al. 2018) which highly motivates studies of events in the TR. This lead to observations that potentially would be coupled with magnetic events, such as UV bursts, especially in AR. UV bursts are typically observed from short-lived Si IV 1394 Å and Si IV 1403 Å lines and originate due to convergence and cancellation of magnetic flux in opposite polarity where the plasma obtain a motion of > 100 km/s. The temperature of these events are heated in low-chromosphere to temperature that resembles the TR. Hence, UV bursts events typically extends over multiple layers in the atmosphere. Analysis of spectral line diagnostics conclude that UV bursts emits at very high density at the intersection of upper-photosphere and lower-chromosphere. This could potentially be interlinked with AR environments as the plasma is accelerated from the photosphere towards the corona along the strong magnetic field.

Small-scale energetic events in the TR has been given various names during the modern history of Solar physics as they have been observed by multiple instruments and many scientists. Young et al. (2018) wrote a review of similar events that covers

UV spectral lines formed in the interface region, i.e. the chromosphere, TR and corona. Such energetic events observed by IRIS and formed at the interface region has been given a general term, namely UV bursts. The International Space Science Institute (ISSI) team that contributed to this review paper introduced the following properties that define UV bursts. The core brightening are $\leq 2''$, even when sourced by extended structures such as jet, fibrils or loop. The duration varies between tens of second to more than a hour. The intensity was not specifically set but compared to the surrounding UV radiation, it is significantly brighter. The motion of UV burst are typically ≤ 10 km s^{-1}. A conclusion from the mentioned properties, is that UV bursts are typically observed through a sequence of images in the UV passbands as a complex, non-Gaussian TR emission line. Their review among others strongly believe that UV bursts are small-scale events associated with magnetic-reconnection. Ellerman bombs (EB) are signatures of magnetic reconnection in the photosphere. They are generally observed in AR with strong magnetic flux emergence. UV bursts and EB are occasionally related with each other in AR (Tian et al. 2016). Rouppe van der Voort et al. (2016) found evidence of potential EB located at the non-active QS regions away from the more energetic AR. An interesting scientific topic is to investigate whether UV bursts in the TR is related to EB in the QS.

Strong spectral profiles in the interface region are mainly emission lines due to the low density and high temperature. The majority of the UV emission lines are structured with a smooth Gaussian shape and is characteristic for a simple, not so dynamic atmosphere. The condition on a location where UV bursts are present result in spectral Si IV 1394 Å and Si IV 1403 Å lines that differ strongly from the ideal Gaussian-shaped profile. UV bursts can be identified from single spectrograph snapshots as an asymmetric profile that show characteristic features due to sub-resolution atmospheric dynamics. UV bursts can be identified in e.g. TR Si IV lines with multiple peaks or extended wings compared to the mean observed profile. UV bursts are frequently observed in AR atmosphere as TR emission lines are enhanced significantly. The energetic AR sometimes result in observation of UV bursts with strong and wide TR profiles with superimposed weak absorption blend due to low ionisation species. For example, (Young et al. 2018) observed a Ni II absorption line at 1393.33 Å as a blend on the blue wing of Si IV 1394 Å line. The evidence of such blend locates the UV bursts deep in the atmosphere, below the cool canopy of chromospheric fibrils. In a non-active region, the feature of a Ni II λ 1393.33 Å blend are usually very weak or not observable at all.

Explosive events, previously recognised by the name "turbulent events" by Brueckner & Bartoe (1983), was originally recognised by observing C IV λ 1548 Å with asymmetric broadening or enhanced wings on either side of the core. Today they are determined by observation of Si IV lines in the TR by IRIS in non-active regions. These events are observed with little Doppler shift regardless of the position on the Solar disk. Such isotropic behaviour lead to the addition of "explosive" to its name. The spatial extension of an explosive event are typically 1-2 arcsecond as proposed by Brueckner & Bartoe (1983). The extension on both wings suggests the motion of plasma as both upflow and downflow at same location, i.e. at the same pixel. Such observations is due to lack of spatial resolution as the instrument is not capable of resolving mass flows in isolated

directions. Observations of bi-directional jets which are expected to originate from magnetic reconnection is not separated but blended within a single pixel. Even with the poor spatial resolution, such features in addition to enhancement of radiation, match the predictions of magnetic reconnection well, so is therefore thought to be strongly linked with explosive events (Dere et al. 1989). Explosive events share much of the same structural and physical properties of UV bursts such as complex, non-Gaussian Si IV profiles, but are in principle observed at different atmospheric conditions. Therefore, the aim is to search for events that strictly fit the conditions of a UV bursts type of event in a non-AR data-set.

Chapter 2

Radiative Transfer Theory

To observe electromagnetic radiation from the Sun towards an observer, it needs to be sourced by a physical process. In order to understand this process, it is fundamental to introduce radiative transfer. Radiative transfer is a theory of how energy as a form of EM radiation is transported from afar. The radiative energy transportation are unique throughout the electromagnetic spectrum, hence it is possible to observe unique formation of spectral lines. Radiative transfer theory is influenced and explained by contribution from Rutten (2003).

2.1 Intensity

Observers use telescopes to measure the intensity from a source at different wavelength λ or frequency ν. The conversion between wavelength intensity and frequency intensity is $I_\lambda = I_\nu c/\lambda^2$ where c is the speed of light in vacuum. The local energy transportation is given as:

$$dE_\nu(s) = I_\nu(s)\ dA\ dt\ d\nu\ d\Omega \tag{2.1}$$

where dE_ν is the transported energy of frequency ν, I_ν is the monochromatic intensity, dA is area of transmitted energy, dt is the time interval of energy transportation, $d\nu$ is the range of frequency and $d\Omega$ is solid angle. As seen from the equation, the intensity depends on the amount of energy that is transported through some small interval. I_ν is intensity for a single frequency so in order to compute the total intensity, one need to isolate the I_ν term in equation 2.1 and integrate with respect to ν.

Equation 2.1 is a infinitesimal case, i.e. energy transported on a very small scale. This is not the case at the Solar atmosphere as there are layers of gas above the Solar surface. The monochromatic intensity beam will either be added or removed with a certain amount of energy. The addition of energy to the infinitesimal length is formally called *emissivity* and the local contribution of intensity is given by:

$$dI_\nu(s) = j_\nu(s)ds \tag{2.2}$$

where j_ν is the monochromatic emissivity and s is the path length where photons is added to the energy beam. Emissivity occurs e.g. due to interaction between particles as ions and electrons or due to spontaneous decay. The removal of energy is formally called *extinction* and explain how energy is removed from the intensity beam. The definition of local removal of energy per path length is given by:

$$dI_\nu(s) = -\alpha_\nu(s) I_\nu(s) ds \tag{2.3}$$

where α_ν is the monochromatic extinction coefficient. With the expression of addition and removal of local intensity throughout an energy beam from equation 2.2 and 2.3, another expression namely *source function* is defined as:

$$S_\nu = \frac{j_\nu}{\alpha_\nu} \tag{2.4}$$

Both j_ν and α_ν can be a combination of multiple physical combination. Therefore both of these terms are highly difficult to explain analytically as processes in the Solar atmosphere is very complex and dynamic. Still, they are crucial concepts for the atmospheric composition of molecules and understanding of observational intensities of the Solar atmosphere.

2.2 Transport Equation

With the two equation 2.2 and 2.3 concerning how intensity is locally added or removed through a beam, the radiative transport, i.e. *transport equation*, through an infinitesimal interval ds is given as:

$$dI_\nu(s) = I_\nu(s + ds) - I_\nu(s) = j_\nu(s) ds - \alpha_\nu(s) I_\nu(s) ds \tag{2.5}$$

Dividing both sides by $\alpha_\nu ds$ and use equation 2.4, the transport equation becomes on differential form:

$$\frac{dI_\nu}{\alpha_\nu ds} = S_\nu - I_\nu \tag{2.6}$$

The transport equation describe how intensity at some frequency ν evolves through changing layers of some medium. It also assumes that photons do not decay spontaneously. The transport equation is then highly dependent on the composition of atoms in the atmosphere. The photosphere is rather dense compared to the layer above and so the emissivity is dominant which result in the intensity being closely equal to the source function $I_\nu \approx S_\nu$. As the density becomes significantly reduced at greater heights in the atmosphere, addition to the intensity beam is reduced and removal of intensity becomes dominant due to interaction of charged particles in the gas. The source function then becomes negligible. When photons escape the Solar atmosphere and enter the vacuum of space, there is nothing added $j_\nu = 0$ nor removed $\alpha_\nu = 1$ to the beam and so the intensity becomes constant along the path.

2.3 Optical Depth

Photons that is sourced from the Sun does not always traverse to the observer. Especially photons at the Solar core due to the dense composition of molecules. The layers above are blocking the emitted photons within the core making it impossible to observe from Earth. This occurs due to the extinction coefficient becomes too large so that the transported energy of a beam s completely removed. To express how extinction coefficient affect the radiative beam is by introducing *optical path length* for some small interval ds:

$$d\tau_\nu(s) = \alpha_\nu(s)ds \tag{2.7}$$

Integrating the equation of a medium with thickness D, e.g. over the Solar atmosphere, it can be express as optical thickness as following:

$$\tau_\nu(D) = \int_0^D \alpha_\nu(s)ds \tag{2.8}$$

Equation 2.7 can be plugged in to equation 2.6 and by applying the initial condition with no emissivity ($j_\nu = 0$) and integrating from 0 to a distance D, the the transport equation becomes:

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu \tag{2.9}$$

$$\Rightarrow \int_0^D I_\nu^{-1} dI_\nu = -\int_0^D d\tau_\nu \tag{2.10}$$

where optical depth from equation 2.8 at distance 0 imply that the beam has not yet propagated through any medium, therefore $\tau_\nu(0) = 0$:

$$\Rightarrow \ln I_\nu(D) - \ln I_\nu(0) = -\tau_\nu(D) \tag{2.11}$$

$$\Rightarrow I_\nu(D) = I_\nu(0)e^{-\tau_\nu(D)} \tag{2.12}$$

The last equation indicates that if the optical depth at a distance D is 0, then the intensity will be constant which imply no matter in between the source and the observer. Optical unity occur when $\tau_\nu(D) = 1$ whereas the intensity at a distance D has decreased by a factor of $1/e$. The medium is considered optically thin for frequency ν if $\tau_\nu(D) < 1$ and optically thick if $\tau_\nu(D) > 1$.

As emissivity is present in the Solar atmosphere, equation 2.12 is a special case that is never actually true when considering the atmosphere. The intensity highly depends on the optical thickness including both extinction and emissivity, so that the transport equation 2.6:

$$\frac{dI_\nu}{d\tau_\nu} = S_\nu - I_\nu \tag{2.13}$$

The intensity thus become on integral form:

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} S_\nu(t_\nu)e^{-(\tau_\nu - t_\nu)}dt_\nu \tag{2.14}$$

As seen from the equation above, the observed intensity might depend on intensity from a source, the removal of energy due to extinction for infinitesimal intervals and also addition of energy to the beam.

2.4 Formation of Spectral Lines

The formation of spectral lines are very complex due to many factors that varies across the Solar layers in terms of velocity, density, temperature, collisional rates and so on. In addition to variation of intensity from emission and extinction, the background continuum is also affecting the intensity at some frequency ν. Hence, the source function 2.4 include contribution from two different processes:

$$S_\nu = \frac{j_\nu^c + j_\nu^l}{\alpha_\nu^c + \alpha_\nu^l} = \frac{S_\nu^c + \eta_\nu S_\nu^l}{1 + \eta_\nu} \tag{2.15}$$

where superscript c denote continuum and l denote line. $\eta_\nu = \alpha_\nu^l/\alpha_\nu^c$ is the line-to-continuum extinction ratio. In order for any emission lines to be evident, it follows that j_ν^l is increased due to atomic and relativistic effects. Such effects are mainly bound-bound transitions or bound-free transitions. Bound-bound transitions occurs when an ion excites from an energy level i to a higher energy level j. This occurs as free electron collide with the ion and transport some of its energy to the ion. The ion is then in a state with higher potential energy and will eventually deexcite to a lower energy level and emit a photon with an energy according to the transition level of the ion.

Atomic excitation in the upper atmosphere are dominated due to collision between ions and free electrons. Other but less important collision are also possible, such as proton-ion collision, photo-excitation, radiative scattering, and ionization and recombination into excited levels (Phillips et al. 2008). The ion-electron collision events are quite frequent allowing spontaneous decay and collisional de-excitations to happen. Whenever ions decay, the excess energy between the states are converted to photon energy. The estimated energy of a photon is found by the Planck-Einstein relation:

$$E_{i\to j}(\lambda) = \frac{hc}{\lambda} \tag{2.16}$$

where h is Planck constant, c is the speed of light in vacuum and $E_{i\to j}$ is the released energy when an ion spontaneously decay from state i to state j. Ionization processes occurs whenever a bound electron is freed from its ion and equivalently, a recombination occurs whenever a free electron is captured by an ion. The ionization/recombination are typically opposite-paired in three distinct processes, as stated by Phillips et al. (2008). The first being photoionization and radiative recombination, i.e. interaction between an ion an ambient radiation field. The second process involves collision between an

ion and two free electrons. The third process involves a collision between ions above the ionization threshold with a free electron. It is the interplay between ionization and recombination that determine the abundance of an element. The spectral lines observed by IRIS are highly dependent on the fractional abundance of particular ions.

2.4.1 Optically Thick Lines

Optically thick lines are formed under the condition where equation 2.8 is greater than a unit value, i.e. $\tau_\nu > 1$. Plasma below this height becomes gradually opaque as $\tau_\nu \gg 1$ due to the thick layer of atmosphere between the source and the observer. Typical example of optically thick lines are lines formed at the photospheric region due to the high density of plasma. From equation 2.14, the first term goes to zero. The Eddington-Barbier relation state that the outgoing intensity is equal to source function if the source function is linear dependent of optical depth. This assumption assume there is transparency above the region, i.e. $\tau_\nu = 0$ above this region. As the optical thick region often are followed by a optically thin region, this assumption remain relatively accurate for outgoing intensity normal to the surface. If the outgoing intensity is at an angle of the surface normal, the beam trajects through a thicker layer of atmosphere and the Eddington-barbier relation breaks. This assumption is quite accurate at the photosphere and therefore the intensity can be directly expressed with the source function:

$$I_\nu(\tau_\nu = 0) \approx S_\nu \tag{2.17}$$

The photosphere is typically the lowest layer that telescopes are able to gather data from and it is here the background continuum is observed. The continuum is optically thick and therefore follows the relation 2.17 quite well. Emission and absorption lines are increased or decreased in intensity at some frequency relative to the background continuum depending on the strength of the emissivity or extinction coefficient.

2.4.2 Optically Thin Lines

Optically thin lines radiates through a medium without being fully absorbed or scattered. The medium in this case is the Solar atmosphere as there is only vacuum between IRIS and the Sun. The optical thin regime result in equation 2.8 less than a unit value, i.e. $\tau_\nu < 1$. In between the origin of the emitted light and towards vacuum, the atmospheric density is relatively low as a result of the transparency. With low density follows very little absorption and therefore only emission is observed for particular frequency in the UV domain.

Emission lines in the UV domain are most importantly formed due to ions deexcite from a higher energy state to a lower energy state. As an ion deexcite to a lower state and due to conservation laws, the excess energy is transformed into photons with frequency that corresponds to the change in the ions energy state. The intensity of any emission line depends typically on the amount of emitting ions from one to another energy state for a particular element. The Solar plasma processes that contribute to

photon energy are formally called ionization/recombination for change of ionization states and *excitation/de-excitation* for populating/depopulating ionization states. In the upper atmosphere where optical thin regime dominates, the latter processes are most common.

The intensity at frequency ν that reach the observer are a summation of energy contribution of emitting photons along the line-of-sight (LOS) beam (equation 2.2). Ignoring excitation, using the frequency to wavelength conversion and reformulating equation 2.14, the total intensity for a given wavelength in an optical thin regime can be expressed as:

$$I(\lambda) = \frac{1}{4\pi d^2} \int_{\Delta V} E_{i\to j} A_{j\to i} N(X_j^{+z}) dV \tag{2.18}$$

where d is the distance between the emitted region and the observer (one astronomical unit (AU) from Sun to Earth), $A_{j\to i}$ is the Einstein coefficient that describes the probability of spontaneous excitation from state $j \to i$ and $N(X_j^{+z})$ is number density of ion X_j^{+z} in the jth excited state. z is the ionization level of element X. Equation 2.18 integrates over a small volume of plasma that have the potential of emitting photons. The ion number density is expressed as following:

$$N(X_j^{+z}) = \frac{N(X_j^{+z})}{N(X^{+z})} \frac{N(X^{+z})}{N(X)} \frac{N(X)}{N(H)} \frac{N(H)}{N_e} N_e \tag{2.19}$$

where the first four factors represent relative level population, relative ionization population, abundance of element X relative to hydrogen and abundance of hydrogen relative to free electrons in the plasma respectively. N_e is the density of free electrons. Merging equation 2.16 and 2.19 is used to define the *contribution function* as following:

$$G(T, N_e, \lambda) = \frac{N(X_j^{+z})}{N(X^{+z})} \frac{N(X^{+z})}{N(X)} \frac{N(X)}{N(H)} \frac{N(H)}{N_e} \frac{A_{j\to i}}{N_e} E_{i\to j}(\lambda) \tag{2.20}$$

With further modification of equation 2.18 where the contribution function is applied, it becomes:

$$I(\lambda) = \frac{1}{4\pi d^2} \int_{\Delta V} G(T, N_e, \lambda) A_{j\to i} N_e^2 dV \tag{2.21}$$

The emitted light is then expressed by the contribution function of elements, the probability of spontaneous decay of element X and electron density over the LOS.

Chapter 3

Observations

3.1 Interface Region Imaging Spectrograph

On 27 June 2013, a Pegasus-XL rocket was launched into space with in its cargo bay the *Interface Region Imaging Spectrograph* (De Pontieu et al. 2014). The telescope was placed in a Sun-synchronous, low-Earth orbit at a perigee of 620 km and apogee at 670 km. The inclination was optimized at 97.9 degree such that eclipse-free observations were possible from early February to late October the first few years. The telescope is shielded so that the data gained from observation is sufficiently protected against the increased radiation over the northern and southern high-latitude zones (HLZs). The energetic field at the South Atlantic Anomaly (SAA) leads to high level of noise so that the data obtained by IRIS at this orbital location are severely affected.

The spacecraft contains a 19 cm Cassegrain UV telescope that transit the Solar radiation towards a spectrograph. First the radiation enters the telescope and hit the primary mirror. The front of the primary mirror is coated with dielectric material such that UV light is reflected to a secondary mirror. The secondary mirror then focus the reflected light towards the spectograph where spectral data is assembled (Podgorski et al. 2012). The spectrograph is slit-based, meaning that it covers a 0.33×175 arcsec^2 range of data for every exposure. There are two observational modes available by IRIS, which are the raster mode and sit-and-stare mode. The raster mode observe over a field of view (FOV) while sit-and-stare focus the slit on a single physical location on the Solar atmosphere over multiple exposures. A raster observation use the slit by moving step-wise perpendicular to the slit until an extended spatial region is fully covered. The step size can be large to cover a wide raster or narrow for coverage of fast dynamic evolution. The raster extend a maximum FOV of 130×175 arcsec^2. IRIS can take a single raster observation or multiple depending on the preference for the scientific study. The spectrograph is able to observe the spectra in FUV passbands from 1332–1358 Å and 1389–1407 Å. These passbands are named FUV 1 and FUV 2 respectively for reference. FUV 1 passband covers the C II 1334 Å and C II 1335 Å lines formed in the upper chromosphere and the FUV 2 passband cover Si IV 1394 Å and Si IV 1403 Å lines formed in the TR. IRIS also observe in the NUV passband that ranges 2783-2834

Å which include the strong Mg II k 2797 Å and Mg II h 2804 Å lines. These lines are formed in the chromosphere. The specification of the FUV and NUV spectrograph onboard IRIS can be seen in table 3.1.

Table 3.1: Characteristics of the IRIS spectrograph instruments.

