Needles in a Haystack: Advanced Statistical Techniques and Large Stellar Spectroscopic Datasets

By

Arvind Hughes

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Physics
November 2017

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Arvind Hughes

Acknowledgements

First of all, I would like to thank my supervisors, Lee Spitler, Daniel Zucker and Chengyuan Li. They provided invaluable support and direction throughout the year, helping me develop my skills as a researcher in the natural sciences. Lee, thank you for starting me off down this path, we didn't search for strange signals, but maybe next time. Lee, I would also like to thank you for your support and encouragement, I don't think I would have been able to write this thesis without your belief in me. Dan, thank you for introducing me to the fascinating world of stellar spectroscopy and metal-poor stars, a welcome change from the property market. I can see myself as a stellar spectroscopist in the future. Finally, Chengyuan thank you for backing the methodology and believing I can go further!

Secondly my fellow Masters students, housemates and friends thank you for the many thesis distractions throughout the year. In particular I would like to thank the 'Boy0s', Connor and Oliver, thank you for the many memes sent, games played and snacks eaten. I suppose I should also thank my desk neighbour and fellow astro partner, Adam, for always answering my random questions.

Lastly, I would like to thank my family and my partner Liv. To Mum, Dad and the Bros, thank you for the constant encouragement, keeping me sane and for the fortnightly movie nights. Liv, thank you for your love, support and numerous library sessions throughout your own thesis year. You helped me stay motivated, and never stopped believing in me.

Abstract

With the development of advanced astronomical instruments, many survey teams are producing datasets that are too large for traditional analysis. Thanks to recent improvements in computing and statistical methods, it is now possible to extract information more efficiently.

In this thesis, data from the GALactic Archaeology with HERMES spectroscopic survey (GALAH), is used to show how machine learning methods can identify rare but interesting stars. This thesis tests a new methodology that employs the t-SNE dimensionality reduction technique with the clustering method, HDBSCAN, and a new tool developed by the researcher, the t-SNE Visualiser. This method was applied to $\sim 200,000$ stars in the GALAH dataset with the aim of detecting extremely metal-poor stars and Solar twins.

Applying this approach lead to the discovery of 66 possible extremely metal-poor stars and 20 Solar twin candidates. A verification of the success of the new methodology is also presented.

Contents

Somewhere, something incredible is waiting to be known.

Carl Sagan

1 Introduction

Astronomy and statistics have a relationship dating centuries. This introductory chapter provides a brief outline of this history and introduces machine learning, with a focus on the field of dimensionality reduction as applied to spectroscopic data. A brief description of the target astronomical objects in this thesis will be presented.

1.1 Modern Astronomy

Astronomy and Statistics have an interwoven history and,'perhaps more than other physical sciences, astronomy is frequently statistical in nature' (Feigelson, 2009). Ancient civilisations carried out many fundamental quantitative measurements of celestial phenomena. According to Plackett (1958), the Greek astronomer Hipparchus developed one of the earliest statistical methods, that of arithmetic mean and variance. By using measurements of the duration of the day and the interval between solstices, Hipparchus estimated the variability in the duration of the day by using half the range of his observations.

Challenges in astronomy led early researchers to develop new statistical methodologies. Statistics in the 18^{th} century was considered the collection and compilation of data but over time became focused on the development of mathematical methods by which to analyse and interpret data. The late 18^{th} and 19^{th} century saw the rise of the statistical method in astronomy, through mathematical pioneers such as Johann Carl Friedrich Gauss and Pierre-Simon Laplace. Both Gauss and Laplace considered the method of least squares in a probabilistic framework and proved that it was the best method for finding orbital parameters from astronomical observations (Feigelson, 2009). Thus, least squares rapidly became the principal tool linking astronomical observations with celestial mechanics. Gauss additionally showed methods to handle observational measurement errors, and introduced his famous Gaussian, or "Normal" distribution. For much of the 19^{th} century this distribution was known as the "astronomical error function" (Gauss & Stewart, 1995).

The first few decades of the 20^{th} century saw diminished links between astronomy and statistics research, as statisticians became more focused on issues such as social behaviour and life insurance (Feigelson, 2009). At the same time a portion of astronomy became centred around theoretical astrophysics rather than observational astronomy. This shift to a more theoretical approach may be to

Survey	Volume
Two Micron All-Sky Survey	10 TB
Sloan Digital Sky Survey	40 TB
SkyMapper Southern Sky Survey	500 TB
SKA Survey	4.6 EB (Predicted)

Table 1.1: Current Data Volumes of Astronomical Surveys

attributed to the work of prominent physicists of the era, such as Albert Einstein with General Relativity (Einstein, 1916), and Max Planck and his solution to the black body radiation problem (Planck, 1901).

For the observational astronomer in the early 20th century the primary statistical tool was still the method of least squares, with the most well known example of an observationally-derived linear relationship being Hubble's Law (Hubble, 1929) for the expansion of the universe. Hubble's Law is governed by a linear equation between recessional velocity of external galaxies and proper distance. Two significant statistical advances, Bayes Theorem by Thomas Bayes (Bayes & Price, 1763) and Maximum Likelihood Estimation by Ronald Fisher (Fisher, 1922) were sporadically applied in astronomy throughout the 20th century with maximum likelihood having a major impact on areas such as image restoration (Lucy, 1974) and the calculation of the galaxy luminosity function in extragalactic astronomy (Efstathiou et al., 1988). Nonetheless, the diverse statistical methods developed in the 20th Century remained sparsely used in astronomy.

Early in the 21st century, however, astronomy gained prominence as one of the main observationally driven sciences. Astronomy has seen dramatic developments in the quantity of data available due to the construction of detailed surveys using the rapid advancement in astronomical instruments and data storage capacities. Data-driven disciplines are more and more faced with the problem of how to store, organise, use, and interpret the enormous amounts of data being generated by new research infrastructure (Szalay & Gray, 2001). The development of both ground- and space-based instrumentation spanning the electromagnetic spectrum from gamma rays to radio is taking astronomy into the era of "big data" science. Table 1.1 gives and indication of the quantity of data being acquired in these targeted surveys. The Sloan Foundation 2.5m Telescope, Hubble Space Telescope and the High Efficiency and Resolution Multi-Element Spectrograph (HERMES) are examples of advanced engineering used to generate large surveys, such as the Sloan Digital Sky Survey (SDSS) (Albareti et al., 2016), the Cluster Lensing And Supernova survey with Hubble (Postman et al., 2012) and the GALactic Archaeology with HERMES survey (GALAH) (De Silva et al., 2015), respectively. The SDSS in particular, is a well known and highly successful example of one of the first "big data" endeavours in modern astronomy. The SDSS group directly anticipated data storage and access requirements and very successfully deployed infrastructure for astronomers around the world to do "big data" analysis with its data products. The current state of the statistical method in astronomy is somewhat in flux, as many studies still use well established traditional methods such as least squares, χ^2 fitting or principal component analysis when more accurate methods, especially for coping with significantly large datasets, are available. However, this push into "big data" has seen the astrostatistical methodology grow rapidly and emerge as an active area of research, with an increase in the application of more precise statistical methods such as artificial neural networks (e.g., Lahav, 1996; Collister & Lahav, 2004; Yeche et al., 2010; Eatough et al., 2010) and Bayesian modelling (e.g., Benítez, 2000; Park et al., 2006; Trotta, 2008; Gagné et al., 2014).

1.1.1 Galactic Archaeology and GALAH

The survey data analysed in this thesis were obtained by the GALactic Archaeology with HERMES (GALAH) survey collaboration, comprising members from different universities across Australia, the

Australian Astronomical Observatory (AAO), and institutions around the world . The science goal of GALAH is to produce a comprehensive view of the formation and the evolution of the Milky Way (De Silva et al., 2015). GALAH aims to achieve this by obtaining $\sim 1,000,000$ high - resolution stellar spectra ($\Delta\lambda \sim 28,000$) for elemental abundance analysis using the HERMES instrument on the 3.9 m Anglo-Australian Telescope (AAT), fed by the 2dF fibre positioner system. GALAH is targeting a 0.05 dex accuracy for elemental abundance, and to achieve this goal requires a signal to noise ratio of ~ 100. This effectively limits the magnitude of stars able to be observed to $12 < V < 14$ towards the Galactic plane. Given these restrictions, the final GALAH survey sample is expected to contain approximately 77% thin disk stars, 22% thick disk stars, 0.8% bulge stars and 0.2% halo stars (Martell et al., 2017). By tracing chemically distinct groups of stars, the aim is to learn more about the history and assembly of our galaxy by reconstructing its original stellar substructures. Surveys like GALAH will produce enormous amounts of data, on orders of magnitude more than could be analysed by survey teams using traditional methods. A more efficient and scalable approach is needed to extract new astrophysical constraints from these survey outputs.

1.2 Machine Learning

In order to obtain useful and meaningful science from very large survey datasets, more optimised and efficient data analytical methodologies are essential. Methods that require minimal or no human interaction (Babu & McDermott, 2002). Machine learning is the science of creating algorithms and techniques that can learn rules from data, adapt to the needs of the user and improve with use.

Machine learning methods are composed of three major components:

1. The model
2. The parameters
3. The learner algorithm

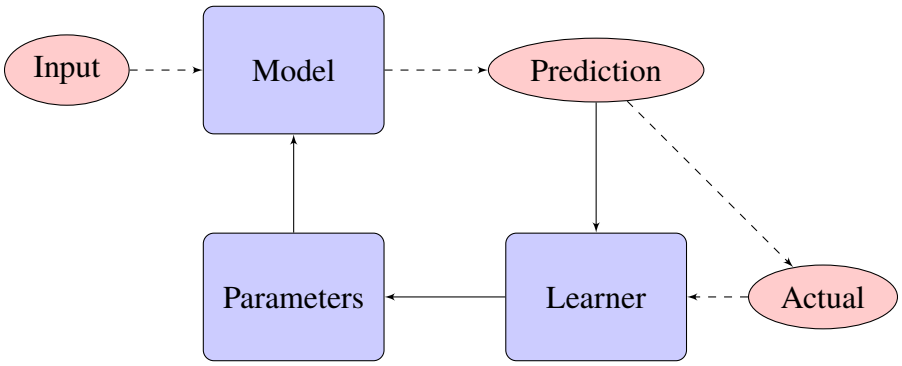


Figure 1.1: Flow Diagram of a typical machine learning process. Blue defines the system and pink the data. The dashed and solid lines represent the flow of feedback/inputs in to the system and the system iterations respectively.

A *learner* is the system that adjusts the parameters in a model by assessing differences in predictions versus actual outcomes. Machine learning algorithms are applied to two types of problems: *classification*, which predicts discrete categorical variable outcomes (a category such as type of animal), and *regression*, which involves estimating or forecasting continuous numerical real valued responses (the value of a house). Classification uses pattern recognition to label objects that have yet to be classified whereas regression uses a mathematical expression to estimate the response. A *classifier* is a function that maps a *feature vector* (a vector that contains the characteristics of a given object), into a discriminant vector containing likelihoods that the objects belong to the different considered classes.

The algorithms may be classified as *supervised learners* and *unsupervised learners*. Supervised learning uses *training data* (data that have already been classified) to determine a relationship between

the input data and the desired outputs. Algorithms in this group rely on having some prior knowledge via the training data, that was given, usually by a user or some previously classified sample. Expressed formally, let $(X, Y) = (x_1, ...x_n, y_1, ...y_n)$ be a set of n points, where $x_i \in X$ & $y_i \in Y, \forall i \in 1, \ldots, n$. The aim of supervised learning is to learn a mapping from x to y, where $y_i \in Y$ are defined as the labels or targets of x_i. This task is well-defined, as a mapping can be evaluated through its predictive performance on *test samples* (random slices of training data purposefully withheld from learning). If y_i is discrete the task is known as classification and if y_i is continuous then the task is defined as regression.

Unsupervised learning does not utilise training data, but rather attempts to infer structure in the data by clustering based on relationships among variables. Without any prior knowledge of the data , these algorithms are designed to explore the data themselves and uncover underlying relationships. Unsupervised learning is inherently more difficult than supervised learning, but can potentially lead to more unique results as these algorithms may find relationships in the data that a human may not have seen and/or considered.