	FUV	NUV
FOV slit	0.33×175 arcsec^2	0.4×175 arcsec^2
Spatial pixel scale	0.167 arcsec	0.167 arcsec
Spatial resolution	0.33 arcsec	0.4 arcsec
Spectral pixel scale	12.8 mÅ	25.6 mÅ
Spectral resolution	26 mÅ	53 mÅ

IRIS is uniquely installed with the slit-jaw image (SJI) instrument that allow observers to obtain regional context to the spectral data. The SJI add the position of the spectrograph slit to the images so that it is possible to a see nearby features in different filters. The most common filters are SJI 1330 where the upper chromospheric C II lines are formed, SJI 1400 that covers the Si IV lines, SJI 2797 that is centered on the Mg II k core and SJI 2832 that cover the Mg II h wing. The SJI images enable study of neighbouring events close to the slit that potentially could be correlated with events from spectral data.

3.2 IRIS Spectral Diagnostics

Some of the strongest optically thin lines observed by IRIS are the Si IV 1394 Å and 1403 Å lines in the FUV. They are very popular among the community as it produce data from the TR. Occasionally, the Si IV doublet fall under the optically thick regime, i.e. if the intensity ratio 1394 Å / 1403 Å is different from 2 (Peter et al. 2014). The ratio can also be used for density diagnostics. Most of the emission lines formed in optically thin regimes are similar to a Gaussian-profile. Parameters of observations such as Doppler shift, line width and intensity can produce results such as motion at LOS, non-thermal motion and amount of emitting plasma respectively. If the observed line is offset with respect to a rest wavelength, i.e. the core of an observed line is not placed at 0 km/s, it is Doppler shifted toward blue or red spectrum. Such observation inform whether the plasma is moving towards or away from the observer and possibly at what velocity.

Two of the strongest optically thick lines observed by IRIS are the single ionized C II and Mg II. The prior ion allow three transitions producing a multiplet of 1334.54 Å, 1335.66 Å and 1335.71 Å lines (Rathore & Carlsson 2015). 1335.66 Å line is blended by

the 1335.71 Å line meaning that only two C II lines are truly observed. The strongest allowed Mg II lines are the Mg II k 2795.53 Å, Mg II h 2802.70 Å and the Mg II triplet lines (2790.78 Å, 2797.93 Å, 2798.00 Å) lines (Leenaarts et al. 2013a).

C II lines are observed for diagnostics of atmospheric conditions at the upper chromosphere or lower TR. Leenaarts et al. (2013b) found through modelling that asymmetry between peaks of Mg II lines give information about chromospheric velocity gradient. They concluded that stronger blue peak imply a downflow of the plasma. A stronger red peak yields upflow of the plasma. C II lines is closely resembled with Mg II lines implying that such diagnostic could be implemented for both lines. If double peaks are observed among the C II lines, it is possible to estimate the velocity gradient based on asymmetry of the peaks. Another motion diagnostic is the Doppler shift of the core that is found relative to the rest wavelength of the C II lines in vacuum. Blue Doppler shift indicate a downflow while red Doppler shift indicate upflow of the plasma.

The C II line widths are well correlated with non-thermal velocities as tested through numerical modelling by Rathore et al. (2015a). The correlation are approximately linearly which imply a greater non-thermal velocity at the formation height where the widths of the observed profile are broader. In a QS simulation by Rathore et al. (2015a), the average profile at network and internetwork had double peaks while the average profile had single peak in umbra, penumbra and plage. All the average profile were slightly red-shifted with the vast majority 1335 Å / 1334 Å ratio being lower than 1.8 implying almost exclusively an optical thick regime. The latter was supported by Rathore et al. (2015b) in later observations by IRIS of QS atmosphere.

The Mg II h&k lines are prominent and observed from IRIS in the NUV domain and the lines are very popular among the community to study the Solar chromosphere. The Mg II h&k lines are quite easily recognized with their clear emission with central reversal surrounded by the slowly increasing continuum. Both lines are given abbreviations of features close to the line centers. k_{1R} (h_{1R}) and k_{1B} (h_{1B}) refers to wing minimum of Mg II k (Mg II h) on red and blue side respectively. k_{2R} (h_{2R}) and k_{2B} (h_{2B}) refers to emission peaks of Mg II k (Mg II h) on red and blue side respectively. Finally, k_3 (h_3) refers to the central depression. The continuum contain many absorption lines that provide potential diagnostics of the photosphere.

3.3 IRIS Level 2 Data

The raw data observed by IRIS are initially transported from the Mission Operation Center at Ames Research Center to the Joint Science Operations Center (JSOC) at Stanford University and Lockheed Martin Solar and Astrophysics Laboratory (LMSAL) (De Pontieu et al. 2014). Here the data is archived with a backup at LMSAL site before converted to Level 0 data. Level 0 IRIS data are raw images with housekeeping and overscans that is depacketized. The data is then converted to Level 1 where the images are reoriented so that the top are closer towards north and left to right indicate increasing position in units. Headers are added to inform the user about temperature, roll and pointing of IRIS. Then the data is converted to Level 1.5 which are considered

as the lowest scientifically data of usefulness. More detailed information is added to the headers and additional calibration procedures are applied to it, such as dark-current offsets removal, flat-field corrections and geometrical corrections.

Eventually, Level 1.5 data is converted to Level 2 data that is being used in the current study. The data from Level 1.5 is grouped into logical units, i.e. spectral raster and SJI that is stored in seperate Flexible Image Transport System (FITS) file. While the rasters are localized within the same FITS file, the different passbands are separately allocated, enabling easy access to each imaging passband. The number of passband depends on the type of observation which is specified by the 10-digit observational identification (OBSID) integer that is provided for all IRIS observations.

For a single wide raster, a Level 2 IRIS data is typically structured as a $m \times n \times h$ matrix $\mathbf{X}$ where m and n is the number of horizontal and vertical pixels in the raster respectively and h is the number of wavelength points. When obtaining the observational data, the correct units in all three dimension are not included. In order to extract the units, the user must look into the header and then correctly compute the units. The header contain details on the FOV in x and y direction including the central position of the raster and SJI. The distance between pixels in x and y direction are equidistant. With this information, the spatial coordinates can be appropriately computed. Extracting wavelength values can be obtained using the *World Coordinate System* (WCS) procedure that are applicable on FITS files that are widely used among the astrophysical community. Greisen & Calabretta (2002) describe the procedure in sequential steps where initially the pixel coordinates is converted to intermediate pixel coordinates through matrix operations. It is then rescaled to physical units into intermediate world coordinates and finally converted to real world coordinates through proper configurations using information from the FITS header.

Table 3.2: Specifications of various spectral passbands obtained from the IRIS FITS header from the coronal hole observation as explained in section 3.3.1. The vacuum rest wavelength of C II lines and Si IV lines are taken from Peter et al. (2014). Rest wavelengths of Mg II h&k and Mg II triplet lines are taken from Kerr et al. (2015).

Spectral passband	Spectral range	Rest wavelength (λ_{rest})
C II 1336	1332.72 - 1337.67 Å	1335.71 Å
C II 1335	-	1334.54 Å
Si IV 1394	1390.92 - 1396.26 Å	1393.76 Å
Si IV 1403	1398.64 - 1406.41 Å	1402.77 Å
Mg II k 2796	2792.93 - 2806.58 Å	2796.34 Å
Mg II h 2803	-	2803.52 Å
Mg II triplet	-	2798.75 Å 2798.82 Å

Wavelength is converted to Doppler shift so that potential mass flow can be diagnosed properly. To be able to compare shifts with other observations, a reference wavelength is required. In this study, the reference wavelengths are rest wavelengths in vacuum and are given in table 3.2. The rest wavelength of C II 1335 Å, C II 1336 Å, Si IV 1394 Å and Si IV 1403 Å lines are extracted from Peter et al. (2014). The Mg II k 2797 Å, Mg II h 2804 h Å and Mg II triplet between the h&k lines are extracted from Kerr et al. (2015). The the Doppler shift is computed with the appropriate rest wavelength from the table using the following equation:

$$\Delta\nu = \frac{\lambda - \lambda_{\text{rest}}}{\lambda_{\text{rest}}} c \tag{3.1}$$

where λ is an arbitrarily wavelength point and c is the speed of light in vacuum. The units are given in km/s and is a convenient conversion to set specific boundaries of any profile.

3.3.1 The Coronal Hole Observation

The IRIS observation used in this study is of a CH taken during the early stages of the IRIS mission in october 2013. The initial exposure started 9 October at 23:26:07 UT and ended 10 October at 02:56:53 UT. The unique identification of the data-set is given by its OBSID, i.e. 20131009_232607_3820012146 and is described as a 400-step very large dense raster. The center coordinates of the raster and SJI were $(x, y) = (511, 296)$

arcsec. The observation contain a single raster that span a FOV of 141×182 arcsec^2 area. The y-dimension of the raster refer to the length of the spectrograph slit and the x-dimension covers the full stepping length of the slit, i.e. 400 steps of a 0.33 arcsec interval. Each slit has a step cadence of 31.6 seconds where 30 seconds was dedicated to the exposure time. The SJI FOV covers a $167'' \times 174''$ area in four wavelength passbands, i.e. 1330 Å, 1400 Å, 2796 Å and 2832 Å. The information about the observational properties are encoded in the last number sequence in the data-set identifier OBSID, which in this study is 3820012146. Even with the SJI data disposable in the using data-set, this study is focusing on the spectral raster data only.

Figure 3.1 visualise the raster FOV map, displaying the scaled total intensity of Si IV 1394 Å line. The initial slit exposure is at the left-most column of the map while the right-most column is the final exposure of the raster sequence ≈ 3.5 hours later.

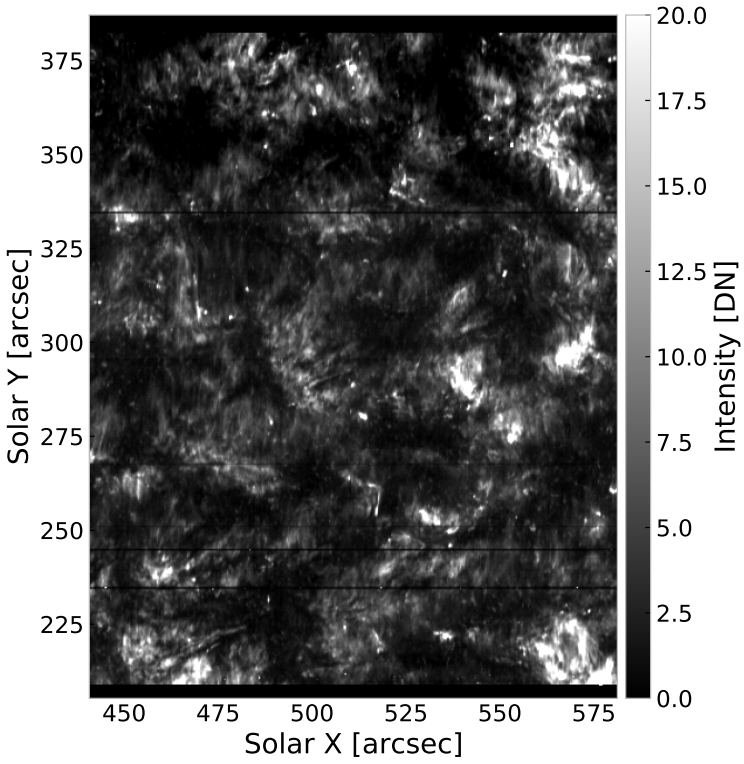

Figure 3.1: Raster map of the coronal hole observed by IRIS on 9 October 2013. The map show scaled total intensity of the Si IV 1394 Å line. The intensity is in detector units, i.e. in units of Data-Number [DN]. The initial slit exposure is at the left-most column.

Chapter 4

Methods

4.1 Data Preparation

The IRIS telescope is a state-of-the-art observatory that produce high quality data in the FUV and NUV passbands. Even with the extraordinary observation with great success on its mission, factors that affect the end-data typically as noise is inevitable. The level of noise varies and ever so often, the noise extend to such degree that it should not be used scientifically. Therefore the data should be appropriately prepared before further analysis. Anomalies such as spikes, that are due to random cosmic ray hits on the CCD, should be properly taken cared of. This section focus on such implementation of algorithms with the goal of reducing the risk of data being ill analysed when machine learning techniques is applied later.

In addition to data spikes, certain pixels in the data-set are of no interest and should be disregarded. These pixels typically falls at the edges of the slit where no observations are made, such as the broad dark stripes visible at the upper and lower edge in figure 3.1. Few horizontal thin dark stripes between the edges indicate location of pixels not able to observe real data. Two of the dark stripes are purposely marked and they are named fiducials that blocks light from entering the spectrograph. The purpose of the fiducials is for preparation of level 2 IRIS data, where different channels will be properly aligned.

4.1.1 Interquartile Range

In order to remove bad pixels from the observational data, a statistical approach called *Interquartile Range* (IQR) method is implemented. The structure of the IQR algorithm is fairly simple and can be divided into four steps. Assume the data-set $\mathbf{X}$ as described in section 3.3. For each array of some data, e.g. each pixel $\mathbf{x}_{i,j}$ where $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$ in $\mathbf{X}$, do the following:

1. Compute first (Q1) and third (Q3) quartile of the array $\mathbf{x}_{i,j}$.
2. Compute $IQR = Q_3 - Q_1$.

3. Define the upper and lower limits as $IQR_{\text{upper}} = Q_3 + \alpha \cdot IQR$ and $IQR_{\text{lower}} = Q_1 - \alpha \cdot IQR$ where α is an adjustable parameter to fine tune the limits.

4. Mask or replace data points that lies outside the IQR limits.

The advantage of implementing IQR is that outliers far from the average data point is neglected and therefore does not affect the range of the IQR limits. Only 50 % of the mid-data points are considered. Data points that fall outside the IQR range will be adjusted accordingly. The procedure is versatile so that it can be applied to the entirety or part of any array. The IQR procedure will be implemented as a masking and despiking routine as explained in the following sections.

4.1.2 Masking Pixels

It is evident from figure 3.1 that there are pixels that separates from the remaining of the raster. From the figure, these appear as dark horizontal regions at the top and lower edges and also a few thin dark horizontal stripes in between are evident. The anomalies are present for all wavelengths in the data-set. Therefore, the data should be masked based on this paradigm as true or false by some threshold. True imply the pixel is suitable for further analysis, false means otherwise.

After inspection of the data 3.3.1, and due to calibration of IRIS Level 2 data, some of the wavelength points hold negative data values. In particular a significant amount of the points at the wings of C II and Si IV profiles. This imply that some negative value must be set as minimum. Further inspection suggest that a minimum value of -10 is feasible and should clear out the more extreme negative values.

Assume the data-set $\mathbf{X}$ as described in section 3.3. Another $m \times n$ matrix $\mathbf{M}$ with integer entries is defined that represents the masking of the original data-set. Iterate through the wavelength points $l = 1, 2, ..., h$ so that for each iteration, the reduced matrix $\mathbf{X}_l$ is evaluated. Whenever a pixel value at the wavelength point l is less than the minimum threshold -10, the corresponding pixel value in $\mathbf{M}$ is raised by 1. The procedure is repeated until the wavelength axis is cycled. This method does not take care of the few thin darker rows that are visible from figure 3.1. These values are not necessarily negative but still fainter than the average row. They are darker possibly because the sensitivity of that pixel has comparatively deteriorated. In order to mask these rows as false, iterate through the wavelength points h as above. Compute the row mean in $\mathbf{X}_l$, i.e. all pixels perpendicular to the slit, to form a mean sample array $\overline{\mathbf{x}}_l$. $\overline{\mathbf{x}}_l$ has the same length as the number of measured points in the spectrograph, or n in this case. Then the IQR procedure as described in section 4.1.1 is implemented to compute IQR of $\overline{\mathbf{x}}_l$. The IQR is subtracted from the mean of $\overline{\mathbf{x}}_l$ by a factor α_{low} and α_{high} to form an upper and lower limit respectively. If any value in $\overline{\mathbf{x}}_l$ falls between these limits, the corresponding row in $\mathbf{M}$ is raised by 1.

An example of the masking routine of darker thin lines perpendicular to the slit is visualised in figure 4.1. The darker rows in the raster for this particular wavelength computes a mean value that fall within the IQR limits as red lines and then raise the row value in $\mathbf{M}$ by 1. α_{low} and α_{high} were optimized to be 0.5 and 5.0 respectively.

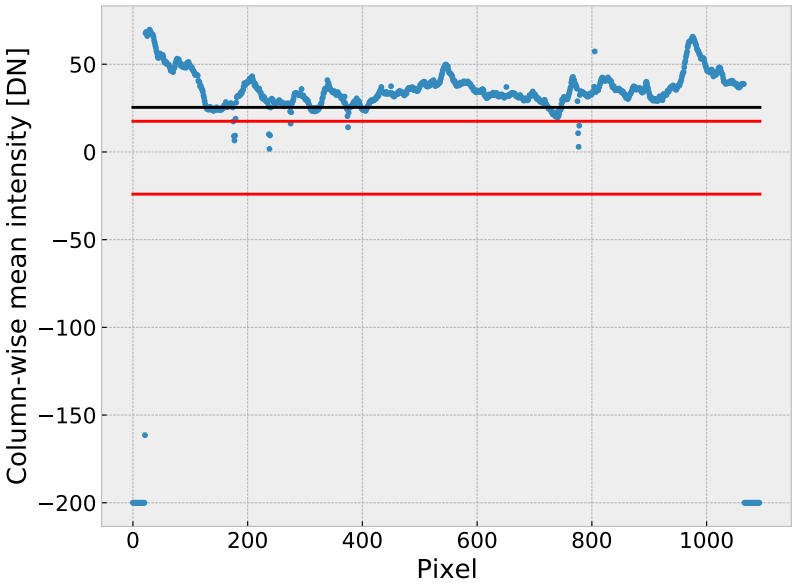

Figure 4.1: Row-wise mean $\overline{\mathbf{x}}_l$ of the raster at $\lambda_l = 1393.753$ Å shown as blue dots. The black solid line show the mean of $\overline{\mathbf{x}}_l$. The red upper and lower solid lines show the mean of $\overline{\mathbf{x}}_l$ subtracted by 0.5 and 5.0 times IQR respectively. Note that y-axis is row-wise mean and not column-wise mean.

Note that points in $\overline{\mathbf{x}}_l \approx -200$ does not fall within this limitations. Such values should be excluded as they will affect the row values in $\mathbf{M}$ twice.

After the iterative routines are completed, the minimum value M_{min} in $\mathbf{M}$ is found to ensure variation of different data-sets does not affect the final threshold. The final threshold that defines a pixel as true or false becomes:

$$T = M_{\text{min}} + t(h - M_{\text{min}}) = M_{\text{min}}(1 - t) + th \tag{4.1}$$

where T is the final threshold and $t \in [0, 1]$ is a adjustable factor. Note that all pixels will be masked true or false if $t = 1$ or $t = 0$ respectively. Finally, $\mathbf{M}$ is redefined as a boolean matrix and the elements are given the following statement using equation 4.1:

$$M_{i,j} = \begin{cases} \text{False} & \text{if } M_{i,j} > T, \\ \text{True} & \text{else} \end{cases} \tag{4.2}$$

where $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. The masked matrix is returned from a function and will be used to remove anomalies from the original data-set. The data-set then needs to be despiked properly before it is applicable for further analysis without interference from poor pixels.

4.1.3 Despike Profiles

It is evident that the spectrograph tends to gain abrupt high intensity data points that separates greatly from the average data, known as spikes. Such features are typically many orders of magnitude greater in value than its neighbouring wavelength points, e.g. $X_{i,j,l-1} \ll X_{i,j,l} \gg X_{i,j,l+1}$. These features fall beyond the third quartile and imply implementation of the IQR method as explained in section 4.1.1. The data points being masked from the IQR routine will be given new values to simulate a realistic profile instead of removing them. An assumption is that the replaced data value $X_{i,j,l}$ should not deviate much from the values of the neighbouring points given the high-spectral resolution, i.e. $X_{i,j,l-1} \approx X_{i,j,l} \approx X_{i,j,l+1}$. As noise is ubiquitous in the data-set, a more accurate replacing estimation of $X_{i,j,l}$ is to consider a fixed number of nearest points. The spike replacement is therefore computed as the average of n_{filter} nearest points around the wavelength point with spike feature. The despiking were implemented as a two-step routine which is referred to as **segment despiking** and **scanning despiking**. The implementations are explained in the following sections. The despiking routines were not applied to the Mg II h&k lines and Mg II triplet line. The reason is that these lines are generally much stronger and therefore spikes are much rarer for these spectral lines.