Having described the basic method of machine learning, attention must be brought to the input. Under the assumption that the data have been processed correctly and have errors within accepted bounds to ensure that any model built is an accurate representation, the issue of the dimension of the variables describing the objects in the data has to be addressed. The so-called *curse of dimensionality* arises when analysing and organising data with a large number of variables – the idea being, with regards to machine learning, that when each object in a dataset has numerous possible free parameters, the amount of training data required to ensure each combination is covered is large. There are limits to machine learning irrespective of computational resources; indeed the predictive power reduces as dimensionality increases, which is known as the Hughes phenomenon (Hughes, 1968). Thus having more variables describing the object may decrease the performance of the machine learning algorithm. There is no definite answer to variable selection when developing a statistical model which results in multiple possible solutions to the problem being addressed.

There have been many attempts to apply machine learning to astronomical data (e.g., Boroson & Green, 1992; Trager et al., 1998; Houck & Denicola, 2000; Bennett et al., 2013). This thesis focuses on dimensionality reduction and the application of clustering, because, as astronomical datasets are growing rapidly, the methodology required to obtain meaningful scientific results on these large datasets efficiently is a key area of research. Dimensionality reduction is often applied as a first step in the machine learning process to not only alleviate the curse of dimensionality but also to improve computational time and model accuracy.

1.3 Statistical Methods

Statistical methods are a foundation for machine learning algorithms. The following section outlines two techniques in the area of dimensionality reduction used in astronomy, with a detailed description of t-distributed stochastic embedding (t-SNE). Clustering methods are often used in conjunction with dimensionality reduction to quantitatively classify the resulting visualisation. The clustering method of Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is described in detail.

1.3.1 Dimensionality Reduction

At a fundamental level, machine learning is not well understood, not because it is too complicated, but rather because humans have evolved to reason in 2 and 3 dimensions, stretching to 4. Machine learning however, often deals in very high dimensions,which complicates simple concepts. Obtaining meaning from these higher dimensions directly is often fruitless. Therefore the field of dimensionality

reduction was developed, which explores techniques for translating high dimensional data into lower dimensional data and methods for visualising these transformed data.

Astronomical datasets from modern surveys (e.g. SDSS) are both large in size and in dimensionality. Any statistical methods applied therefore need to be time efficient and scalable. One approach is to reduce the number of dimensions considered. The increasing sparseness of data in higher dimensions means geometric-based algorithms that compute distances between points become highly unstable. It is therefore often essential to reduce the dimensionality of a data set prior to classification. Dimensionality reduction can lead to more reliable outcomes when using geometric tree-based methods to give a generalised mapping on account of the reduced number of parameters. Classification with complete spectra can produce good results and dimensionality reduction may be essential in applications when data transmission rates from space based observations are limited.

1.3.1.1 Definitions

Before describing a selection of dimensionality reduction techniques some background statistical definitions are required.

Dimensionality: The dimensionality of a dataset, usually described by d, is the minimum number of variables required to describe a point within a space. For example, if we were interested in describing people in terms of their height and weight, our "people" dataset would have 2 dimensions. If instead we had a dataset of stellar spectra, and each spectrum consisted of a million pixels, then the dimensionality of the dataset would be a million. In many modern machine learning applications, the dimensionality of a dataset is often on the order of tens of thousands.

Manifold: A manifold is an object of dimensionality d that is embedded in some higher dimensional space. The most commonly used example is that of Figure 1.2, a Swiss roll in 3 dimensions, which is a common unfolding problem in dimensionality reduction.

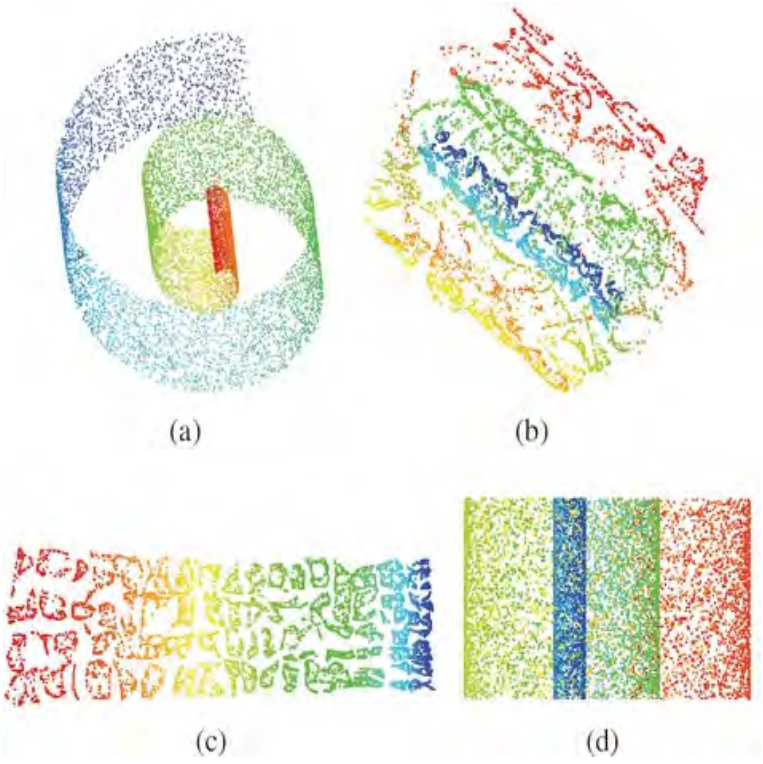


Figure 1.2: (a) The Swiss Roll data set (b)-(d) Unfurling using dimensionality reduction methods. Adapted from (Paulovich et al., 2011)

Embedding: The 2-dimensional surface of a 3-dimensional space is contained within the 3-dimensional

space. This 2-dimensional surface is an embedding of the higher 3-dimensional space (i.e a subgroup).

The above definitions may be formally expressed: Let $M \subset \mathbb{R}^{\mathbb{N}}$ to be a d dimensional manifold that is embedded in an N dimensional space and the explicit mapping between the two is $f : \mathbb{R}^d \rightarrow \mathbb{R}^N$ where $x_i \in \mathbb{R}^N$ are the data points.

1.3.2 Dimensionality Reduction Techniques

Having defined three fundamental terms in the area of dimensionality reduction the following section describes the techniques of Principal Component Analysis, Local Linear Embedding and t-distributed Stochastic Neighbour Embedding.

1.3.2.1 Principal Component Analysis

Prior to the rise of "big data" Principal Component Analysis (PCA) was the standard method for achieving a dimensionally reduced data set. PCA is a form of feature selection, where the input vector is known as a feature vector. PCA is a method of extracting the key features (components) from a large N-dimensional dataset.

PCA deconstructs the N-dimensional dataset into eigenvalues and their perpendicular eigenvectors, where the eigenvalues detail the variance of the data and the eigenvector the direction of the eigenvalue. PCA yields a linear transformation of the originally higher dimensional data by linear combinations of the few eigenvectors with the highest eigenvalues. The dimensionally reduced data is achieved by removing the low eigenvalues and thus keeping only these principal components. Like all dimensionality reduction techniques, PCA assumes that the data represented in a lower dimensional space are the same as in the higher dimension. Deeming (1964) first proposed the use of PCA in analysing volumes of stellar spectra. However, most subsequent work involving PCA and classification focused on using PCA as a method to pre-process the data fed into artificial neural networks (Lahav et al., 1996). A significant drawback of PCA, despite its simplicity, is its tendency to fail when applied to very high dimensional datasets; these then become a problem for machine learning techniques.

1.3.2.2 Local Linear Embedding

More recently Local Linear Embedding (LLE) was utilised by Daniel et al. (2011) on SDSS spectra. LLE is a non-linear dimensionality reduction technique and has been shown to be suitable for classification purposes as it is able to take complex, high dimensional spectra and project the spectra on to a low dimensional space. Daniel et al. (2011) showed that LLE was superior to PCA in the case where the underlying structure of the data is a non-linear manifold rather than a simple linear combination. In this instance PCA was shown to ‘wash’ spectral features away, whereas LLE would reduce the data dimensions while keeping the nonlinear structure.

1.3.2.3 t-Distributed Stochastic Neighbour Embedding

The unique dimensionality reduction technique, t-Distributed Stochastic Neighbour Embedding (t-SNE), introduced by Maaten & Hinton (2008), a modification of Stochastic Neighbour Embedding (SNE), is the selected dimensionality reduction method used in this thesis. SNE employs a probabilistic approach to placing objects described by the high dimensional vector space in a low dimensional space. This optimally preserves how objects are defined in a local area of the higher space, its neighbourhood identity.

Stochastic neighbour embedding converts the high dimensional euclidean distances between data points in to conditional probabilities that represent similarities. In this thesis, there are over 200,0000 spectra, which are considered as data points x_i in the high dimensional space X with 12288 wavelength values each.

A data point x_i is defined as similar to the j-th point, x_j (where i and j represent two different stars), by the conditional probability $p_{i|j}$, if nearby points were selected in proportion to their probability density under a Gaussian distribution with variance σ_i^2.

$$p_{j|i} = \frac{exp(-d(x_i,x_j))/2\sigma_i^2}{\sum_{i\neq k} exp(-d(x_i,x_k))/2\sigma_i^2}, \quad p_{i|i} = 0, \tag{1.1}$$

where $d(x_i, x_j)$ is defined as the euclidean distance $||x_i - x_j||^2$

This is similarly defined in the low dimensional space Y for the projected spectra y_i and y_j. Y may be of any dimension but is easily graphically represented in the two dimensional plane as shown in the plots in Chapter 3. However where t-SNE differs from SNE is in this evaluation of the similarity between two points in the lower dimensional space. t-SNE uses a student t-distribution instead of a Gaussian to evaluate the similarity. The student t-distribution is similar to a Gaussian (as it is an infinite mixture of Gaussians) and, most importantly, is computationally much faster to evaluate compared to a Gaussian.

Therefore in the low dimensional embedded space, for the projected spectra y_i and y_j similarity is defined using a heavy tailed distribution and one degree of freedom,

$$q_{i|j} = \frac{(1+ \parallel y_i - y_j\parallel^2)^{-1}}{\sum_{k\neq l}(1+ \parallel y_k - y_i\parallel^2)^{-1}} \tag{1.2}$$

Embedding then attempts to match these two distributions, $p_{i|j}$ and $q_{i|j}$, making x_i and x_j to be correctly matched to their low dimensional projection x_i and x_j by minimizing a cost function, defined as being the sum of KullBack-Leibler divergences,

$$KL(P|Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{1.3}$$

using gradient descent and the Barnes-Hut algorithm (Barnes & Hut, 1986) for a faster cost function approximation. It should be noted that the cost function is not convex and thus multiple runs of t-SNE would reveal different maps. Furthermore the axes on a t-SNE plot have no physical meaning, hence they have been removed from all plots throughout the thesis.

There is one significant free parameter in t-SNE, perplexity, that affects how the projection on to the lower dimensional space is generated. Perplexity is defined as 2 to the power of Shannon Entropy. Perplexity tunes the variance of the probability distribution in Eq. (1.1) and roughly determines how wide the region of comparison is for a given spectrum. The Barnes-Hut algorithm has a single parameter θ that ranges from 0 - 1 which determines the level of accuracy of the cost function and thus the speed, with a value of 0 favouring accuracy but compromising speed. This was left at the default level of 0.5 for all computations.

While LLE is best suited to unfold a single continuous low dimensional manifold, t-SNE will focus on the local structure of the data and will tend to extract clustered local groups of samples. This ability to group samples based on the local structure is beneficial for visually disentangling a dataset that comprises several manifolds at once. t-SNE is a recent method in astronomy, having been used by Traven et al. (2017) to identify problematic spectra in the GALAH survey and to find twin stars in order to calculate parallaxes for stars in the RAVE survey (Matijevič et al., 2017). t-SNE is a tool for data visualisation; it reduces the dimensionality of data to 2 or 3 dimensions so that it may be easily graphically displayed (at least on the 2-dimensional plane). Local similarities are preserved by this embedding and t-SNE converts distances between data in the original space to probabilities. t-SNE

is well suited to the purpose of finding and visualising the distribution of similar spectra in a large dataset. For this reason, t-SNE is the main algorithm employed in this work.