Segment Despiking

An important issue when despiking is to distinguish the profile core and wings as the shapes are significantly non-identical. Assuming the average Si IV profile is Gaussian, it is clear that the core is more extended in intensity when traversing through wavelength axis. Using the entire profile to compute IQR will result in cases where the upper limit do not exceed the peak of profile cores. Hence IQR should be computed in segments to avoid bad computation of upper and lower limits. This statement leads to the first despiking routine and is called the *segment despiking routine*. The algorithm segregates any pixel $\mathbf{x}_{i,j}$ from $\mathbf{X}$ into segments and then perform IQR analysis on each separately. As the array $\mathbf{x}_{i,j}$ will present a single observed line profile, the array is segmented into a core, red wing and blue wing. The wing segments will be aggregated.

When calling the segment despiking function, the user needs to define the width of the core segment $w_{\text{core}} \in [0, 1]$. If $w_{\text{core}} = 0$ then no core is assumed while $w_{\text{core}} = 1$ imply the entire array as a core. The optimized core width factor of both C II lines was found to be $w_{\text{core, C II}} = 0.38$. For Si IV lines, the core segment was optimized to $w_{\text{core, Si IV}} = 0.45$. The core segment and wing segments are given their adjustable factor α_{core} and α_{wing} to tune the IQR limits appropriately and they were fixed equally for all spectral passbands. The optimized parameters was found to be $\alpha_{\text{core}} = 10$ and $\alpha_{\text{wing}} = 18$. All these parameters had to be tuned through trial and error due to the shape variation of the cores. The operation of the segment despiking routine is shown as an example in figure 4.2. The algorithm marks several points as a spike and then replace the intensity values by the mean of the 7 red and blue non-spiked neighbouring points. The brown dots show the replacing points of the spikes.

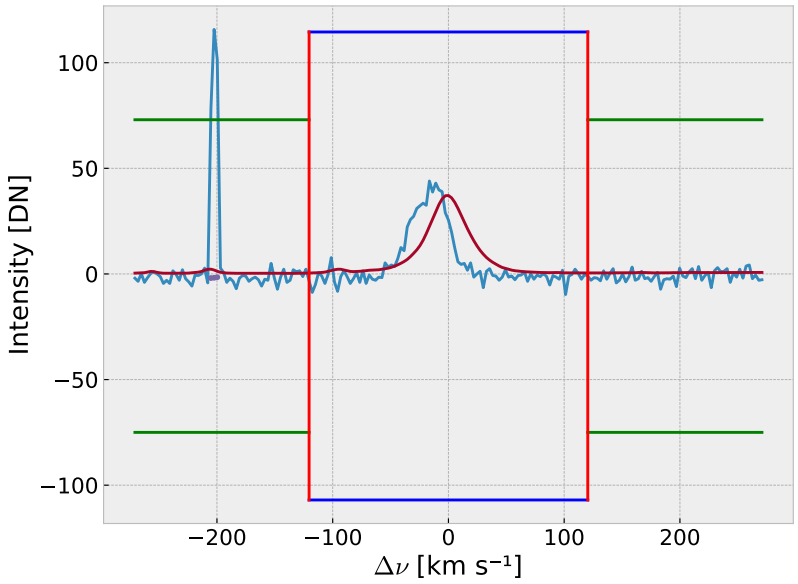

Figure 4.2: Result of segment despiking routine of an arbitrary Si IV 1394 Å profile from the data. The dark blue and green horizontal lines show the upper and lower IQR limits of the core and wings respectively. The red vertical lines indicate the separation between the core and wings. The dark red line is the mean Si IV 1394 Å profile over the FOV for reference. Optimized parameters are $\alpha_{\text{core}} = 10$, $\alpha_{\text{wing}} = 18$ and $w_{\text{core}} = 0.45$. The brown dots are the replacing points of the spikes as the average of the 7 neighbouring non-spike points.

It was found that due to the variety of shapes of the observed profiles, the quartile values had to be fixed with great care for the segment routine. For broad profiles, e.g. broad Si IV 1394 Å profiles, the upper IQR limit were placed at a satisfied distance from the highest valued data point that was not considered as a spike. This were not the case for narrow profiles as the points forming the emission line fall above the third quartile. In order for the emission line to affect the upper quartile range, the 75 quartile (third quartile) were replaced by the 80 quartile. Doing so sufficiently increased the upper limit in a satisfied manner. From figure 4.3 it is clear that adjusting the upper quartile from 75 to 80 increase the IQR range to rise above the steep profile peak. The adjustment of the upper quartile does not affect the limits of broad profile at the same level. That is due to broad profiles not being as steep as the narrow profiles adjacent to the 80 quartile.

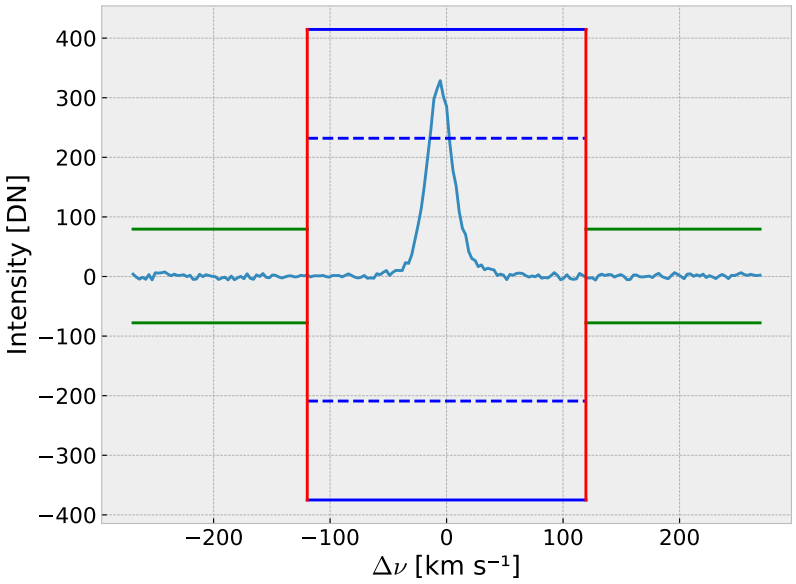

Figure 4.3: Adjusted despiking routine of an arbitrary steep Si IV 1394 Å profile. The third quartile is changed from 75 to 80. The solid blue line indicate the new IQR range using the 80 quartile while the dashed blue line indicate the old IQR range using the 75 quartile. The red vertical lines indicate the separation between the core and wing segments. Green horizontal lines are the IQR limits of the wing segment.

Scanning Despiking

It was found that it is not sufficient to perform the segment despiking routine alone. This issue were evident for small spikes relative to the profiles. Hence IQR_{upper} were placed above such spikes due to this effect as seen from figure 4.4. Therefore a second routine of despiking is implemented as the *scanning despiking routine*. As the segment procedure operate as a global routine, the scanning procedure operates as a local routine iterating through all wavelength points. The scanning routine also utilize the IQR method and is performed on a point $X_{i,j,l}$ along with its n_{filter} neighbouring points. The number of neighbours were optimized to $n_{\text{filter}} = 7$ so that the chosen wavelength points should in theory not vary much in value. From figure 4.5 it is perceptible that the profile core is significantly broadened. The segment routine in figure 4.4 were not able to identify the spike whereas the scan routine did successfully in figure 4.5. As the scanning routine performs well on spikes that is up to two points in size, it is not able to pick up spikes ranged over multiple points such as cosmic ray spikes. Such spikes will heavily affect the IQR range. Therefore, the despiking procedure should not be performed solely by the scanning routine. The most efficient approach of the despiking routines was to first activate the segment routine followed by the scanning routine. This order of implementation removed broad and narrow spikes in a satisfying manner, i.e. the k-means training were significantly improved following this procedure.

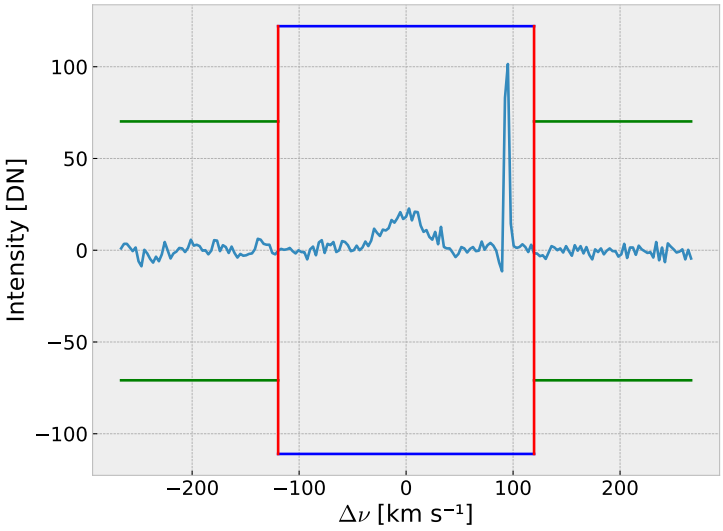

Figure 4.4: Failing of segment despiking routine of an arbitrarily Si IV 1394 profile in the data in solid blue color. The dark blue and green horizontal lines show the upper and lower IQR limits of the core and wings respectively. The red vertical lines indicate the separation between the core and wings. $\alpha_{\text{core}} = 10$, $\alpha_{\text{wing}} = 18$ and $w_{\text{core}} = 0.45$.

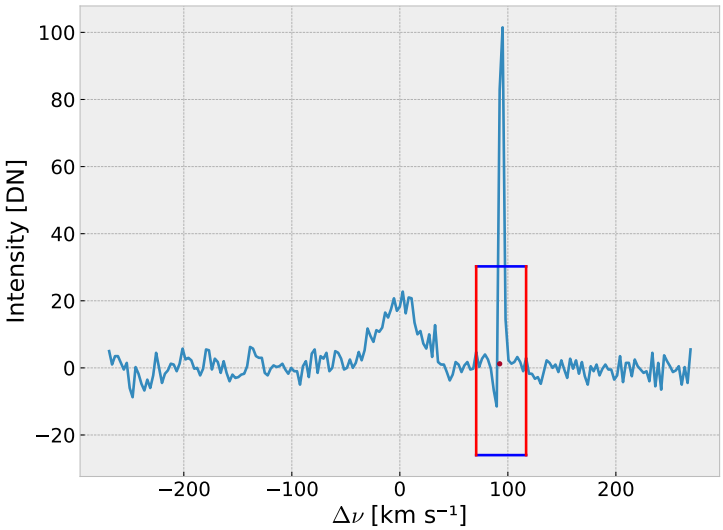

Figure 4.5: Result of a scanning despiking routine perform on the same observation as in figure 4.4. The dark blue lines show the upper and lower IQR limits on a subset of data points. The red lines indicate the subset of 8 points on each side of the brown dot. $\alpha_{\text{scan}} = 7$. The brown dot is the replacing point of the spike as the average of the 7 neighbouring non-spike points.

Parallelization

As explained in sections above, the despiking algorithm utilize two different despiking routines which operates computationally different. Due to the slowness of the scanning routine and that each pixel is computationally independent, parallelization is implemented to speed up the process. Parallelization utilize more of the potential from a machine by activating more central processing units (CPU). While the segment routine estimated time is approximately 30 seconds on a single thread, the scan routine estimated time were approximately 17 minutes using 40 threads. The computation time will vary depending on the specification of the machine used, but parallelization is still heavily recommended.

4.2 Machine Learning

The scientific data-set provided by the IRIS instrument as described in section 3.3.1, consists of 437200 unique profiles for single spectral channel. After the procedure of masking individual data points, the total number of analysable profiles were reduced to 394386. Investigating every single pixel to manually extract the ones that are of interest is clearly a inefficient task. The vast number of profiles yields to implement a machine learning technique to resolve such a problem. Machine learning is typically split into three categories, i.e. supervised learning, unsupervised learning and reinforcement learning Géron (2017). Supervised learning is a technique that is applied if the data-set is labelled. A model is trained on the labelled data-set where its parameters are adjusted based on the predictions of the model. Reinforcement learning is another machine learning technique where the model is continuously looking for patterns in the data environment. The model is given a reward if the prediction is good or punished if the prediction is bad. The final category is unsupervised learning. The answer to the data-set is unknown so it is up to the algorithm to find the optimal solution. There are no predefined answers of the structural shape of the observed profiles. Hence, an unsupervised machine learning technique will be implemented known as the *k-means clustering*. Before a model is trained, the data should be configured such that it reduce features that is of no interest for this study. The aim of this thesis is to locate Si IV 1394 Å and Si IV 1403 Å profiles that are structural complex, i.e. that significantly deviates from the ideal single-Gaussian-shaped profile. The venture of ignoring such profiles is attempted by implementing another machine learning technique called *Principal Component Analysis*.

4.2.1 Principal Component Analysis

The k-means clustering algorithm is a powerful tool to classify observational data into a fixed number of distinct groups. A drawback of this method is that it is not able give comparatively more weight to a specific property of the data, i.e. to specifically enhance profiles that are significantly deviating from the average profile in terms of shape. The average Si IV profile has a symmetric Gaussian profile which is not a characteristic of

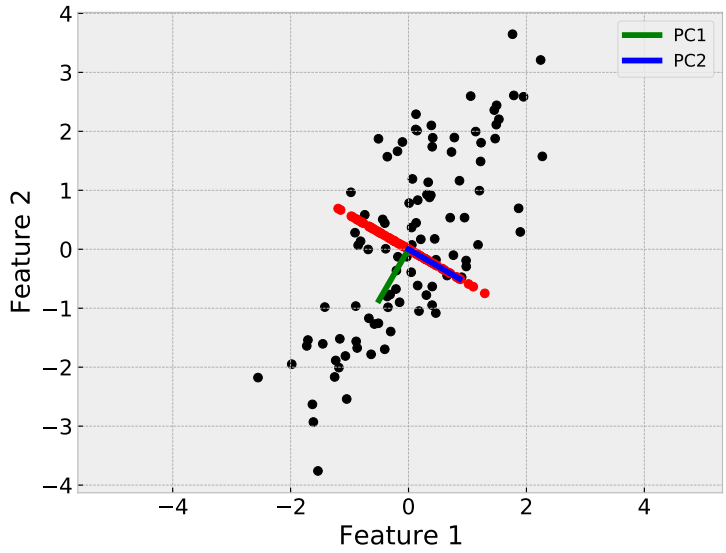

Figure 4.6: Example of a data-set consisting of 100 samples measured by two features. The data is normal distributed along both axis. The first and second principal components are shown as a green and blue line respectively. The red dots shows the data after the first component is removed.

an UV burst. A principal component analysis (PCA) (Pearson 1901) is implemented to effectively remove components that contributes to the characteristic of a single Gaussian profile.

The PCA algorithm have two main advantageous application that will be utilized on the observational data. First advantage is related with the dimensionality of the data-set. PCA manipulates the feature space in a way that the unit vectors points towards the direction where the variance of the data is maximized. These unit vectors are formally called principal components and are all orthonormal vectors. This gives possibility to drop the last principal components and still preserve a significant amount of the data. The dimension manipulation will remove some of the noise and speed up the training of a k-means model as the data is reduced. The second advantage is due to the removal of the characteristic Gaussian profile, the amplitude of a any observational profile will be of little significance.

To give an intuition on how the PCA operates, assume a data-set with 100 samples that contains two features each. A plot of the example data is shown in figure 4.6. The idea behind PCA is to find unit vectors from to the data-set such that the vectors directs toward maximum variance. Before the new basis are made, the data-set is centered for all features by subtracting the mean of each feature. The first vector with the greatest variance is called the first principal component. The second vector that is perpendicular to the first vector is called the second principle component and so on. All

components are perpendicular to each other in order to form a basis using the feature space. From the figure, we note that PC1 and PC2 as green and blue line respectively form a coordinate system that is a linear combination of feature 1 and feature 2.

Assume we have a data-set that can be written as a dense $m \times n$ matrix $\mathbf{X}$. In this case the m rows corresponds to the number of samples and n columns corresponds to the number of features, i.e. dimensions. Initial step is to shift the entire data-set so that the mean of each column equals to zero. Then define a n-dimensional unit vector $\hat{\mathbf{y}}_1$ that crosses the new origin. The entire data-set is projected to $\hat{\mathbf{y}}_1$ so that the euclidean distance between the origin and a sample is $d_i = \mathbf{X}_i \cdot \hat{\mathbf{y}}_1$ where $i = 1, 2, ..., m$. The sum of squared distance is then computed as follows:

$$D_1 = \sum_{i=1}^{m} d_i^2 = \sum_{i=1}^{m} (\mathbf{X}_i \cdot \hat{\mathbf{y}}_1)^2 \tag{4.3}$$

where D_1 is the total squared distance from origin to the samples projected on $\hat{\mathbf{y}}_1$. To obtain the first principal component, $\hat{\mathbf{y}}_1$ is adjusted so that the maximum possible value of $D_1 \rightarrow D_{1,\max}$ is achieved. When $\hat{\mathbf{y}}_1$ is optimized in such a way, the first principal component are then appropriately defined $\hat{\mathbf{y}}_1 \rightarrow \mathbf{p}_1$. Now D_1 is the *Eigenvalue* and $\mathbf{p}_1$ is the *Eigenvector* of the first principal component. Taking the square root of the Eigenvalue produce the singular value of the first component $\sqrt{D_1} = \sigma_1$ that is associated with the variance, i.e. we want to maximise the singular value to obtain the first principal component.

The second, and therefore the rest of the components in turn, is found by subtracting the contribution from the first j-th components from the data-set:

$$\mathbf{X}_{\text{reduced},j} = \mathbf{X} - \sum_{k=1}^{j-1} \mathbf{X}_k \mathbf{p}_k^{\mathrm{T}} \mathbf{p}_k \tag{4.4}$$

where $\mathbf{X}_{\text{reduced},j}$ contains the data but the first $j-1$ components are removed, $\mathbf{p}_k$ is the k-th principal component and $k = 1, 2, ..., j-1$. The superscript T denotes a transposed matrix. The summation on the right-hand side removes only the data in $\mathbf{X}$ parallel to $\mathbf{p}_k$. An example where the first principal component is removed from the data can be seen from the red dots in figure 4.6. This imply that a repetition of equation 4.3 is applicable. Replacing $\mathbf{X}_i$ with the reduced matrix result in the following:

$$D_j = \sum_{i=j}^{m} (\mathbf{X}_{\text{reduced},j,i} \cdot \hat{\mathbf{y}}_j)^2 \tag{4.5}$$

As before, the unit vector $\hat{\mathbf{y}}_j$ is adjusted so that the propagating vector towards maximum variance is repeatedly located until the number of principal components n_c are satisfied. Note that to form a new basis based on the variance of the data-set, the components must be orthonormal. Due to the removal of components in equation 4.4, the statement

$$\mathbf{p}_i \mathbf{p}_j = 0 \tag{4.6}$$

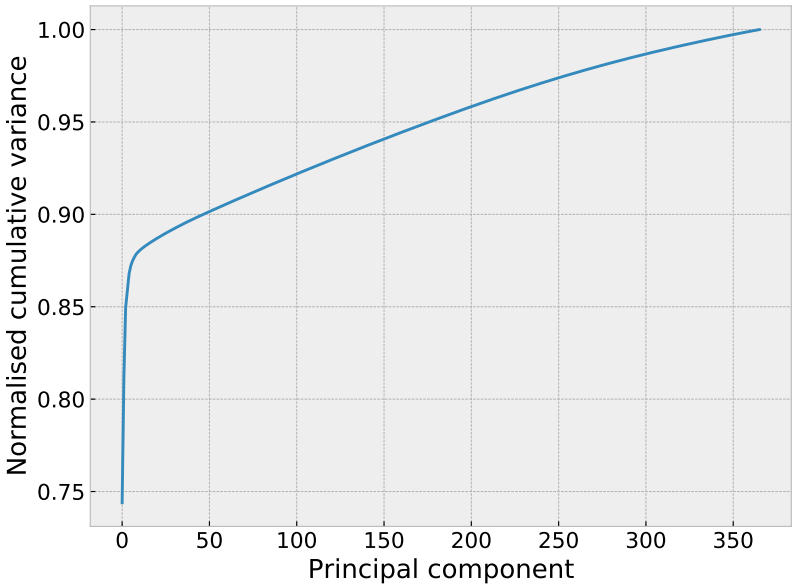

Figure 4.7: Cumulative significance of the variance from the principal components of the CH observation explained in section 3.3.1. The principal component analysis is performed on Si IV 1394 Å and Si 1403 Å lines.

must be true for all $i < j$ where $j = 2, 3, ..., n_c$ as $\mathbf{p}_j$ has no contribution from the first $j - 1$ unit vectors. This statement is clearly seen in figure 4.6 as the red dots have no contribution along the first principal component axis.