1.3.3 Clustering methods

Clustering methods are typically used in combination with dimensionality reduction to quantitatively classify similar objects. Clustering is the organisation of similar unlabelled data items into groups called clusters. A cluster is a collection of data items which are similar to each other but dissimilar to data items in other clusters. A standard clustering method contains: a distance metric to determine similarity between data points; a convergence criterion function to evaluate whether the clustering has been effective; and an algorithm to optimize the criterion function. There are 3 classes of clustering methods:

1. Hierarchical algorithms find successive clusters using previously established clusters;
2. Partitional algorithms typically determine all clusters at once;
3. Bayesian algorithms try to generate a posteriori distribution over the collection of all partitions of the data.

To identify the stellar classes in this thesis, the algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is utilised.

1.3.3.1 k-means

The most popular clustering algorithm is k-means, developed by MacQueen (1967), which is a partitional clustering algorithm. The k-means algorithm divides the given data into k clusters, where k is specified by the user. Each cluster has a cluster centre called a centroid. k-means starts by choosing k random data points to be the initial centroids; each data point is then assigned to the closest centroid. The centroids are recomputed using the newly assigned cluster members. If a convergence criterion is not satisfied the process is repeated. Key advantages of the k-means algorithm are the ease of implementation and its simplicity. However there are two major disadvantages of k-means. First, the Euclidean distance is used for determining the clustering. This could require the observational data to be transformed to give an appropriate range of values. Second, the number of clusters is specified by the user and thus must be appropriate for the dataset being analysed. Too few or too many groupings can lead to less definitive results.

1.3.3.2 Hierarchical Density-Based Spatial Clustering of Applications with Noise

Ester et al. (1996) introduced density based spatial clustering and application with Noise (DBSCAN) to identify clusters of any shape in a data set containing noise and outliers. DBSCAN, unlike k-means does not require every point be assigned to a cluster and does not partition the data. Rather it identifies dense clusters and leaves the remaining unclustered points as noise. DBSCAN is extremely efficient at clustering data of varying structures, e.g., not just circular clusters. However the non-hierarchical version of DBSCAN struggles to cluster data of varying density and can only provide a flatlabelling of data objects based on a global density threshold. Using a *single* density parameter can often not characterise data sets with clusters of very different densities or nested clusters. HDBSCAN is a clustering method developed by Campello et al. (2013). It is an extension of DBSCAN in that the method allows for varying density clusters. A simplified description of the algorithm adapted from (Campello et al., 2013) is given below;

Consider $Y = \{y_1, ..., y_n\}$ to be the set of n projected spectra and D to be an $n \times n$ matrix containing pairwise distance between y_i and y_j, $d(y_i, y_j)$ for a metric $d(.,.)$:

Min_Samples: A smoothing factor in density estimates, which is the neighbour threshold for a spectrum to become a core point;
Min_Cluster_Size: The criteria determining which clusters are kept when applied to the weighted graph;
Core Distance: The core distance of a spectrum $y_i \in Y$ with respect to min_samples, $d_{core}(y_i)$, which is the distance from y_i to its min_samples nearest neighbour;
Mutual Reachability Distance: The mutual reachability distance between two spectra y_i and $y_j \in Y$ with respect to min_samples, defined as $d_{mreach}(y_i, y_j) = \max\{d_{core}(y_i), d_{core}(y_j), d(y_i, y_j)\}$. Under the metric $d(.,.)$, points with a low core distance (dense) remain the same distance from each other;
Mutual Reachability Graph: A complete graph, $G_{min\_samples}$, in which $y_i \in Y$ are vertices and the weight of each edge is the mutual reachability distance between y_i and y_j.

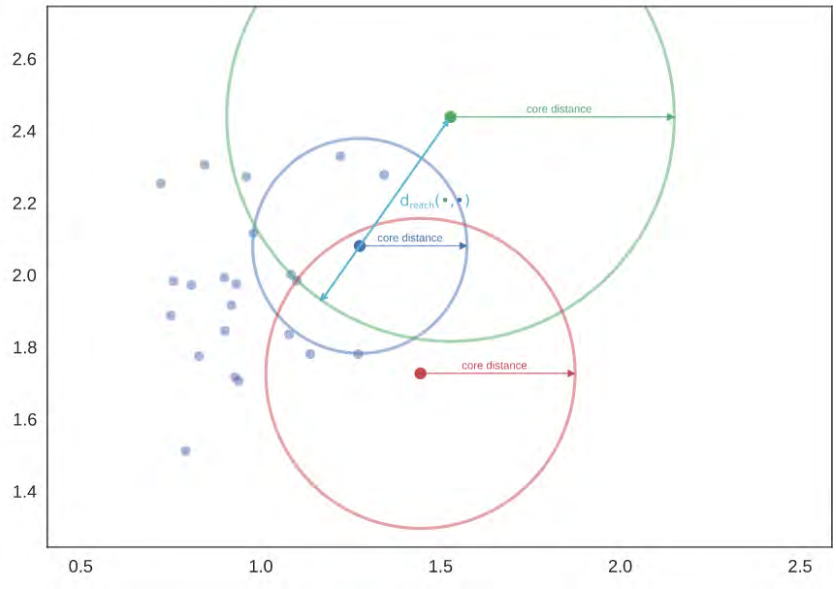


Figure 1.3: Plotting test points on a 2-D plane, where the axes represent distance. An illustration of the key HDBSCAN definitions, core distance and mutual reachability distance are shown. Taken from http://hdbscan.readthedocs.io/

HDBSCAN starts by finding d_{core} of each spectrum. In Fig. 1.3, as the blue centre lies within the green circle, the blue centre is a member of the green cluster and the mutual reachability distance may be measured.In contrast, the red centre lies outside the green circle, hence they form two independent clusters. Given more observations, however, these centres may be linked.

This is repeated for the n spectra to build a weighted graph, $G_{min\_samples}$, with the spectra as vertices and an edge between any two spectra with weight equal to the mutual reachability of those points. Using Prim's algorithm, a minimum spanning tree may be constructed for the mutual reachability distance metric (equivalently potential clusters). The minimum spanning tree is converted to a hierarchy of connected components represented by a dendogram. Finally, to identify the clusters this dendogram (as seen in Figure 1.4) is condensed by "trimming" the off shoots, keeping clusters greater than the min_cluster_size parameter.

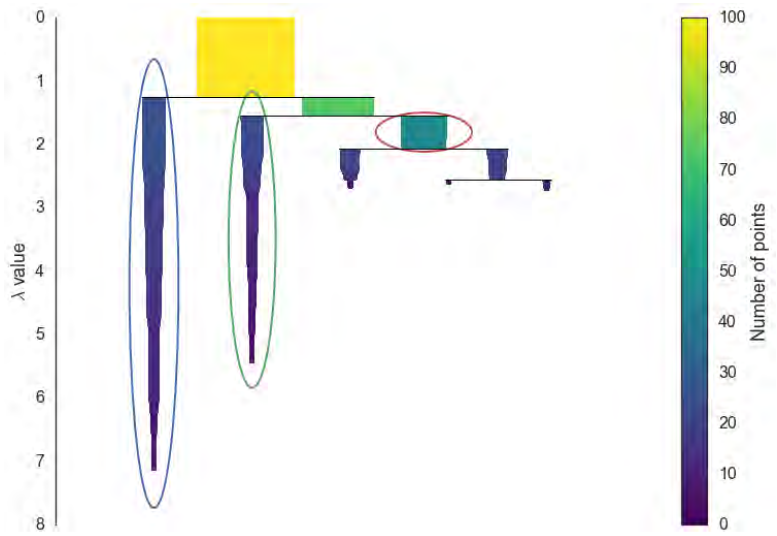


Figure 1.4: Identified clusters are highlighted in this condensed dendogram. λ is defined as the inverse of the distance between clusters, and represents the lifetime of a cluster. Clusters are selected which have the greatest lifetime or 'ink' size. Looking at the blue rectangle circled in red, once a cluster has been selected, then any cluster that is a descendant of it cannot be selected. Adapted from http://hdbscan.readthedocs.io/

HDBSCAN was chosen over other clustering techniques because it has been shown to be effective at identifying clusters of varying density, and the number of clusters to be identified does not have to be specified prior to execution.

1.4 Stellar Objects

The HERMES spectrograph produces data that allow ready identification of features in stellar spectra. Based on the specific features in the spectrum, the type of stellar object can be classified. In this work, we attempt to search for three types of objects: extremely metal-poor stars, solar twins, and "zero-metallicity" stars.

1.4.1 Metal-Poor Stars

The first stars are known as Population III stars. They were born from metal-free material, the primordial gas left after the Big Bang, and are thought to have been very massive and short-lived (Palla, 1983). Population III stars have yet to be observed and their existence is inferred from cosmology and the properties of present-day stars. Population III stars were devoid of any metals and are thus thought to have had relatively smooth spectra aside from hydrogen features. A few nearly zero metallicity stars, like SMSSJ031300.36670839.3 (Keller et al., 2014) have been discovered which may have been formed from material produced in Population III stars. These stars in turn, polluted the surrounding interstellar medium with chemical elements through both stellar winds and core-collapse supernovae. From this newly enriched medium and the effect of additional cooling mechanisms, subsequent generations of stars had higher metallicities and longer lives. This is based on the hypothesis that Population III stars had a top-heavy initial mass function (IMF) (Larson, 1998). Although direct evidence is lacking, a top-heavy IMF for Population III stars is plausible because: 1) They would have been free of metals, resulting in poor cooling efficiency, and hence inhibiting the collapse of their progenitor molecular cloud. As a result these clouds would preferentially form a massive, individual star rather than a group of less massive stars; 2) The initial cosmic environment would have had a very high Jeans mass (Bromm et al., 1999), because the early universe was dense and hot; and 3) The high density environment would have triggered many merger events between protostars or clouds (Abel et al., 2000).

The metal-poor stars that are observable today are Population II objects, like HE 1424-0241 (Cohen et al., 2007), and belong to the stellar generations that formed from non-zero metallicity gas. In their atmospheres these objects preserve chemical signatures of early supernova events. By identifying and studying these stars, it is possible to constrain the types of nucleosynthetic events responsible for the chemical abundances observed. Metal-poor stars thus provide archaeological evidence of the earliest times of the universe. This knowledge is invaluable for our understanding of cosmic chemical evolution and the onset of star and galaxy formation. As metal-poor stars are the local equivalent of the high-redshift Universe, they provide us with observational constraints on the nature of the first stars and supernovae (Frebel, 2008).

Looking at our Galaxy, an important question that may be addressed by finding more metal-poor stars is the nature of the metallicity distribution function (MDF) of the Galactic halo. With a larger sample of metal-poor stars, one may answer the question of whether the limit of low metallicity in the Galaxy has been reached and test whether the MDF remains constant as a function of distance within the Galactic halo (An et al., 2013). Furthermore by finding metal-poor stars, an accurate measurement of abundances useful for determining the ages of astrophysical objects is possible, a technique known as cosmo-chronometry. Metal-deficient stars have been shown to contain uranium (Cayrel et al., 2001) and thorium; by comparing the U/Th ratio, an age can be estimated. Lastly, by combining metallicities and ages with locations in the Galaxy it is possible to constrain how the Milky Way formed, as the current theory suggests the Milky Way formed inside-out, with the most metal-poor stars located near the Bulge (Tumlinson, 2010).

1.4.2 Solar Twins

The search for stars similar to our Sun was first conducted by Hardorp (1978) using 77 solar type stars scanned with 20Å resolution in both the northern and southern hemispheres. Hardorp (1978) only found two stars that were similar to the sun (V = 6.6 mag) in the ultraviolet spectrum (3640 - 4100Å): HR7504 in the northern sky and HR2290 in the southern sky. Cayrel de Strobel et al. (1981) provided the first definition of a sun like star or a solar twin as,'a star having within acceptable observational uncertainties, properties such as surface gravity, effective temperature and metallicity identical to the sun.' As a result, the spectrum would be indistinguishable from that of the sun. In contrast, solar analogues are stars which closely resemble the spectroscopic and photometric properties of the sun (de Strobel, 1996).