Using the same assumption of the data-set $\mathbf{X}$ as previously but now stating that the rows and columns correspond to observation and wavelength respectively. Each row has its own set of features with unique values, or in our case each observation has its own set of wavelength points. The matrix can be factorized into a product of multiple matrices using the singular value decomposition (SVD) Golub & Kahan (1965):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{4.7}$$

where $\mathbf{U}$ is a $m \times m$ orthogonal unit matrix, $\mathbf{\Sigma}$ is the $m \times n$ singular value matrix and $\mathbf{V}$ a $n \times n$ orthogonal unit vector matrix. $\mathbf{U}$ and $\mathbf{V}$ are known as the *left* and *right singular vectors* respectively as the columns on each forms the unit vectors. From the diagonal entries of $\mathbf{\Sigma}$, we have the strict-positive singular values $\sigma_1, \sigma_2, ..., \sigma_{\min(m,n)}$ and the remaining are zeros.

The rank of $\mathbf{X}$ is determined by the total number of non-zero σ_i. There are several approaches to solve equation 4.7 depending on the shape of $\mathbf{X}$ and case of interest. One way is to solve the full equation. Due to nature of $\mathbf{\Sigma}$ that the number of observations are much greater than the number of wavelength points, i.e. $m \gg n$, this method will be computationally inefficient. In addition, we reduce the number of singular values by defining the number of components n_c from the PCA procedure. A more efficient

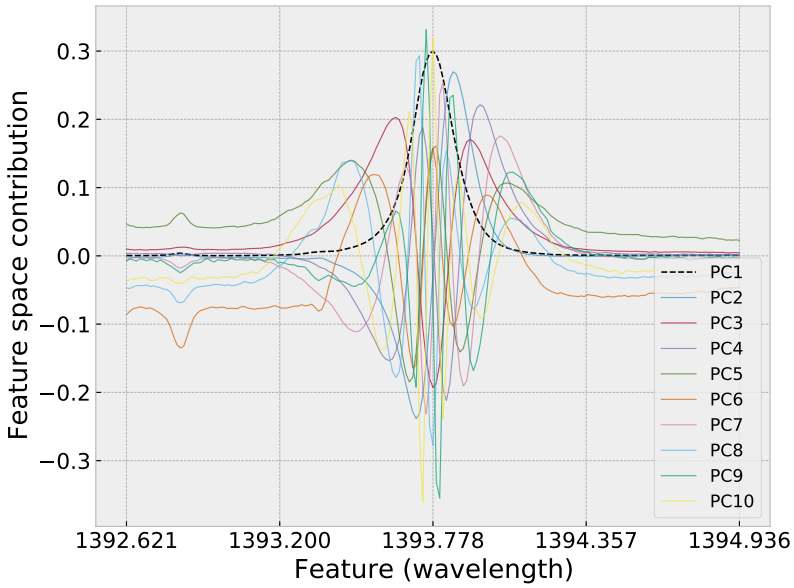

Figure 4.8: The first 10 principal components as PCA is performed on the CH observation as explained in section 3.3.1. The analysis is performed on the Si IV doublet profiles. The plot shows principal component contributions to the Si IV 1394 profile.

solution is to truncate the matrix $\mathbf{\Sigma} \rightarrow \mathbf{\Sigma}_{n_c}$ using a method of randomization (Halko et al. 2009). The singular value matrix is reduced to an $n_c \times n_c$ matrix where n_c is smaller than the rank of $\mathbf{X}$.

With the truncated decomposition of $\mathbf{X}$, the PCA procedure will sort the matrices such that the variance is of a decreasing order without interfering the data itself. To determine which components that preserves the majority the data, a plot explaining the variance is made as shown in figure 4.7. The variance of each component is scaled by the total variance so that the sum equals to 1. It is clear that the first 50 components preserves ≈ 90 % of the data and so the remaining components will not be used when training a k-means model. The first component covers approximately 74 % of the data-set implying that the referring component describe the majority of the data-set.

Figure 4.8 show the first 10 principal components of the Si IV 1394 line from the trained PCA model. The first component is shown as a black dashed line and it is clear that it establishes a single Gaussian-like profile as expected from the average Si IV profile. Therefore, to enhance profiles that have the characteristics of UV bursts, i.e. excessively broad non-Gaussian profiles possibly with multiple peaks, the k-means model should exclude the first principal component shown in figure 4.8.

4.2.2 k-means Clustering

The IRIS instrument perform raster observations where the data consists of a few hundred thousand pixels and each pixel contain several spectras where each spectra varies

in data points. Classifying each pixel is a huge task that should be executed by implementing an automatic method utilizing the potential from a machine. How each pixel will be classified is also not straight forward. The shape of the profiles is highly dynamic and varies due to many factors such as temperature, atomic abundance in solar atmosphere, velocity of plasma, optical thickness and so on (Peter et al. 2014; Young et al. 2018). Solving such classification problem leads to implementation of the *k-means clustering* algorithm (Forgy 1965; Lloyd 1982). k-means clustering is an unsupervised machine learning technique and its popularity is widely used among the astrophysical community (Bose et al. 2019; Rouppe van der Voort et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022). The intention of this algorithm is to optimize centroids that minimizes the euclidean distance to the observational profiles. These optimized centroids is referred to as representative profiles (RP) as they in fact represent closely relatable profiles within the data set. The observations will be grouped into k clusters and to determine the cluster size is not an easy task when the data-sets dimensionality equals to the total number of wavelength points. How the number of clusters were fixed is discussed below.

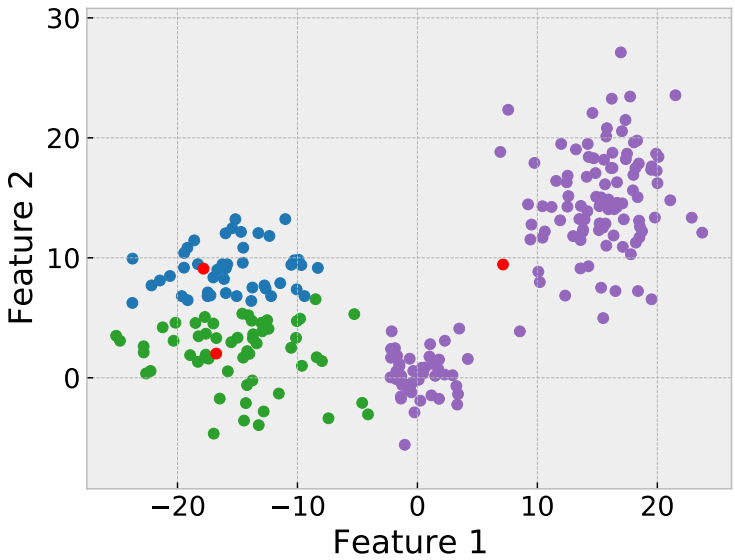

Figure 4.9: k-means clustering on a example data-set. The data-set consists of 250 data points and is measured by two features. The data is produced by three different random 2D-Gaussian distributions. The red dots show the centroids after a initial random guess.

Assume a simple $m \times n$ data-set $\mathbf{X}$ that is measured by two features as shown in figure 4.9. The data-set contains 250 samples so that $m = 250$ and $n = 2$. It is in general not obvious how many clusters the data-set could be split into. That is another problem and will be discussed later. In the example data-set mentioned, there is indication of possibly three clusters and therefore $k = 3$ is defined by the user. First, a random guess

of centroids $C = [c_1, c_2, c_3]$ are made and are shown as red dots in the example figure. Each sample is classified into the cluster that is closer in terms of euclidean distance between the sample and all cluster centroids. The algorithm will evaluate the fitting of the centroids based on the within-cluster sum-of-squares $WCSS$, also known as the *inertia*. The goal is to optimize the centroids by minimizing the inertia that is given by the following:

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{m_i} ||\mathbf{x}_j - c_i||^2 \tag{4.8}$$

where $\mathbf{x}_j$ is a measure of sample j, c_i is the mean of cluster i, i.e. the centroid of cluster i and is initially a random guess. m_i is the number of samples within cluster i. The centroid vector C are updated as the mean of samples within the corresponding clusters and so the same procedure is repeated until the inertia is almost non-changing. A general procedure can be broken down into 6 steps as follows:

1. Define the number of clusters k the user wants to group the m data into.
2. Make a random guess of the centroids $C = [c_1, c_2, ..., c_k]$.
3. Find the minimum distance between a sample and all clusters. The sample is assigned to the closest centroid. This is done for all samples.
4. Evaluate the inertia using equation 4.8.
5. The new centroids $C = [c_1, c_2, ..., c_k]$ are computed as the mean of their respective clusters.
6. Repeat step 3-5 until the inertia does not change more than some threshold.

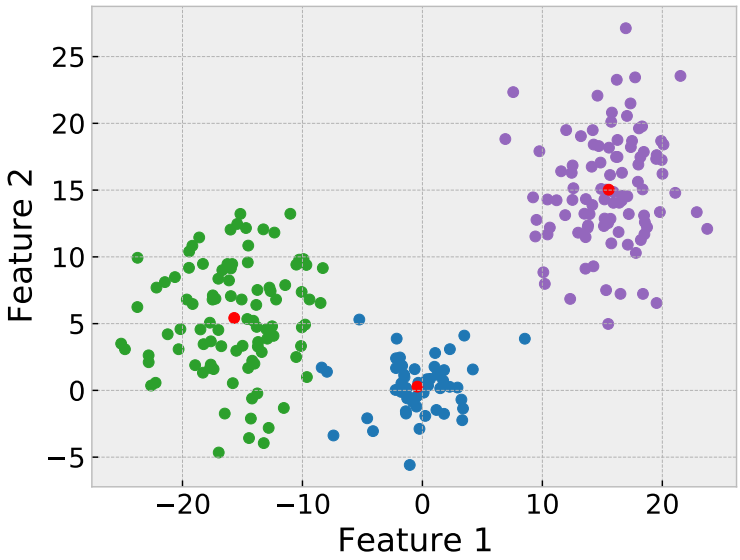

Figure 4.10: k-means clustering on the same data-set as in figure 4.9. The red dots show the centroid of clusters after 6 iteration of configurations. The data points are color-coded into their respective clusters.

The result when applying k-means on the example data is shown in figure 4.10. It is clear that the algorithm successfully classifies the samples into three distinct groups that are visually distinguishable. As an arbitrarily data-set might be a lot more complex, deciding the number of clusters k is not a straight forward task. Now consider the data-set $\mathbf{X}$ as explained in section 3.3. The number of features are determined by the number of wavelength points h meaning that clustering is now physically unobservable in contrast with the example data. In order to solve this problem on the data-set $\mathbf{X}$, the training procedure is repeated for $k \in [k_{\min}, k_{\max}]$ to estimate the number of clusters K. For each k, the algorithm computes the inertia $WCSS_k$ using equation 4.8 and stores the values in an array $\mathbf{I} = [WCSS_{k_{\min}}, WCSS_{k_{\min+1}}, ..., WCSS_{k_{\max}}]$. The inertia will always decrease with $k_i > k_{i-1}$ as the variance between a centroid and its cluster samples gets smaller with more clusters. The optimal solution would be $WCSS = 0$ but that would result in $k = h$ which makes the k-means procedure useless. It is better to consider the relative change of inertia $|I_i/I_{i-1} - 1|$. The absolute of the relative change is only to visualise the result in a decreasing order. k is then found by a determination of where the relative change appear to flatten. Figure 4.11 showcase how the relative inertia evolves with increasing value of k on the CH data-set. It is estimated from the figure that the curve seem to be flat beyond $k \approx 120$ and is then chosen to be the optimized number of clusters in this study.

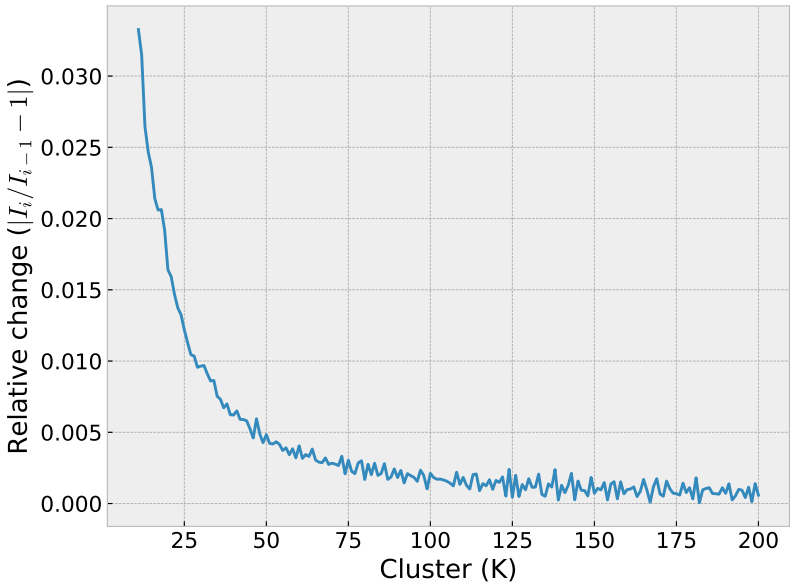

Figure 4.11: Relative inertia from trained k-means cluster models for increasing number of clusters k.

The k-means clustering is a robust algorithm for classifying data into distinct groups, but raise a problem of finding local minima if the initial centroids are not placed properly. Even though the inertia might not change by much for two consecutive iterations, the potential of centroids being ill-placed is plausible. To reduce the risk of this issue, the algorithm is repeated several times with different initialization of centroids. Each initialization is independent of each other, hence the algorithm is open for parallelization which significantly speed up the training process. Whichever model that result with the lowest inertia after convergence is then chosen as the best model. Another additional method to reduce the risk of local minima is to initiate the centroids in a more structured manner instead of complete randomization. The initialization method *k-means++* (Arthur & Vassilvitskii 2007) is implemented. Each new centroid are then chosen based on the previously selected centroids. Let $D(x)$ denote the distance between an arbitrary point $x \in \mathbf{X}$ and the closest centroid that currently has been defined. The initialization goes as follows:

1. First centroid is randomly chosen $c_1 \in \mathbf{X}$ by a uniform probability distribution amongst the data-set.
2. The following centroids $c_i = x' \in \mathbf{X}$ is randomly chosen based on the weighted probability distribution $P(x') = \frac{D(x')^2}{\sum_{x \in \mathbf{X}} D(x)^2}$.
3. Step 2 is repeated until $i = k$.

As seen from the probability distribution $P(x')$ from step 2, the numerator increase in value as the distance between a point x' and its nearest centroid is further apart.

The denominator is the summation of distances between all data points $x \in \mathbf{X}$ and its closest centroid $c_i \in C$ so that the weighted probability never exceed 1. From this, a new centroid that is randomly chosen further away from all centroids is favourable from the weighted probability distribution. According to Arthur & Vassilvitskii (2007), the k-means++ algorithm is proven to be better in terms of computation time and accuracy. The initialization is slower than randomization due to the necessary computation from step 2 above but it was found that the inertia converge towards a minima after less iterations.

4.3 Implementation of the Procedures on the Coronal Hole Data

This section aims to go though the sequential steps on the methodical approach of the IRIS observations used in this study. The procedure is listed as bullet point for easy repetition in potential future work.

1. Extract the IRIS level 2 observational data of interest from the IRIS data search engine. In this study, a CH atmospheric condition was of interest. To avoid degredation of the instruments, the data-set was acquired close to launch date.

2. Using the rest wavelength of all profiles to convert wavelength to Doppler shift by the conversion equation 3.1, each profile passbands were sliced into a satisfying range in terms of Doppler shift and so excess wavelength points were removed. The range for Si IV profiles are ±250 km/s, C II profiles are ±150 km/s, Mg II h&k are ±150 km/s and Mg II triplet in between h and k line is ±70 km/s.

3. The IRIS level 2 data-set went through a despiking routine. The despiking was performed on the most-sensitive spectral lines, i.e. Si IV 1394 Å, Si IV 1403 Å, C II 1335 Å and C II 1336 Å. This routine is essential due to outliers in the data-set are potential of affecting the training of RPs. This implementation is highly necessary as some observed profiles with unrealistically large spikes managed to produce their own isolated cluster. This issue occur due to the nature of training a k-means model as explained in 4.2.2. As Si IV and C II profiles are very different in shape, the despiking routine had to be tested with varying parameters for optimal result.

4. The decision on which spectral lines to apply machine learning methods on needs to be considered. Since this study aim to locate UV bursts type of event, only Si IV 1394 Å and Si IV 1403 Å lines will be used for training a PCA and k-means model.

5. Next is to apply PCA on the CH data-set. The data-set is reorganized into components where the components are arranged based on the strength of the variance in the data-set. Each component is vectorized using the feature-space where feature means wavelength in this study. As the most common profile is of

no interest, the first component, i.e. the profile that closely resembles the mean of the FOV, is removed. An example where the first component is removed is seen from figure 4.6. The red dots represent data without their first PCA component. Then, to reduce the dimension of the data-set but still preserve significant amount of information from the profiles, only 50 components were used (excluding the first component). The 50 components preserved ≈ 90 % of the data as seen in figure 4.7.

6. Before training a k-means model, it is essential to determine how many clusters the data-set should be reduced in to. The number of clusters k was determined by training multiple models iteratively with increased number of clusters. For each increase in k, the inertia gradient was computed and plotted as seen from figure 4.11. The number of clusters were chosen based on where the inertia gradient seem to flatten, i.e. at $k \approx 120$.

7. k-means training was performed on the manipulated CH data-set and did successfully produce RPs that closely resembles real asymmetric profiles. Each pixel in the data-set were labelled based on which cluster they were fitted in. The pixels that fall under a particular cluster of interest is further studied. The RPs that form complex non-Gaussian profiles are further studied.

Chapter 5

Results

Here is presented the results of k-means clustering and PCA applied on the CH data-set as explained in detailed in section 3.3.1. The results showcase the representative complex, asymmetric Si IV 1394 Å and Si IV 1403 Å lines along with C II 1335 Å, C II 1336 Å, Mg II k 2797 Å, Mg II h 2804 Å and Mg II triplet 2799 Å lines. Each pixel is labelled by the cluster number they fall in. It is then possible to localize and analyse events that is of interest in this study, i.e. events that fit the characterization of a UV bursts and then study their dynamics.

5.1 Representative Profiles

The k-means training were done on a CH observation by IRIS as explained in section 3.3.1. The model successfully segregated the representative profiles (RP) of Si IV 1394 Å and 1403 Å lines with various complex, asymmetric shapes. Most of the RPs were similar to a Gaussian. This is expected as the dynamics in the Solar atmosphere varies across the field of view but in general form a Gaussian-like profile for the TR Si IV spectral lines. Figure 5.1, 5.2 and 5.3 showcase the 30 Si IV 1394 Å and Si IV 1403 Å RPs which are the least frequent RPs over the FOV. The number of pixels with the specific observed profile that fall under the respective RP cluster is reported in square brackets from the RPs label.

RP 104 from the left panel in figure 5.2 indicate that the despiking routine did not handle all the spikes in the data-set. The label state that the cluster contains 67 similar profiles, which are statistically not very significant over the entire FOV. The spike here is relatively small and so the upper limit of the despiking routine did not correct such small anomalies. The RP 104 and the profiles in its cluster will be neglected in further analysis.

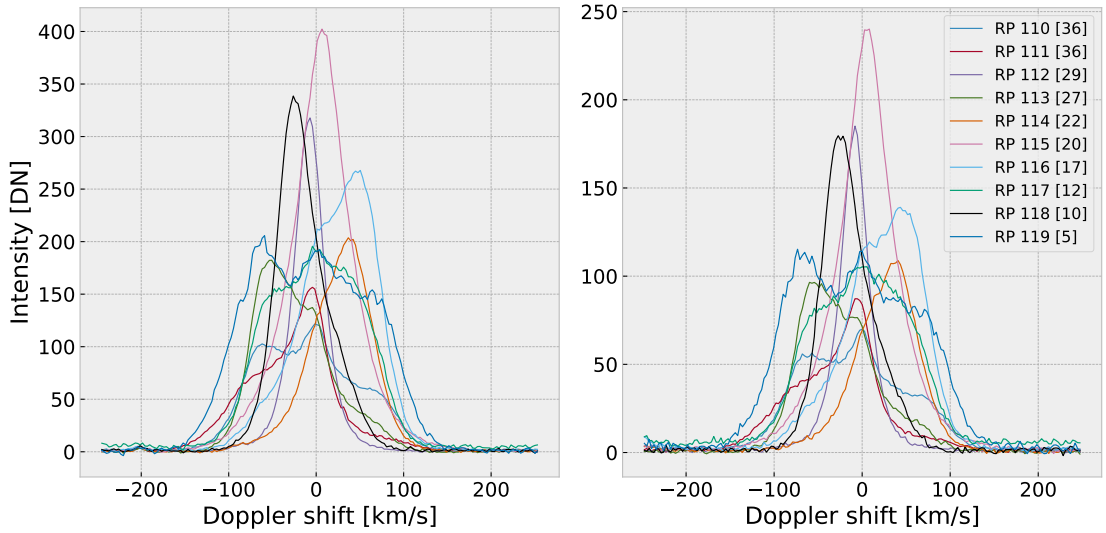

Figure 5.1: 10 least frequent Si IV 1394 Å RPs in left panel and Si IV 1403 Å RPs in right panel for the optimized k-means model. Colored curves in both panels represents the same cluster and is labelled in the right panel. Each RP is labelled with the number of observations within that cluster in square brackets.

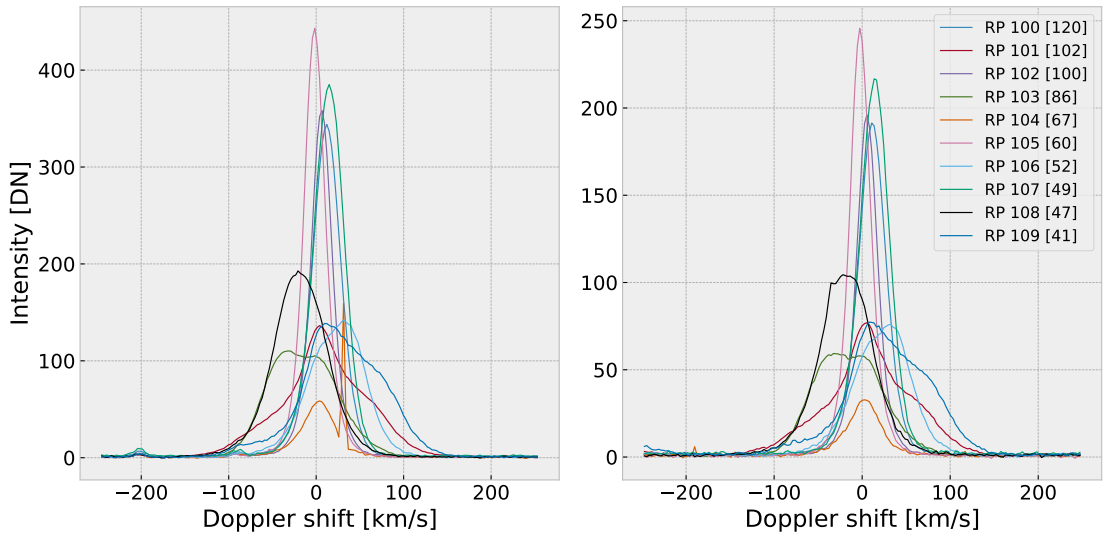

Figure 5.2: Same as figure 5.1 but with 11th to 20th least frequent RPs.

Each RP are clustered with a number of observed profiles that are similar in shape and intensity. To visualise how the observed profile fits with their respective RP, a probability distribution function (PDF) plot are made as seen from the example of both Si IV 1394 Å and Si IV 1403 Å profiles in figure 5.4. Along the PDF, the observed

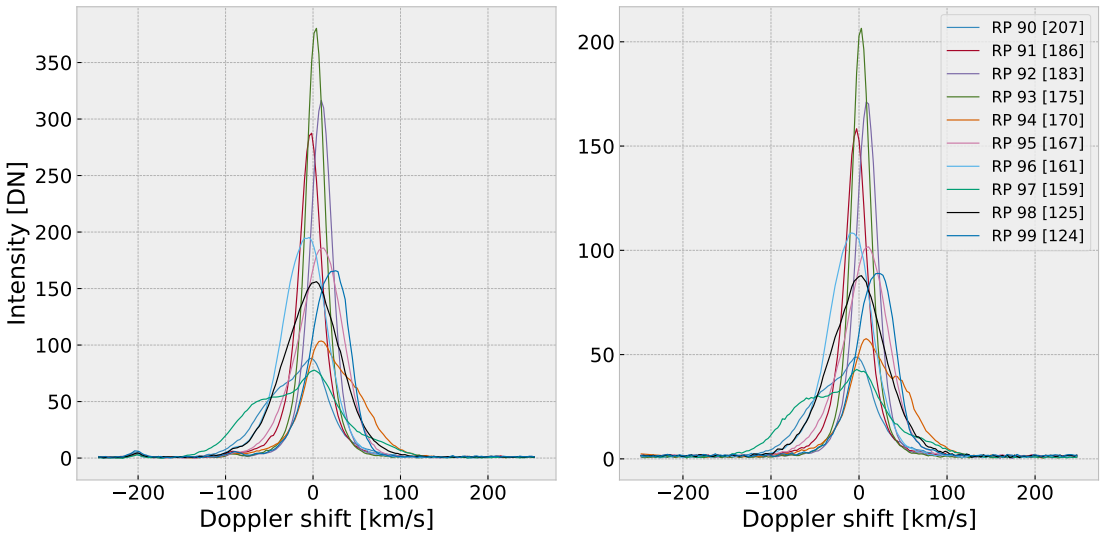

Figure 5.3: Same as figure 5.1 but with 21th to 30th least frequent RPs.

profile that deviates the most from the RP is overplotted as a dashed-dotted line. The general shape of the background shading is consistent with its respective RP. Also the most deviating profile match the RP in terms of shape quite well. Note that since the despiking routine did not handle all spikes in the data-set, some of the most deviating profile with an unwanted spike was observed. The issue became frequent for clusters holding more than 2310 observed profiles, i.e. cluster 47 and lower.

The result after applying k-means clustering on the CH data where each pixel is labelled with their best fitted RP, is shown as a map-plot on right panel in figure 5.5. It is evident from the map that the less frequent RPs are clustered at location where stronger total Si IV 1394 Å line intensity is observed, as seen from the left panel. The red and white, i.e the less frequent profiles, are typically present at bright network and its surroundings, i.e. network-internetwork boundary regions, while blue more frequent profiles are present in the dark internetwork. Bright network at the FOV are usually less frequent than the surrounding internetwork. Therefore, these seeings verify that the implemented method performs well.

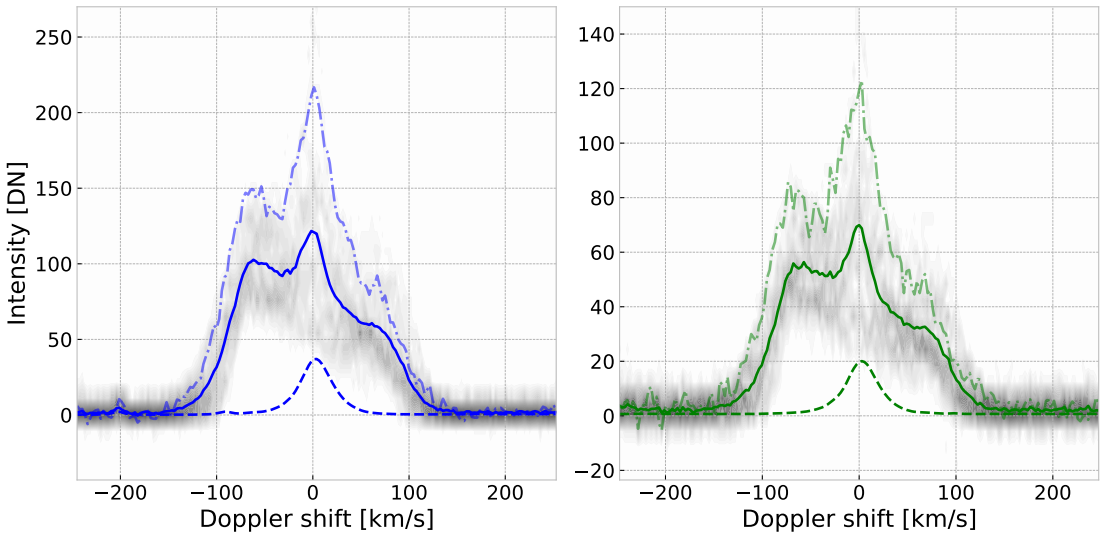

Figure 5.4: Distribution of Si IV 1394 Å profiles in left panel and Si IV 1403 Å in right panel in cluster 110. The solid line is RP 110, dashed line is the FOV average, dashed dotted line is the most deviating observed profile relative to RP 110 and the background shading is the distribution of observed profiles within cluster 110.

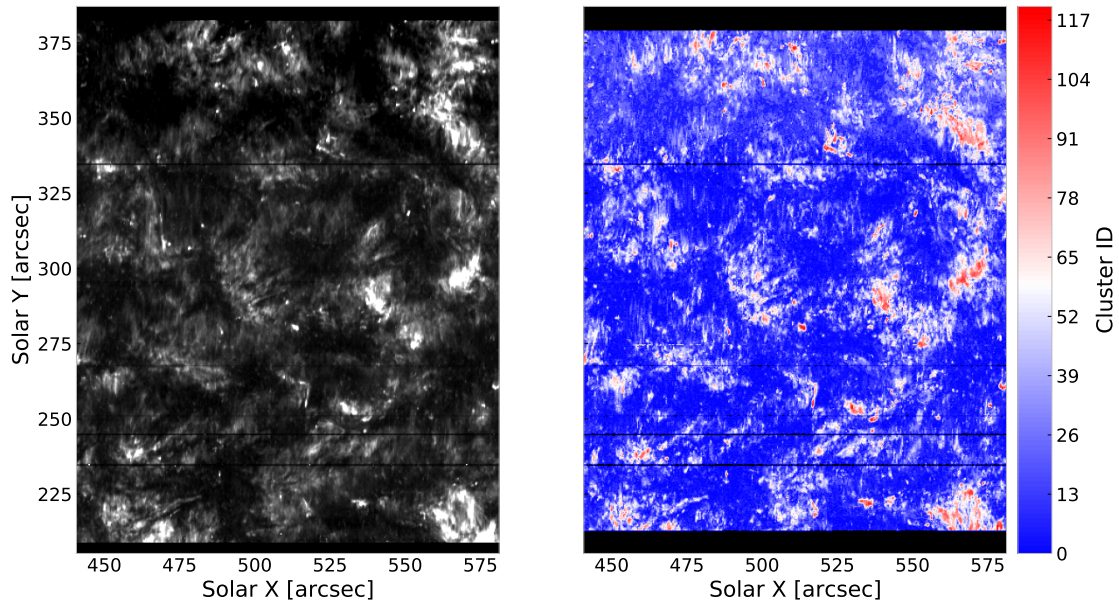

Figure 5.5: The left panel show the Si IV 1394 Å scaled total intensity profiles. Right panel is map of RPs from the data-set as explained in section 3.3.1. The dark horizontal lines are pixels that were masked as bad, hence not used for training the K-means model. The darker blue pixels are the most frequent RPs in terms of cluster size. The darker red pixel are the least frequent pixels. The cluster ID from each pixel can be seen from the color bar on right side.

5.2 Diagnostics of UV bursts

All 120 RPs of Si IV 1394 Å and Si IV 1403 Å line profiles were carefully characterized by eye that fits the category of UV bursts as described in section 1.3. Si IV 1394 Å and Si IV 1403 Å RPs were specifically picked out by looking for shapes of profiles with multiple peaks or extended wings in one or both sides. A combination of the two features are not considered as profile with multiple peaks are naturally extended in one or both wings.

5.2.1 Detection of UV Burst Profiles

The shapes are determined based on comparison with the FOV mean profile from the CH data-set. Figure 5.6 visualise four examples of Si IV 1394 Å RPs with multiple peaks. Note that the intensity is scaled for easy comparing with the mean, hence the amplitude of the intensity is not considered when UV bursts were categorized and therefore may vary between the four examples. It is expected that most profiles with multiple peaks are naturally very broad. The upper left panel from figure 5.6 show RP 103 as solid line along with the FOV mean as dashed line. The core of the RP seem to be slightly reversal that could suggest an effect of absorption and therefore two distinct peaks are evident. In the upper right plot RP 110 contain two enhancements on both sides which result in a red and blue components to the intensity. The blue peak is quite clear while the red peak is small but still present. The lower left panel show RP 113 that is blue shifted to ≈ -52 km/s with a small peak at ≈ -4 km/s. The lower right panel show RP 119 which is the most dynamic profile with two clear peaks on both side of the core. Note the blue peak is more intense than the core peak. The common behavior of all four panels is that they are profiles with large excess broadening compared to the FOV average profile as expected. Profiles that fall under the category of extended wings are those with extended wings on one or both sides and do not show sign of superimposed peaks. Example of such profiles are seen in figure 5.7. The upper left panel show RP 51 that is enhanced in blue wing relative to the core while RP 77 in upper right panel show an enhanced wing on red side. They indicate an upflow and downflow of the plasma respectively. Lower left panel show RP 86 with enhance wings on both sides and are close to a symmetric line. The lower right panel show RP 101 which contain enhanced wings on both sides but stronger on the red wing. The lower panels both indicate a bi-directional flow of plasma.

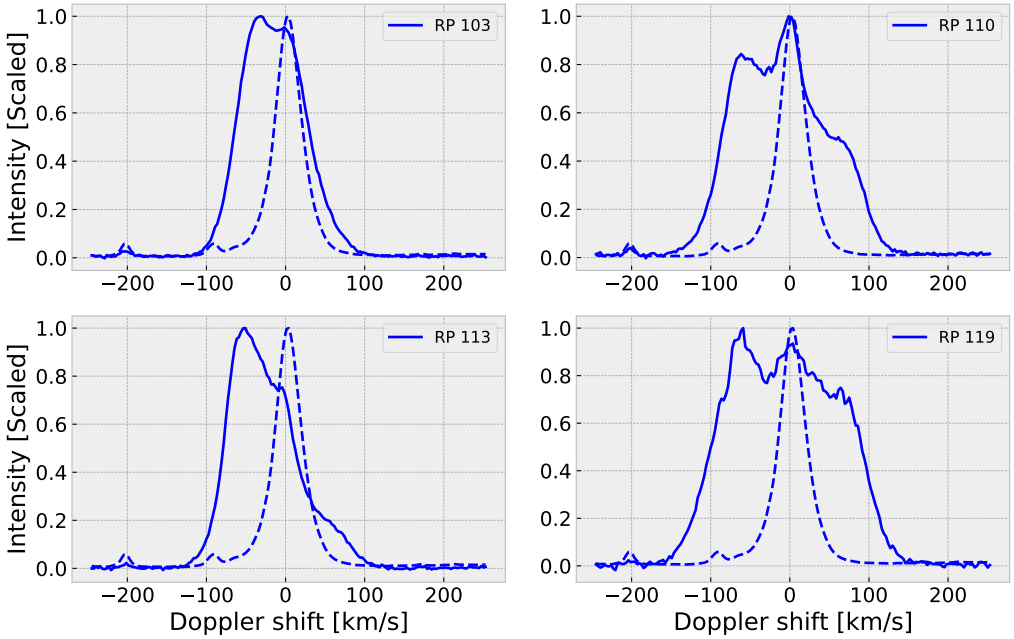

Figure 5.6: Examples of Si IV 1394 Å RP as solid line that fall under the category of UV burst with multiple peaks compared to the FOV average profile as dashed line. The intensities are scaled for easy structural comparison with the FOV average. Doppler shifts are given in units relative to rest wavelength in vacuum given in table 3.2.

The RPs that were categorized as UV bursts are tabulated in table 5.1. There are in total 10393 unique profiles that fall under the characteristics of UV bursts, resulting in total 2.6 % of the entire data-set. Out of the UV bursts type of events, 9 % show multiple peaks and 91 % show extended wings with 8 and 18 RPs respectively. It is clear that profiles with extended wings are the dominant UV bursts feature in the CH data-set.

Table 5.1: Tabulated RP that match signatures corresponding to a UV bursts feature. There are in total 10393 observations that fit the definitions.

	Multiple peaks	Extended wings
Representative profile	74, 94, 103, 110, 113, 116, 117, 119	51, 54, 59, 61, 67, 68, 77, 86, 89, 90, 97, 101, 106, 109, 111, 114, 115, 118
Total	937	9456

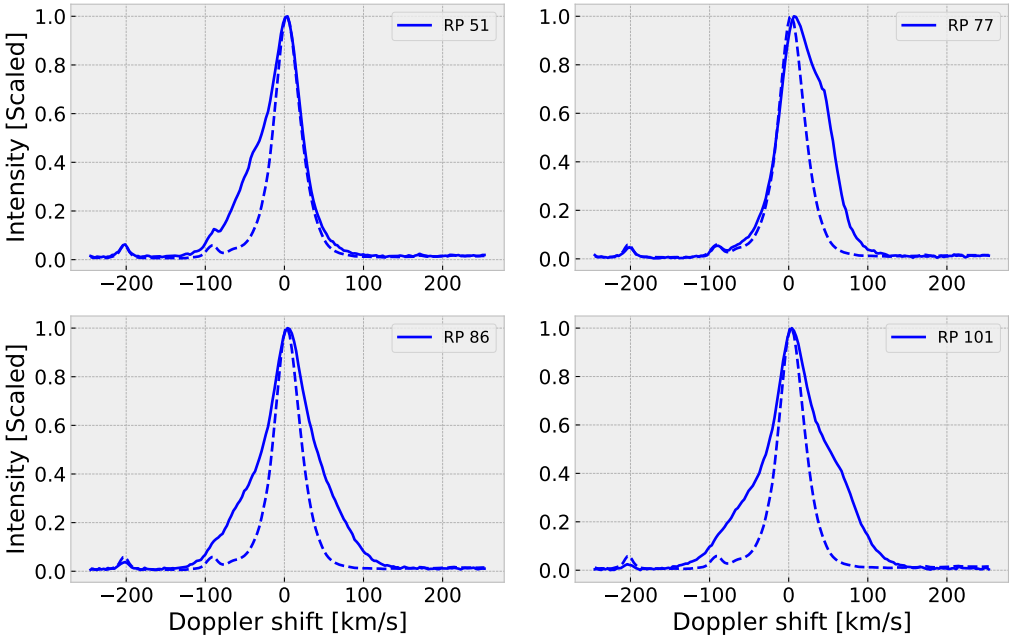

Figure 5.7: Example of a Si IV 1394 Å RP with extended wings. The FOV average and units are similar as in 5.6.