While there is currently no perfect solar match, the approach of Porto de Mello & da Silva (1997) to finding solar twins was to χ^2 match over a large range in the spectral energy distribution (SED), comparing the equivalent widths of specific iron lines relative to the sun and the line depth of specific iron lines. A more sophisticated approach was undertaken by Mahdi et al. (2016) analysing the ELODIE archive. They selected a dataset of ≈2800 spectra and measured the degree of similarity between two spectra using a minimum distance criterion implemented in TGMET code.

Finding solar twins has implications for exoplanetary system research. As planets form from proto-planetary disks, and these disks interact with the outermost layers of its host star, the composition of the planets formed may be influenced by the chemical composition of the star. Combining this with the knowledge of Earth as the only known region of space that supports life, orbiting in the "Goldilocks region" of our sun, by finding solar twins we increase the probability of finding not only exoplanetary systems but life similar to our own.

1.4.3 Outliers

Outlier detection is a key research area in machine learning, with a diverse range of applications from fraud (Konijn & Kowalczyk, 2012) to cancer detection (Wang & Rekaya, 2010). The identification of

anomalies is of relevance as such anomalies can represent both artefacts - i.e observational errors, poor pipeline processing - as well as strange and novel findings in astronomical spectra. In many cases, the discovery of outliers is of more interest than standard data points. For example, fraudulent activity, such as money laundering, should have different characteristics in the data to regular behaviour. Genomic variations, such as differential gene expression in cancerous cell lines, are also anomalies when compared to benign cells. In astronomy the detection of anomalies can lead to the discovery of strange and previously unknown objects in the universe. The method described in this thesis can be extended to detect observations that have been poorly processed or are of types of objects previously unknown.

2

Methodology

This chapter provides a step by step description of each stage of the analysis. First, the techniques used to process the spectral FITS files, and prepare the data for dimensionality reduction and clustering are outlined. Then the execution of the t-distributed Stochastic Neighbour Embedding algorithm, as well as the clustering algorithm, Hierarchical Density Based Scan (HDBSCAN) are explained. The method presented in this thesis is a novel way of finding similar spectra in a large spectroscopic survey.

2.1 Data

Each raw spectrum was processed by the GALAH data reduction pipeline (Kos et al., 2017) which applies bias subtraction, flat fielding, optimal extraction and wavelength calibration through an ar-clamp, in addition to sky subtraction, telluric absorption line removal and a barycentric velocity correction. An uncertainty spectrum, which describes the standard deviation of each pixel in the reduced spectrum, is produced by propagating an initial photon counting standard deviation for each pixel through each step of the pipeline. Observation nights are stored in individual folders, which have the pipeline reduced spectra for the stars observed; in addition each folder has a parameters text file that contains the stellar parameters estimated by GUESS, a stellar template-fitting code which applies a χ^2 minimisation to stellar templates on a grid of Teff, log g and [Fe/H] (Lin, 2015). It should be noted that this approach to estimating stellar parameters produces values that are not perfect.

Each reduced spectrum is stored in a FITS file named with a 16 digit number, where the first 6 digits describe the observation date (yymmdd) and the last digit represents the channel of the spectrograph. Each FITS file contains four extensions: i) the reduced spectrum; ii) the error spectrum; iii) the spectrum without sky subtraction or telluric removal; iv) the error spectrum without sky subtraction or telluric removal. There are four FITS files for one object, reflecting the 4 channels of the HERMES spectrograph, shown below in Table 2.1:

Channel	Name	Wavelength (nm)
1	Blue	471.5 - 490.0
2	Green	564.9 - 587.3
3	Red	647.8 - 673.7
4	NIR	758.5- 788.7

Table 2.1: HERMES Wavelength Ranges

The dataset used in this work is the GALAH reduction version 1.3. To construct a unique star identifier, the first 15 digits of the FITS file name were used; this is denoted as starID in the following.

2.1.1 Labelled Dataset

To construct the labelled dataset, the stellar classification by Traven et al. (2017) was cross matched to the GALAH survey data (Martell et al., 2017) by s_object_ID (starID), a unique star identifier common to both.

Using the SIMBAD astronomical database[1] 12830 objects were classified by Traven et al. (2017) into 5 stellar classes:

1. Binary stars
2. Cool metal-poor giants
3. Halpha/Hbeta emission
4. Hot stars
5. Stars with molecular absorption bands

After cross matching, this resulted in 9510 stars with labels. In addition, for this thesis three other categories of labelled objects were added: 4 Metal-poor stars; 356 Solar observations; and 300 "Zero"-metallicity spectra, with the labels ExtMetalPoor, Solar and ZeroMetallicity, respectively. The labelled data set used in this thesis thus contains 10170 objects. This final sample does not include poor-quality spectra with low signal-to-noise (S/N < 25 per resolution element), and spectra which, after being processed by GUESS, gave invalid stellar parameter estimates (the invalid parameter estimates were denoted by 9999).

As a validation of the t-SNE method, Figure 2.1(a) shows the t-SNE map coloured by the stellar parameter, effective temperature ([t_{eff}]), as estimated by GUESS. Figure 2.1(b) is the same map coloured by the classification label from SIMBAD. The hot stars t-SNE cluster is clearly distinguishable by both temperature and label, indicating both that the SIMBAD label are valid and that the t-SNE is sensitive to both the fundamental stellar parameters and overall SIMBAD stellar classification.

2.1.2 Unlabelled Dataset

The observations that do not have a stellar classification label after cross-matching by s_object_ID are given the label "unlabelled". These observations are combined with the known labelled sample to give the full GALAH dataset. The total number of stars analysed in this work is 203,357, of which 193,187 are unlabelled.