5.2.2 Localization of UV Bursts

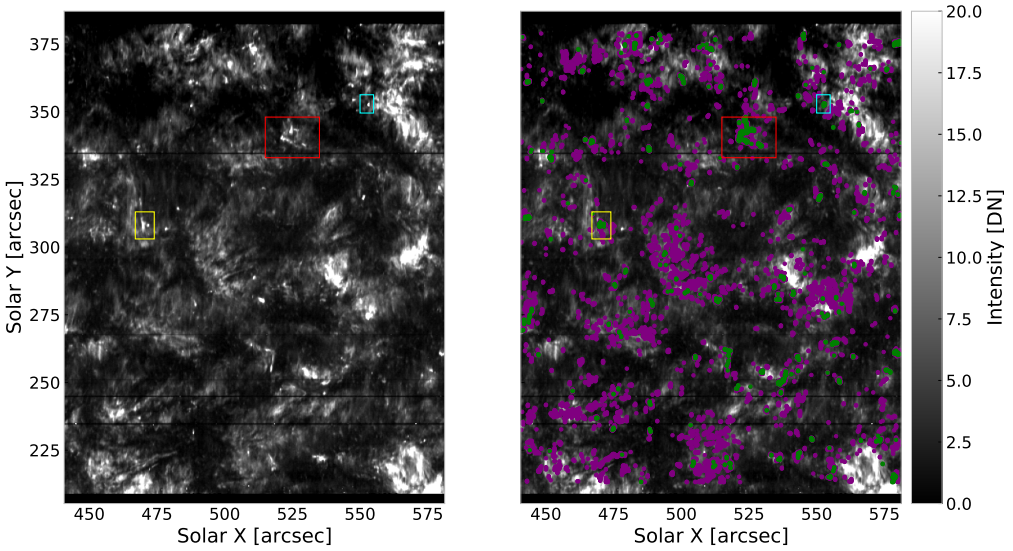

Figure 5.8: Raster map of scaled total intensity of Si IV 1394 Å profile. Right panel is the same raster map with superimposed UV bursts events. Purple pixels are observations of Si IV 1394 Å profiles that have extended wings and green pixels are profiles with multiple peaks.

The location of pixels that is categorized as UV bursts are superimposed on a Si IV 1394 Å raster map in figure 5.8 in the right panel. The pixels with UV bursts signatures are color-coded according to characterization from table 5.1. It is clear that the events are ubiquitous but the location seem to be favourable at the network-internetwork boundary regions. Additionally, some small and distinct intense brightenings are also observed to be associated with UV burst profiles as shown in enclosed boxes in figure 5.8. One example is the red box at Solar X 530 arcsec and Solar Y 340 arcsec where a small patch of UV bursts were located. At this region a combination of Si IV profiles with multiple peak feature and extended wings are present. The location of the rectangle are also superimposed on C II 1335 Å and Mg II k 2797 Å raster map in figure 5.9 and 5.10 respectively, for comparing between the location of UV bursts signature at different heights. There were indications of brightenings also in lower- and upper chromosphere as seen from the figures, at the same coordinates. This findings imply energetic events that extends over a range of heights. The appearance of such events seem to be more prominent at the TR region as seen from the Si IV 1394 raster map.

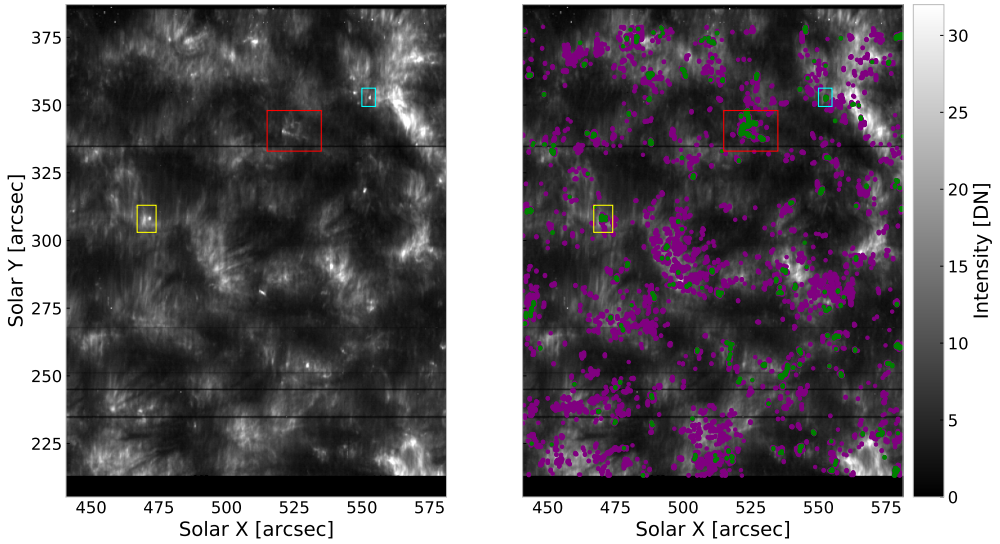

Figure 5.9: Raster map of scaled total intensity 1336 Å profile. Right panel is the same raster map with superimposed UV bursts events. See caption in figure 5.8 for details on colored dots.

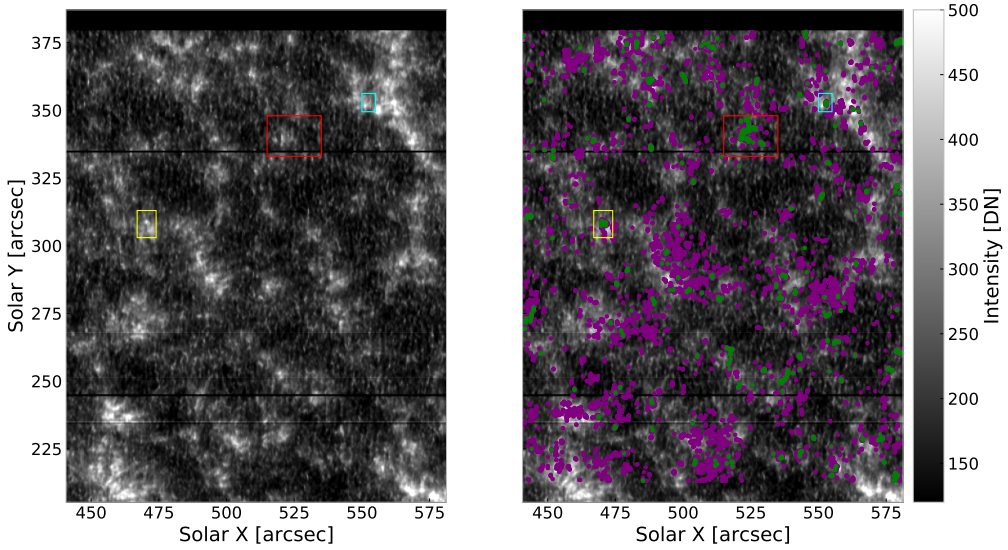

Figure 5.10: Raster map scaled total intensity of 2797 Å profile. Right panel is the same raster map with superimposed UV bursts events. See caption in figure 5.8 for details on colored dots.

Another small-scale brightening in figure 5.8 is found in the yellow box at solar X 475 arcsec and Solar Y 310 arcsec. There are indication of subsequent small-scale brightening in the upper and lower chromosphere in figure 5.9 and 5.10 respectively at the same coordinates implying that these energetic events also occurs over a range of heights. These events at the yellow box are discussed in detail in the next section.

5.3 Case Study of UV Bursts Associated with Compact Brightening and Network Jet

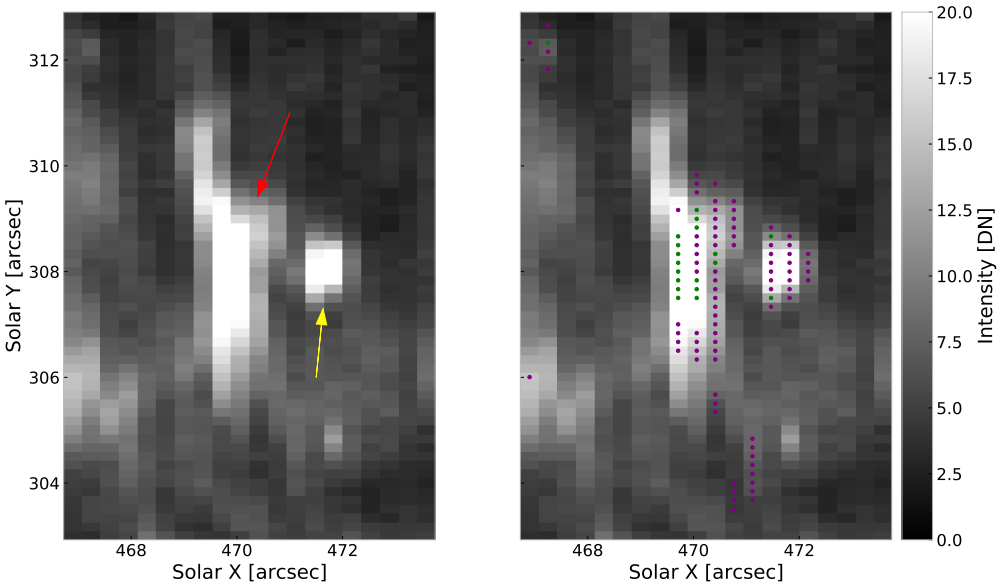

Figure 5.11: Close-up raster on the scaled total intensity Si IV 1394 Å profile. See caption in figure 5.8 for details on colored dots. Red arrow points to a jet-like feature and yellow arrow point to a small-scale compact brightening.

Figure 5.11 is a close-up raster map on an intense Si IV 1394 Å brightening region where multiple of pixels were categorized as UV bursts. The close-up is the same region as the yellow box from figure 5.8. The close-up raster stipulate an elongated event that could be considered as network jet. The jet is pointed by a red arrow. This small-scale jet like feature is also evident in C II 1336 Å raster map from figure 5.12, but not as clearly visible in Mg II k raster map from figure 5.13. The central green dots of the jet from the right panel insinuate that the Si IV 1394 Å profiles have features with multiple peaks. The purple dots on right side of the jet suggests profiles with extended wings. The vertical flow of the jet is not obvious but the RP map 5.5 inform that the spectral lines under the green dots fits with RP 103. Figure 5.2 manifest RP 103 as a broad emission line with two peaks where the blue dominant peak have a Doppler shift of

≈ -30 km/s implying a significant amount of upflow of the plasma.

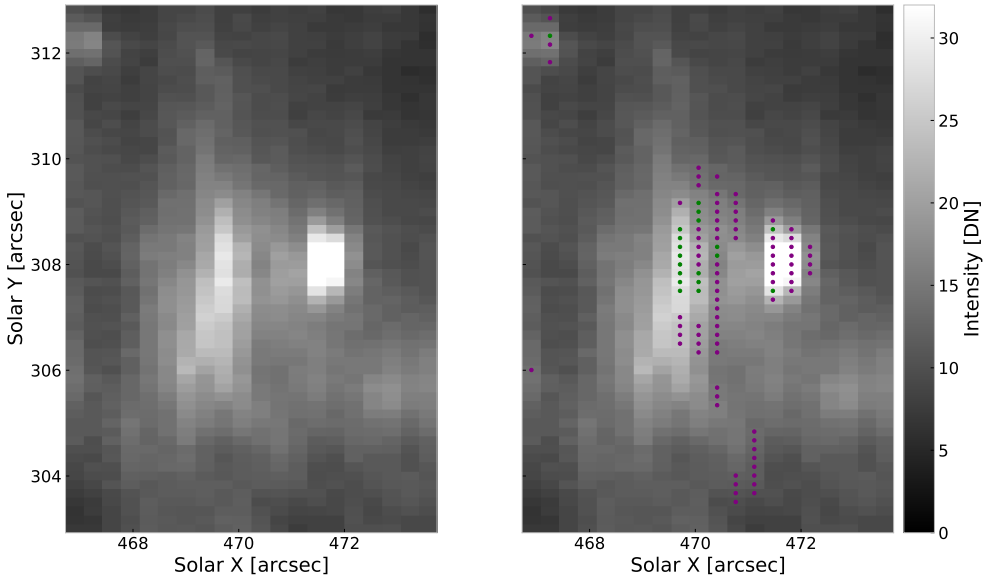

Figure 5.12: Close-up raster on the scaled total intensity C IV 1336 Å profile. See caption in figure 5.8 for details on colored dots.

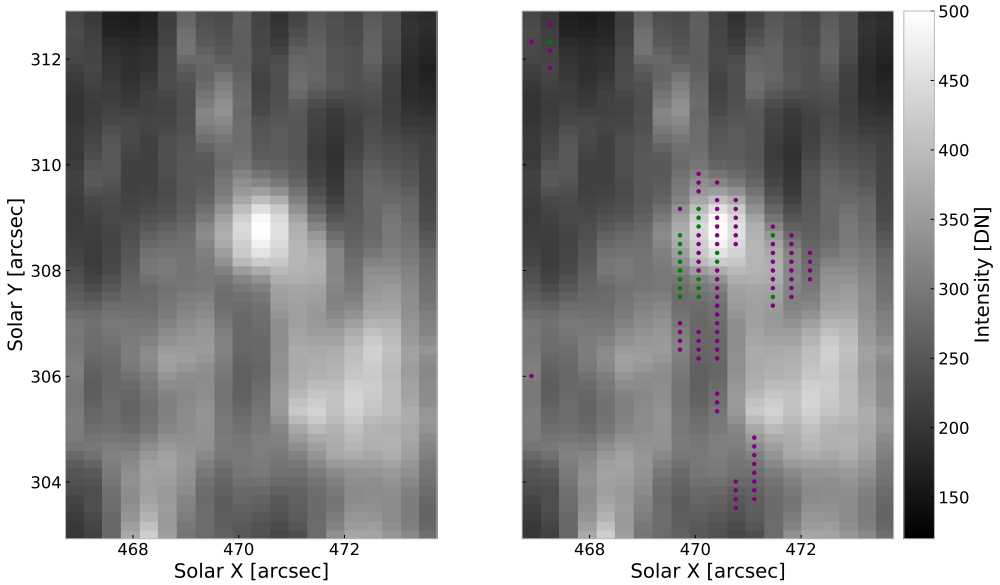

Figure 5.13: Close-up raster on the scaled total intensity Mg II k 2797 Å profile. See caption in figure 5.8 for details on colored dots.

In figure 5.14 the RP 103 of TR Si IV 1394 Å and Si IV 1403 Å are shown along with the lower and upper mean chromospheric lines within the same particular cluster. As mentioned in previous paragraph, the Si IV lines have a peak Doppler shift at ≈ -30 km/s that indicates an upflow of plasma. It is evident that there are stronger C II 1335 Å and C II 1336 Å lines compared to the FOV average which indicate that the upper chromosphere might be contributed to the particular events. In contrast, the Mg II k 2797 Å and Mg II h 2803 Å lines are slightly enhanced compared with the FOV average which suggest that the mid chromospheric region are not as energetic, hence not necessarily linked with the events. This is also the case for the low chromosphere Mg II 2799 triplet line as the solid line closely resembles the FOV average.

Next to the jet, a small-scale compact brightening in Si IV raster map is evident marked with a yellow arrow in left panel from 5.11. The compact brightening span a 1.0×1.7 arcsec^2 area. The features are mainly Si IV 1394 Å lines with extended wings as most of the pixels are marked with purple dots. The RP map 5.5 inform that this compact brightening is labelled mainly with RP 97 and RP 101. Both of the RPs have extended wings towards red and blue as seen in figure 5.3 and 5.2 respectively, indicating a bi-directional flow of the plasma in TR. The compact feature is strong and is at similar location in the C II 1336 Å raster from figure 5.12 but appear to be slightly uncoordinated in position of a possibly compact brightening in Mg II k 2787 Å raster map as seen from figure 5.13.

Figure 5.15 show profiles in the corresponding cluster 97. The Si IV lines are enhanced on both wings but especially in the blue wing indicating a dominant upflow stream in TR. The C II lines are enhanced in the core compared to the FOV average implying a denser plasma. Also, red peak are stronger which imply that the core are slightly blue shifted and again suggest a upflow of plasma in upper chromosphere. Mg II h&k line closely resembles the dashed FOV average and also k_{2R} (h_{2R}) and k_{2B} (h_{2B}) peaks are near equal in amplitude. The low-chromosphere has therefore no indication of flow of the plasma in contrast to the FOV average which have slight downflow as a result of the stronger blue peak. The low chromospheric Mg II triplet line resembles almost exactly the mean.

Figure 5.16 show similarly all profiles but in cluster 101. The TR Si IV profiles are enhanced on both wings indicating a bi-directional flow but slightly more downflow of the plasma as the red wing is stronger. The C II profiles show similar behavior as in figure 5.15 which indicate a denser region compared to FOV average with an uplfow of the plasma. The major difference between cluster 97 and 101 is that the latter have stronger Mg II profiles and there are also evidence of stronger k_{2B} (h_{2B}) peak compared to the k_{2R} (h_{2R}) peak implying downflow in the mid chromosphere. The low chromospheric Mg II triplet profiles are enhanced compared to the FOV average. The two RPs that dominates the compact brightening contradicts in terms of the extent of the event. It is indication of the present events at these Solar locations extends in the upper chromosphere and TR, still inconclusive at lower regions. Note that RP 97 is only evident on left part of the brightening while the RP 101 is only evident on the right part of the compact brightening seen from figure 5.11. One possible reason for the contradiction might be a result of the time of exposure from the instrument, i.e.

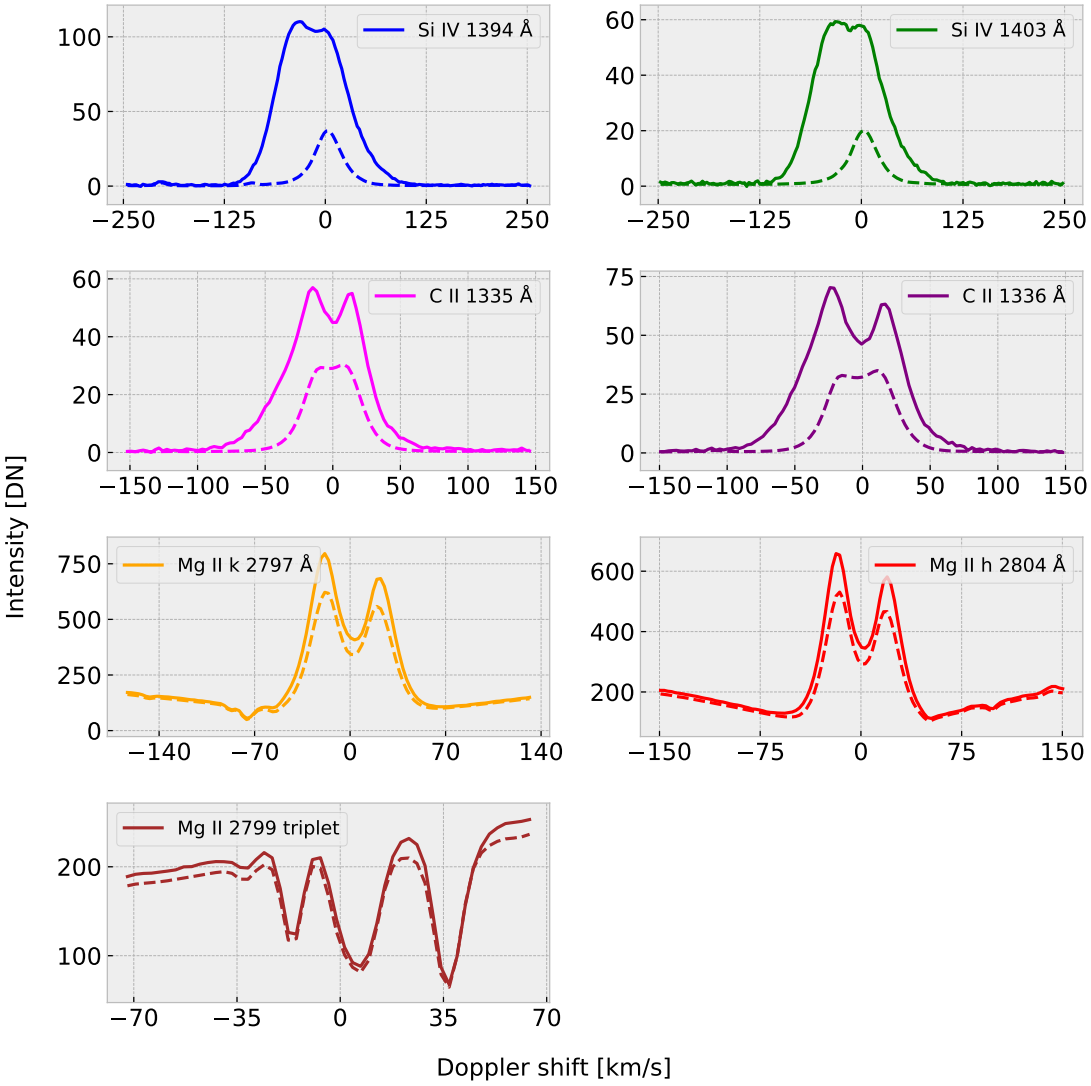

Figure 5.14: RP 103 of Si IV 1394 Å and Si IV 1403 Å at the upper row respectively as solid lines. The second and up to the fourth row show solid line as the mean of the profiles within that particular cluster. The name of the profiles are given in the label. The dashed line in every subplot are the FOV average profile for reference. Doppler shift is given in units relative to their respective rest wavelengths in vacuum from table 3.2.

the left part of the brightening was present at an earlier time than the right part of the brightening.

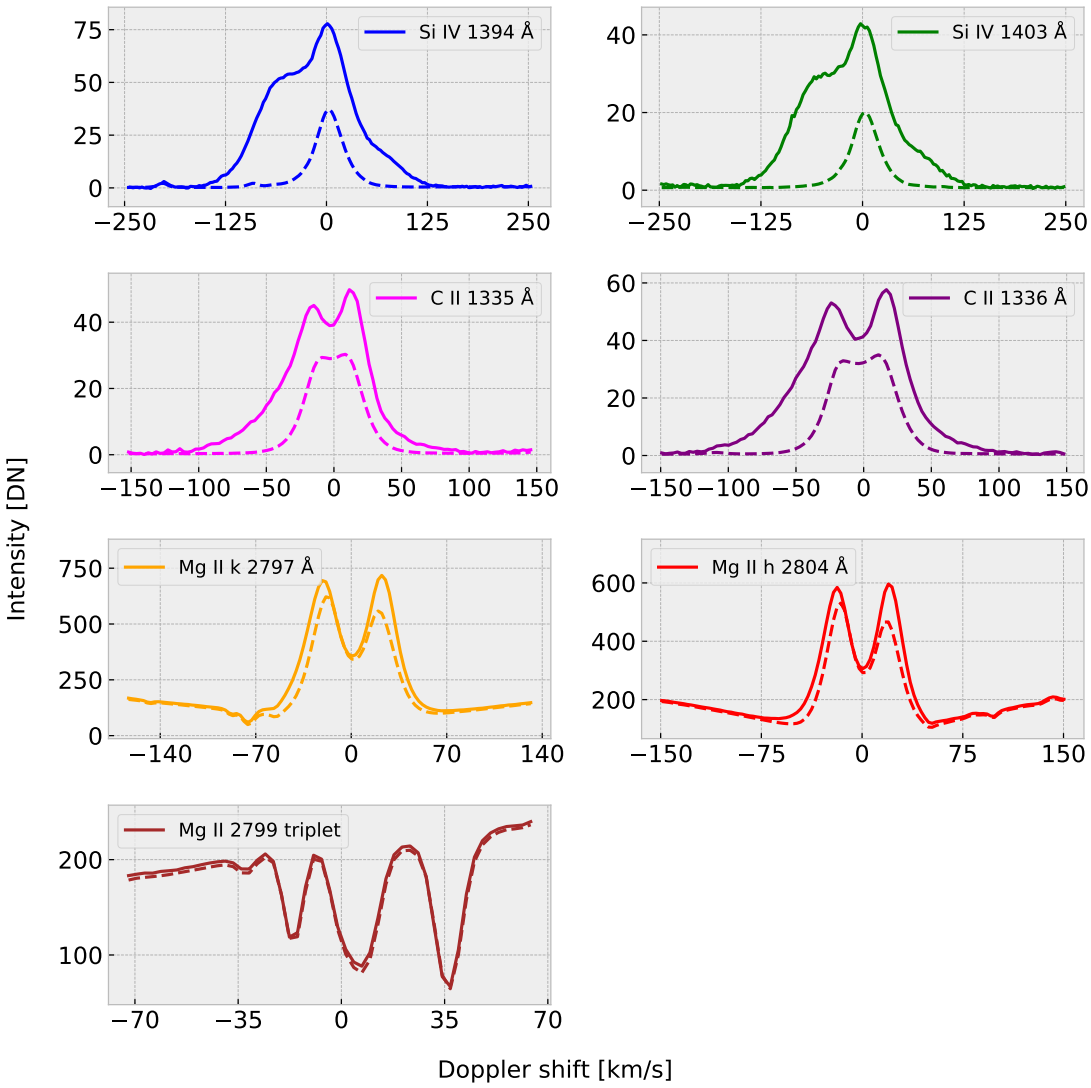

Figure 5.15: RP 97 of Si IV 1394 Å and Si IV 1403 Å at the upper row respectively as solid lines. The second and up to the fourth row show solid line as the mean of the profiles within that particular cluster. The name of the profiles are given in the label. The dashed line in every subplot are the FOV average profile for reference. Doppler shift is given in units relative to their respective rest wavelengths in vacuum from table 3.2.

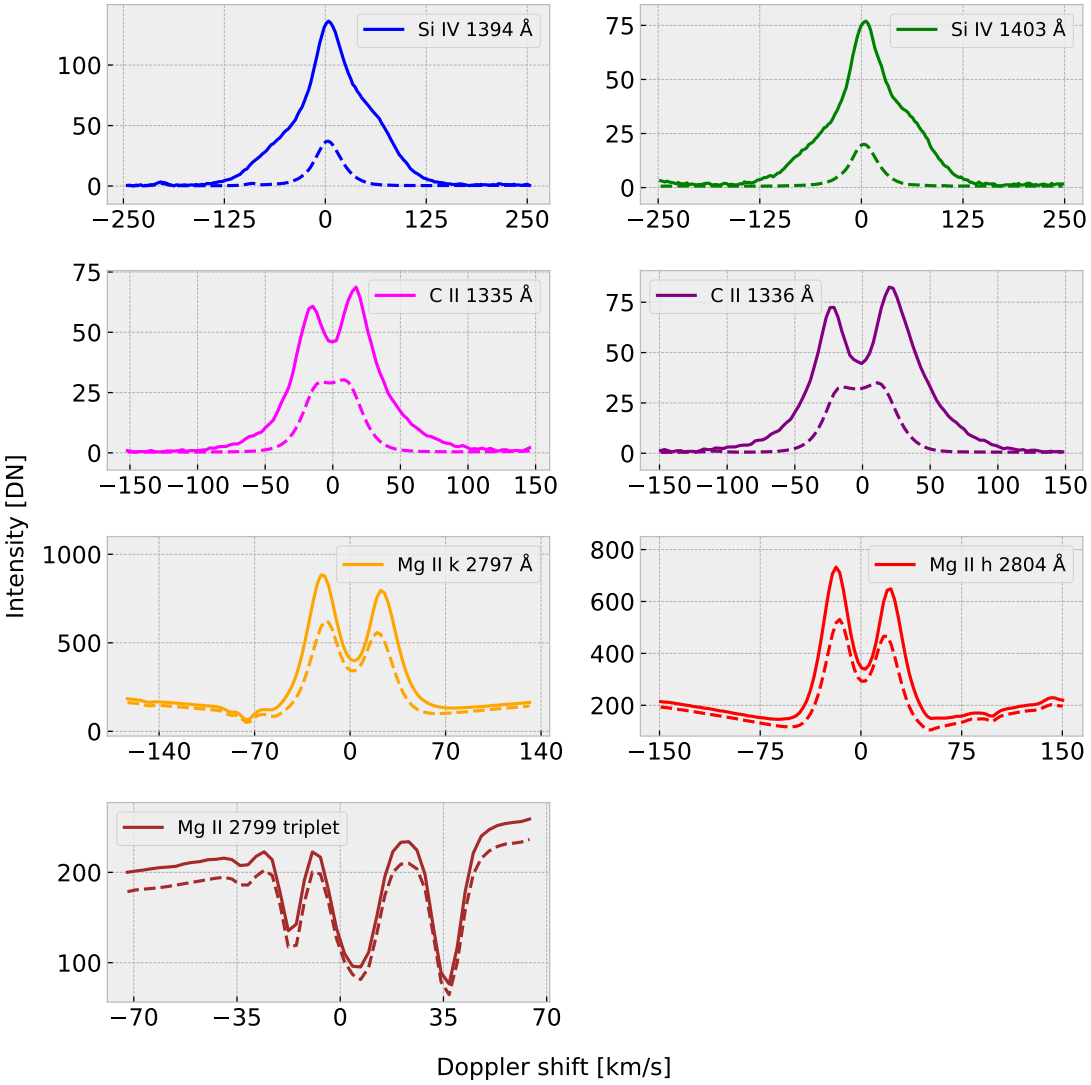

Figure 5.16: RP 101 of Si IV 1394 Å and Si IV 1403 Å at the upper row respectively as solid lines. The second and up to the fourth row show solid line as the mean of the profiles within that particular cluster. The name of the profiles are given in the label. The dashed line in every subplot are the FOV average profile for reference. Doppler shift is given in units relative to their respective rest wavelengths in vacuum from table 3.2.

5.4 Case Studies of the Least Frequent Representative Profiles

5.4.1 Case 1: RP 119; Ni II λ 1393 blend

RP 119 from figure 5.1 is the only Si IV 1394 Å RP that contain a faint blend at approximately -91 km/s which is due to absorption of Ni II λ 1393.33 Å line (Peter et al. 2014; Young et al. 2018). The observed profiles within the cluster are shown in figure 5.17. The blend seem to be slightly evident in the purple, blue and orange lines and they are adjacent at Solar X 553.0 arcsec and Solar Y 352.6 arcsec as seen in the cyan box in figure 5.8, 5.9 and 5.10. A close-up raster map of the cyan box is seen from the TR, upper and mid chromosphere in figure 5.18, 5.19 and 5.20 respectively.

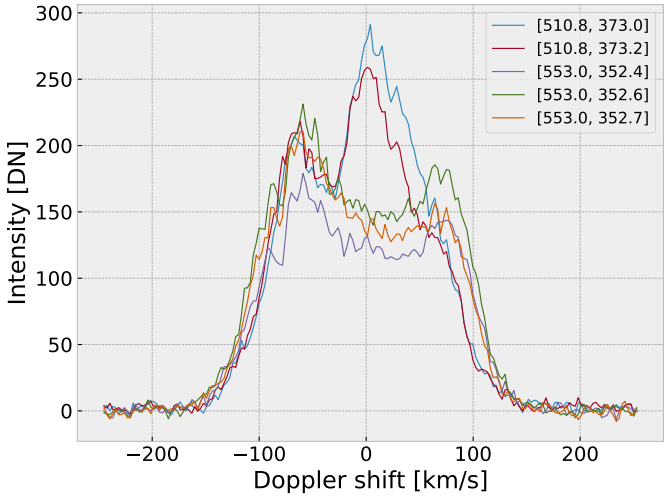

Figure 5.17: Observational profiles of Si IV 1394 Å that fall under cluster 119. The Solar coordinates of the observed profiles are given in the legend.

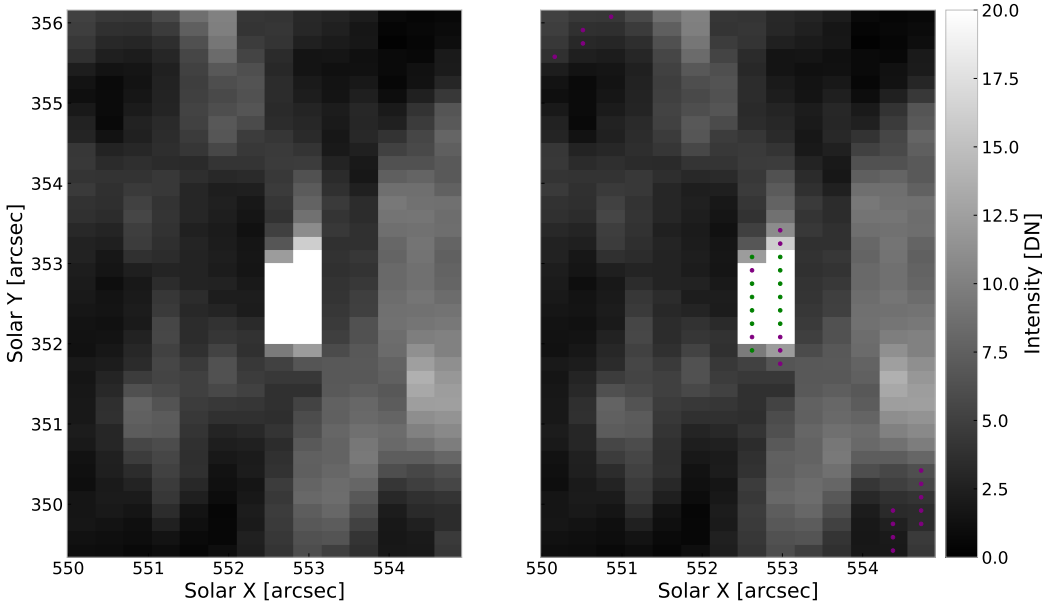

Figure 5.18: Close-up raster on the scaled total intensity Si IV 1394 Å profile with a Ni II blend feature. See caption in figure 5.8 for details on colored dots.

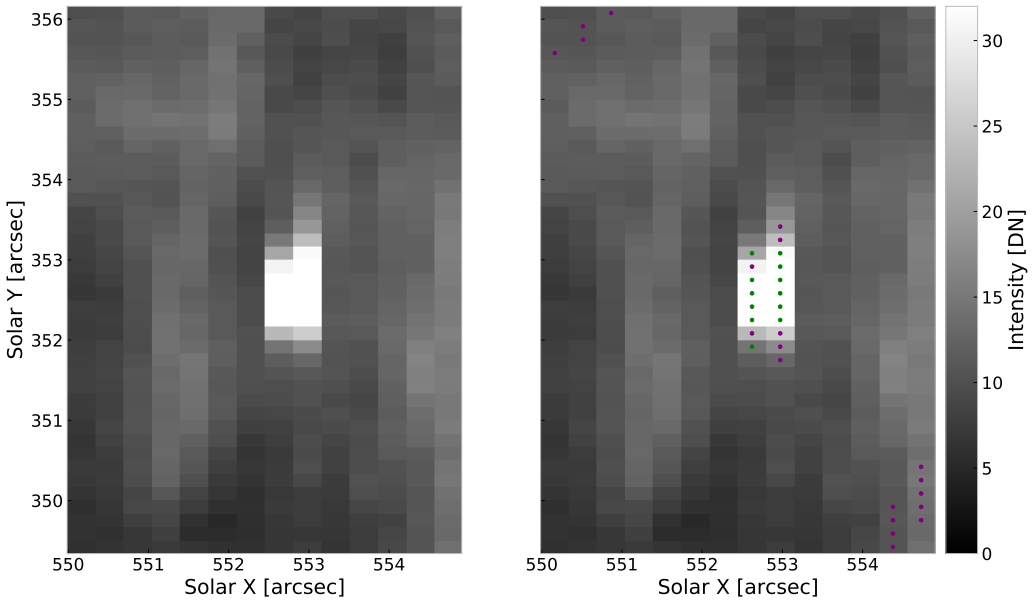

Figure 5.19: Close-up raster on the scaled total intensity C II 1336 Å profile with a Ni II blend feature. See caption in figure 5.8 for details on colored dots.

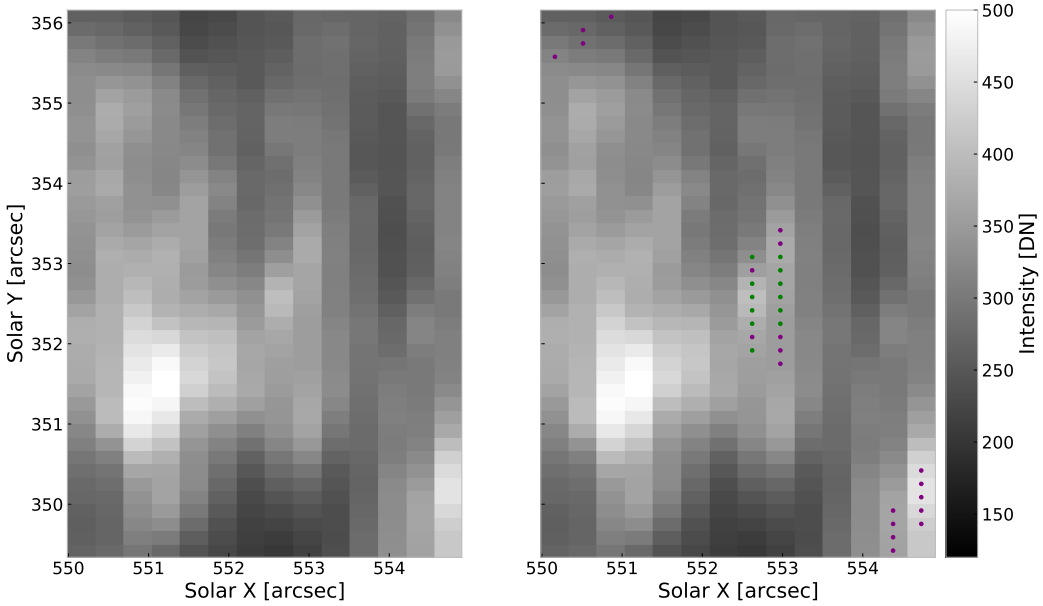

Figure 5.20: Close-up raster on the scaled total intensity Mg k II 2797 Å profile with a Ni II blend feature. See caption in figure 5.8 for details on colored dots.

From the Si IV 1394 Å raster 5.18, the event is observed as small-scale intense brightening and localized next to a network of strong Si IV 1394 Å lines as seen from figure 5.8. The strong brightenings are also evident in upper chromosphere in figure 5.19 but absent in mid chromosphere in figure 5.20. Similar kind of profiles with Ni II line blends are generally reported to be observed in condition such as Solar AR UV bursts (Rouppe van der Voort et al. 2017). The finding here in a CH environment is unique. The Ni II blend is faint possibly due to comparatively small abundance of the cool Ni II λ 1393.33 Å ion in the CH atmosphere as compared to the AR. Similar events with the blend were found within other clusters for certain observed profiles. The Ni II blend were not as easy to locate from the data due to being quite weak compared to reported blends in AR and also the RPs being averaged over the cluster so that blends were smoothed out.

RP 119 form a complex non-Gaussian shaped profile that fits the characteristics of an UV burst. Additional spectral lines of C II and Mg II, averaged over the locations corresponding to the same cluster can be seen in figure 5.21. All lines are compared to the FOV average profile plotted as a dashed line. The upper row show Si IV 1394 Å and Si IV 1403 Å line respectively and the prior has a intensity approximately twice the amount of the latter. The lines can be considered as optical thin if the ratio is $\approx$ 2 (Peter et al. 2014). The computed 1394 Å / 1403 Å ratio is 1.71 which conclude a separation from 2 and therefore suggests deviation from optical thin formation of the lines at the locations of potential UV bursts. Note that RPs are computed as mean of profiles within that particular cluster, implying the conclusion of optical thickness of

RPs must be considered with caution.

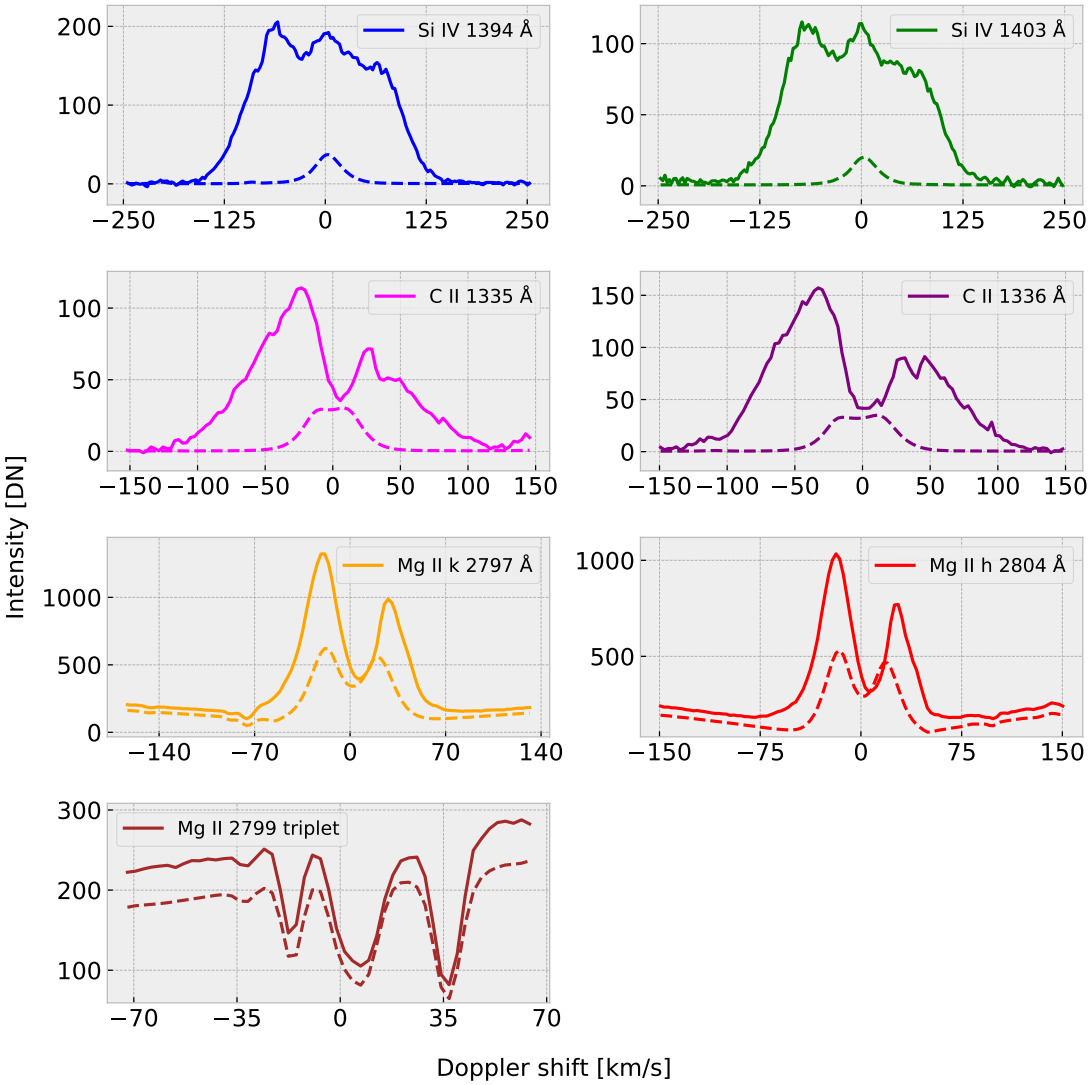

Figure 5.21: RP 119 of Si IV 1394 Å and Si IV 1403 Å at the upper row respectively as solid lines. The second and up to the fourth row show solid line as the mean of the profiles within that particular cluster. The name of the profiles are given in the label. The dashed line in every subplot are the FOV average profile for reference. Doppler shift is given in units relative to their respective rest wavelengths in vacuum from table 3.2.

The strong excess emission peaks for the Si IV lines in figure 5.21 at ±60 km/s indicate the presence of strong small-scale bi-directional flows in the TR. The complexity of the Si IV profiles can possibly be explained by having three components that contribute to the structure of the emission line. The initial component is the background

component that produce the emission peak near 0 km/s. The other components lies at − 60 km/s and 60 km/s indicating an upflow and downflow components respectively. Among the three components, the upflow is the strongest and indicate that the motion of plasma is outwards toward the higher less dense atmospheric layers, i.e. into the corona. The downflowing component indicate that plasma flow inward to the lower, more dense regions which give its remarks in the C II and Mg II h&k spectral lines in the chromosphere. The absorption core of all the C II and Mg II are ≈ 6 km/s shifted which imply a downward motion of the plasma in the chromospheric layers. The enhanced emission lines from the mid chromosphere to the TR give indication of intermittent heating of the atmosphere where an UV bursts event is evident. The peak separation of the C II and Mg II h&k profiles are increase by a factor ≈ 2 compared to the FOV average profile which suggests excessive heating for the mid to upper chromosphere. The low chromospheric Mg II triplet profile do not show excess enhancement of the line which suggest that the origin of the UV burst event is possibly originated at the high chromosphere or TR.

5.4.2 Case 2: RP 117; Enhanced Mg II Triplet Line

Cluster 117 is the only cluster that have an emission feature of the Mg II triplet lines near 2798.75 Å. Pereira et al. (2015) used numerical modelling to produce synthetic Mg II triplet profiles and they concluded that emission in these lines are rare and are possibly caused by rapidly increase of temperature in chromosphere where the chromospheric temperature is minimum. Figure 5.22 showcase the other profiles within cluster 117. As seen from the Mg II 2799 triplet panel, there are evidence of enhanced emission at ≈ 21 km/s which should indicate a steep temperature rise in the low chromosphere. There is also evidence of strong enhanced emission in the other chromospheric and TR lines as shown in other panels from the figure. This suggests intermittent heating through the lower chromosphere to TR at the locations of such profiles as compared to the FOV average profile.

The Si IV lines are bright and enhanced in both wings with multiple peaks. As both wings are enhanced with distinct peaks, the plasma at the TR is considered to have bi-directional flow. The red peak component is ≈ 31 km/s Doppler shifted whilst the blue peak component are ≈ −45 km/s Doppler shifted. The C II lines are blue shifted at ≈ −3 km/s which indicate an upflow plasma in the high chromosphere region. The core of Mg II h%k profiles are approximately unshifted which imply a mass flow to be dominant in the upper chromosphere and TR only whilst no indication of plasma flow in mid chromosphere. Signatures of heating are observed in all spectral profiles as the emissions are excessively enhanced. In addition, the peak separation of C II and Mg II h&k profiles are increased by a factor of ≈ 2 compared to the FOV average profile. The location of such profiles conclude that UV bursts events can potentially be associated with intermittent heating from the low chromosphere to TR.

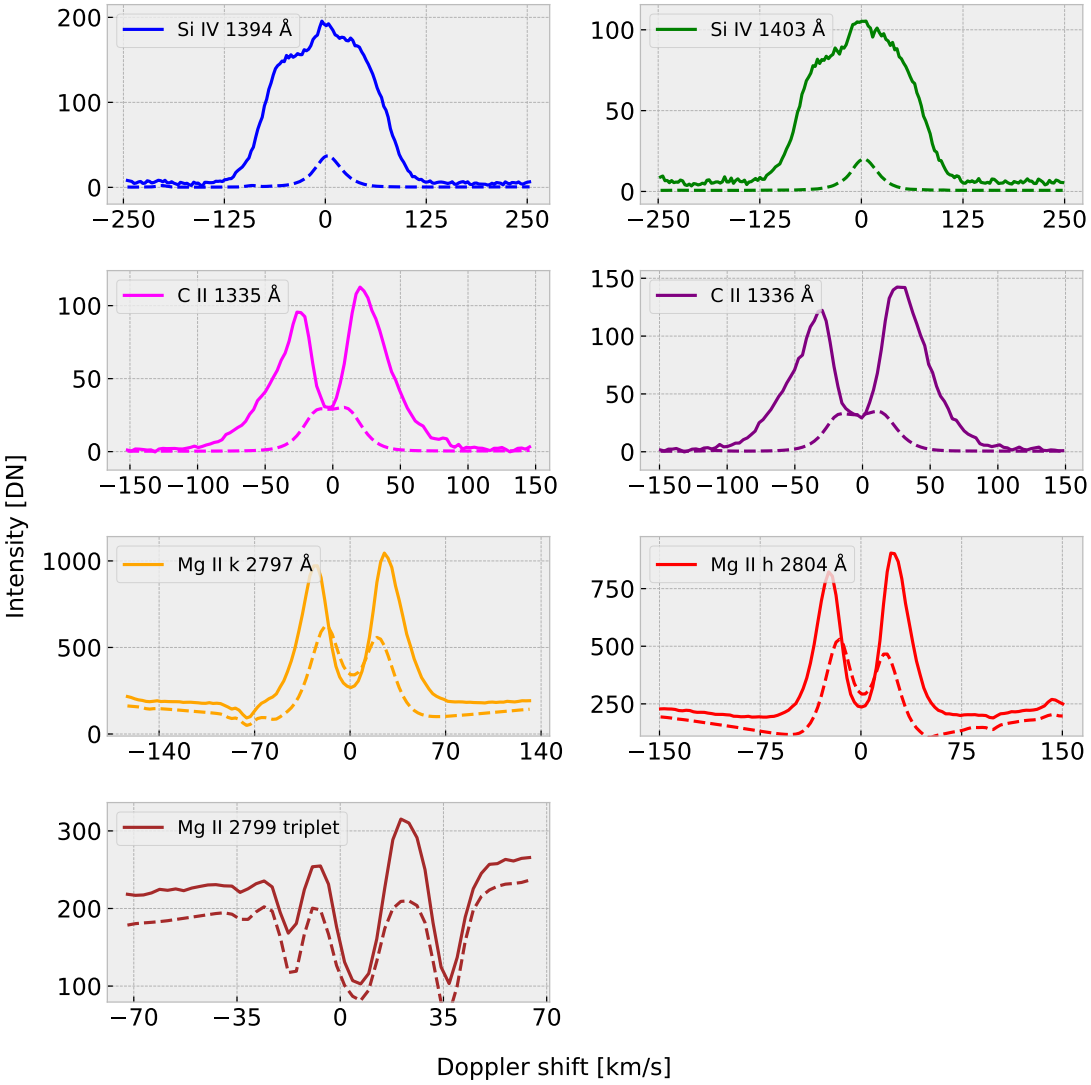

Figure 5.22: RP 117 of Si IV 1394 Å and Si IV 1403 Å at the upper row respectively as solid lines. The second and up to the fourth row show solid line as the mean of the profiles within that particular cluster. The name of the profiles are given in the label. The dashed line in every subplot are the FOV average profile for reference. Doppler shift is given in units relative to their respective rest wavelengths in vacuum from table 3.2.

Chapter 6

Summary

The Solar atmosphere consists of many different types of dynamic signatures and one type of interests are small-scale dynamic processes that potentially give rise to complex spectral lines. This study focused on to bring forth Si IV 1394 Å and Si IV 1403 Å spectral lines in a CH data-set that deviate from the typical Gaussian-profile. Such profiles were detected and localized using machine learning techniques. The localization suggested interesting features in the atmospheric condition, such as events like compact brightening and a hot jets of plasma. The TR Si IV RPs that match the characteristic of UV bursts were studied in lower to upper chromosphere by using C II 1335 Å and C II 1336 Å, and Mg II k 2797 Å, Mg II h 2804 Å and Mg II triplet 2799 Å profiles. It was observed evidently that certain events were potentially prevalent in multiple atmospheric layers. The following section summaries the main findings in this study.

6.1 Profile Clustering

The implementation of PCA improved the k-means clustering algorithm to classify the features of interest in a scientifically motivated way. The PCA was implemented to remove the most significant component of the Si IV 1394 Å and Si IV 1403 Å lines, i.e. remove the component that form the typical single Gaussian-structure of the emission lines. Without this component, the k-means clustering algorithm did in fact enhance the frequency of RPs that fit the characterization of UV-bursts in the data-set. If the k-means clustering were directly applied on the despiked data without implementing PCA, the algorithm was able to cluster the data-set into a fixed number of clusters, but only a few RPs were able to fit the characteristics of UV bursts. This comparison motivated the implementation of PCA in this study. Therefore, in conclusion, the combination of k-means clustering and PCA is a strong numerical tool to identify asymmetric lines in spectral data.

6.2 UV Bursts in a Coronal Hole

UV bursts are typically found and identified in Solar AR as small compact brightenings as summarized by Young et al. (2018). With the techniques implemented in the present thesis, it was shown that UV bursts-like events are also present in a Solar CH environment, though not as frequent as in AR. The RPs that correspond to a UV bursts-like event were ubiquitous in the CH data-set. In total 2.6 % of the data-set were characterised as a UV bursts type of event.

Si IV 1394 Å and Si IV 1402 Å profiles with evident multiple peaks were found for several clusters. Signatures of bi-directional strong flows over the rest-background were evident from such profiles. These findings were ubiquitous in the FOV but evidently quite rare. Profiles with no clear multiple peaks but instead clear enhanced and extended wings were dominant in the population of UV bursts. This include enhancement in red wing, blue wing or both. Such profiles also show signatures of small-scale bi-directional mass flow over the rest background when enhancement in both wings were prevalent. Additionally, if only one wing were enhanced, the mass flow upward or downward over the rest background could be concluded.

The general conclusion is that UV bursts profiles were mostly associated with network jet or compact brightenings observed in the Si IV 1394 Å raster map. By simultaneously studying the Si IV 1394 Å and Si IV 1403 Å in the TR, C II 1335 Å and C II 1336 Å in the upper chromosphere, Mg II k 2797 Å and Mg II h 2804 Å in the mid chromosphere and finally Mg triplet 2799 Å between h&k lines in low chromosphere, UV bursts were generally found to be associated with enhanced heating from the mid chromosphere to TR. Heating signatures in the low chromosphere was rarely found over the studied FOV.

6.3 Future Work

This study focused on locating UV bursts events on a single spectral raster of a non-AR atmosphere. The statistical results were clear that UV bursts-type of events are prevalent in a CH atmosphere but the life time of such events were not possible to estimate. To expand on the statistical analysis of such events in a CH atmosphere, a narrow, fast raster or possibly a sit-and-stare observation could be added to the study. The evolution of UV bursts in the low chromosphere to the TR on such atmosphere would be an interesting addition to the study.

This study focused only on raster data obtained by IRIS. In addition to the raster data, IRIS also provides near-simultaneous SJI data which add context of physical processes at and near the observed spectral UV bursts event. As the analysis already is done on the spectral raster from the chromosphere to the TR, it is only natural to add such diagnostics to the expansion of this study.

Another addition to the spectral raster, is to include coordinated observation from other instruments, such as the Swedish 1-m Solar Telescope (Scharmer et al. 2003). This telescope can provide high-quality context from the photosphere and low chromosphere

to the studied IRIS data. Such coordination could add insight on reported potential correlation between photospheric and chromospheric events such as spicules and Ellerman bombs with the TR energetic UV bursts events, and also possibly magnetic field evolution (Rouppe van der Voort et al. 2017).

The AIA/SDO observatory (Lemen et al. 2012) provide full scale imagery in various passbands which could provide more context to the study. Using these data could proved study of the response from UV bursts spanning from the photosphere to the chromosphere. The resolution is nevertheless at a lower level. To add magnetic field observations during a UV burst event, the Helioseismic and Magnetic Imager (HMI) from SDO (Scherrer et al. 2012) could also be provided to increase the understandings of photospheric magnetic field evolution.

List of Figures

Bibliography

Arthur, D., & Vassilvitskii, S. 2007, in , 1027–1035, doi: 10.1145/1283383.1283494

Bose, S., Henriques, V. M. J., Joshi, J., & Rouppe van der Voort, L. 2019, A&A, 631, L5, doi: 10.1051/0004-6361/201936617

Brueckner, G. E., & Bartoe, J. D. F. 1983, ApJ, 272, 329, doi: 10.1086/161297

De Pontieu, B., Title, A. M., Lemen, J. R., et al. 2014, Sol. Phys., 289, 2733, doi: 10.1007/s11207-014-0485-y

De Pontieu, B., Polito, V., Hansteen, V., et al. 2021, Sol. Phys., 296, 84, doi: 10.1007/s11207-021-01826-0

Dere, K. P., Bartoe, J. D. F., & Brueckner, G. E. 1989, Sol. Phys., 123, 41, doi: 10.1007/BF00150011

Forgy, E. W. 1965, Biometrics, 21, 768

Golub, G., & Kahan, W. 1965, Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2, 205

Greisen, E. W., & Calabretta, M. R. 2002, A&A, 395, 1061, doi: 10.1051/0004-6361:20021326

Géron, A. 2017, Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems (Sebastopol, CA: O'Reilly Media)

Halko, N., Martinsson, P.-G., & Tropp, J. A. 2009, arXiv e-prints, arXiv:0909.4061. https://arxiv.org/abs/0909.4061

Huang, Z., Madjarska, M. S., Scullion, E. M., et al. 2017, MNRAS, 464, 1753, doi: 10.1093/mnras/stw2469

Kerr, G. S., Simões, P. J. A., Qiu, J., & Fletcher, L. 2015, A&A, 582, A50, doi: 10.1051/0004-6361/201526128

Kleint, L., & Panos, B. 2022, A&A, 657, A132, doi: 10.1051/0004-6361/202142235

Leenaarts, J., Pereira, T. M. D., Carlsson, M., Uitenbroek, H., & De Pontieu, B. 2013a, ApJ, 772, 89, doi: 10.1088/0004-637X/772/2/89

—. 2013b, ApJ, 772, 90, doi: 10.1088/0004-637X/772/2/90

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, Sol. Phys., 275, 17, doi: 10.1007/s11207-011-9776-8

Lloyd, S. 1982, IEEE Transactions on Information Theory, 28, 129, doi: 10.1109/TIT.1982.1056489

Nóbrega-Siverio, D., Guglielmino, S. L., & Sainz Dalda, A. 2021, A&A, 655, A28, doi: 10.1051/0004-6361/202141472

Pearson, K. 1901, LIII. On lines and planes of closest fit to systems of points in space, Zenodo, doi: 10.1080/14786440109462720

Pereira, T. M. D., Carlsson, M., De Pontieu, B., & Hansteen, V. 2015, ApJ, 806, 14, doi: 10.1088/0004-637X/806/1/14

Peter, H., Tian, H., Curdt, W., et al. 2014, Science, 346, 1255726, doi: 10.1126/science.1255726

Phillips, K. J. H., Feldman, U., & Landi, E. 2008, Ultraviolet and X-ray Spectroscopy of the Solar Atmosphere, Cambridge Astrophysics (Cambridge University Press), doi: 10.1017/CBO9780511585968

Podgorski, W. A., Cheimets, P. N., Golub, L., Lemen, J. R., & Title, A. M. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8443, Space Telescopes and Instrumentation 2012: Ultraviolet to Gamma Ray, ed. T. Takahashi, S. S. Murray, & J.-W. A. den Herder, 84433D, doi: 10.1117/12.925491

Priest, E. 2014, A Description of the Sun (Cambridge University Press), doi: 10.1017/CBO9781139020732.002

Rathore, B., & Carlsson, M. 2015, ApJ, 811, 80, doi: 10.1088/0004-637X/811/2/80

Rathore, B., Carlsson, M., Leenaarts, J., & De Pontieu, B. 2015a, ApJ, 811, 81, doi: 10.1088/0004-637X/811/2/81

Rathore, B., Pereira, T. M. D., Carlsson, M., & De Pontieu, B. 2015b, ApJ, 814, 70, doi: 10.1088/0004-637X/814/1/70

Rouppe van der Voort, L., De Pontieu, B., Scharmer, G. B., et al. 2017, ApJ, 851, L6, doi: 10.3847/2041-8213/aa99dd

Rouppe van der Voort, L. H. M., Joshi, J., Henriques, V. M. J., & Bose, S. 2021, A&A, 648, A54, doi: 10.1051/0004-6361/202040171

Rouppe van der Voort, L. H. M., Rutten, R. J., & Vissers, G. J. M. 2016, A&A, 592, A100, doi: 10.1051/0004-6361/201628889

Rutten, R. J. 2003, Radiative Transfer in Stellar Atmospheres

Scharmer, G. B., Bjelksjo, K., Korhonen, T. K., Lindberg, B., & Petterson, B. 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4853, Innovative Telescopes and Instrumentation for Solar Astrophysics, ed. S. L. Keil & S. V. Avakyan, 341–350, doi: 10.1117/12.460377

Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, Sol. Phys., 275, 207, doi: 10.1007/s11207-011-9834-2

Tian, H., Xu, Z., He, J., & Madsen, C. 2016, ApJ, 824, 96, doi: 10.3847/0004-637X/824/2/96

Young, P. R., Tian, H., Peter, H., et al. 2018, Space Sci. Rev., 214, 120, doi: 10.1007/s11214-018-0551-0