2.1.3 Extremely Metal Poor Stellar Spectra

A sample of 7 metal-poor stars was manually identified (J.Simpson 2017, personal communication, 22 March) by cross-matching stars observed in GALAH with stars in SIMBAD determined to have [Fe/H] ≈ -3; this latter sample yielded a table with 538 unique stars. A 10 arcsec positional cross-match of this table against the GALAH data gave the list of 7 possible metal poor stars. It should be noted that

[1] http://simbad.u-strasbg.fr/simbad/

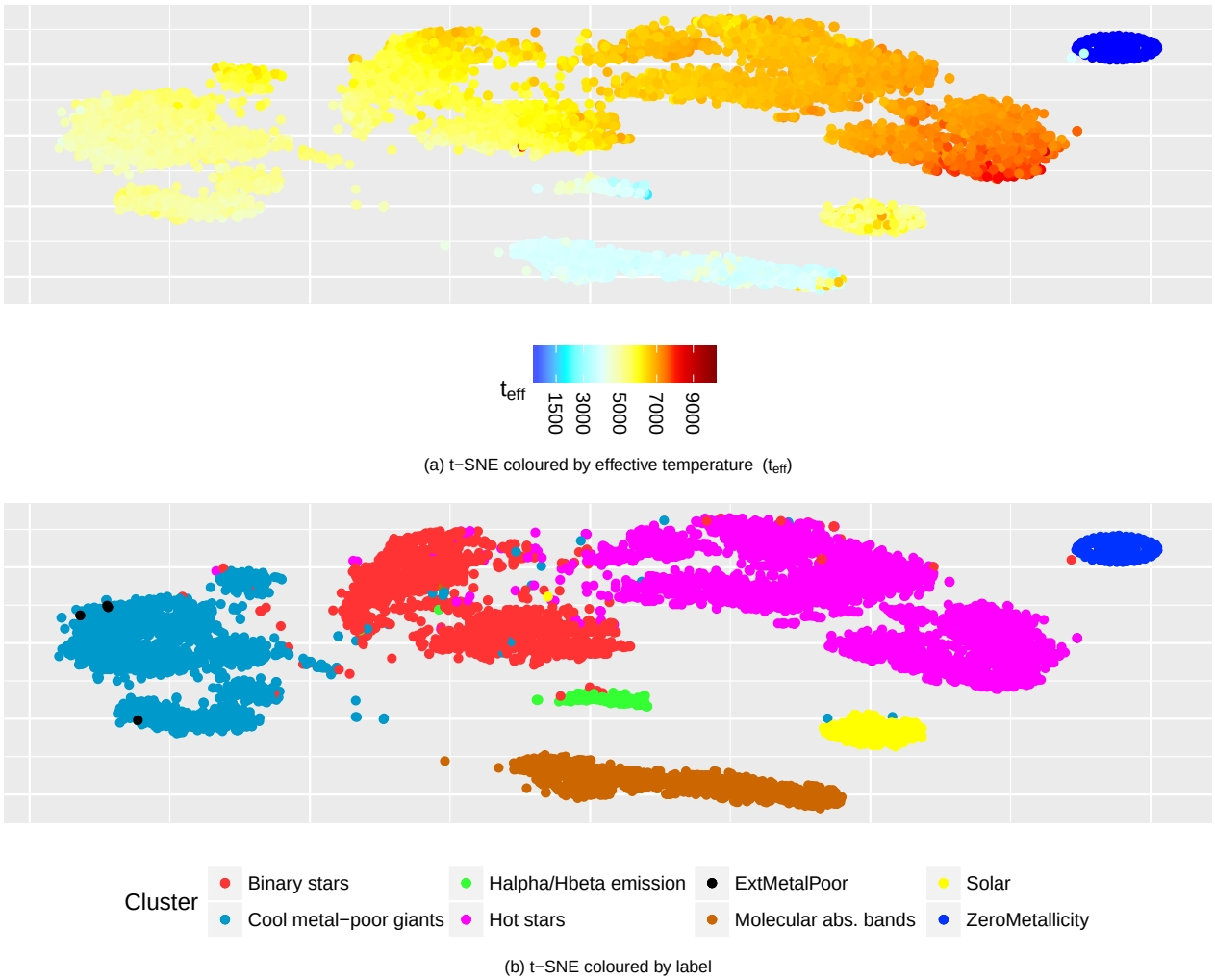


Figure 2.1: (a) t-SNE map coloured by effective temperature from GUESS (Lin, 2015). (b) t-SNE map coloured by stellar classification labels from SIMBAD. Each point represents a star, which has had its spectral information collapsed into two points. (a) shows the distribution of effective temperature; when compared with (b) the hottest stars are clearly located in cluster labelled Hot stars. Hence the classification system from SIMBAD, defined by Traven et al. (2017), is valid. It also shows that t-SNE is sensitive to effective temperature and the stellar SIMBAD classification.

a 10 arcsec cross-march is quite wide, and may potentially lead to faulty cross-matches especially in crowded fields. However 2 of the stars failed to make it through the processing stage because of poor spectrum normalisation, and one of the identified stars did not satisfy the metallicity requirement of [Fe/H] ≈ -3 as a review of the literature revealed it was actually a [Fe/H] = -2 star. The final sample of 4 extremely metal-poor stars is presented in Table 2.2 and is given the label "ExtMetalPoor" in the labelled dataset.

2.1.4 Mock Zero-Metallicity Spectra

The object SMSSJ031300.36670839.3 (Keller et al. (2014), see Section 1.4.1) has not been observed by GALAH. However, the spectrum when plotted is almost featureless in the GALAH channels aside from hydrogen absorption. Therefore to simulate this spectrum, a flat spectra with all flux values set to 1 was generated. The stellar parameter of effective temperature was set to an artificial value of 1500K. Three hundred of these spectra, and not just one, were constructed and given the label "ZeroMetallicity" in the labelled dataset. Three hundred spectra were generated to ensure that a clearly identifiable cluster would form. The flat mock spectrum has no features and thus we call it and similar stars to it "zero"-metallicity, even though we do not expect to find bona fide "zero"-metallicity

galah_id	SIMBAD name	RA	Dec	t_{eff}	log g	[Fe/H]	NASA/ADS
-3	HD 122563	210.632689	9.6860967	4367	0.6	-3.15	2010ApJS..191..352K
4472243	BPS CS 22892-0052	334.256908	-16.657519	4850	1.6	-3.03	2013A&A...551A..57H
5779095	HE 0124-0119	21.744042	1.587583	4330	0.1	-3.57	2015ApJ...798..110L
5432063	2MASS J21260896-0316587	321.537333	-3.283	4725	1.15	-3.22	2011ApJ...742...54H

Table 2.2: Four Extremely Metal-Poor Stars

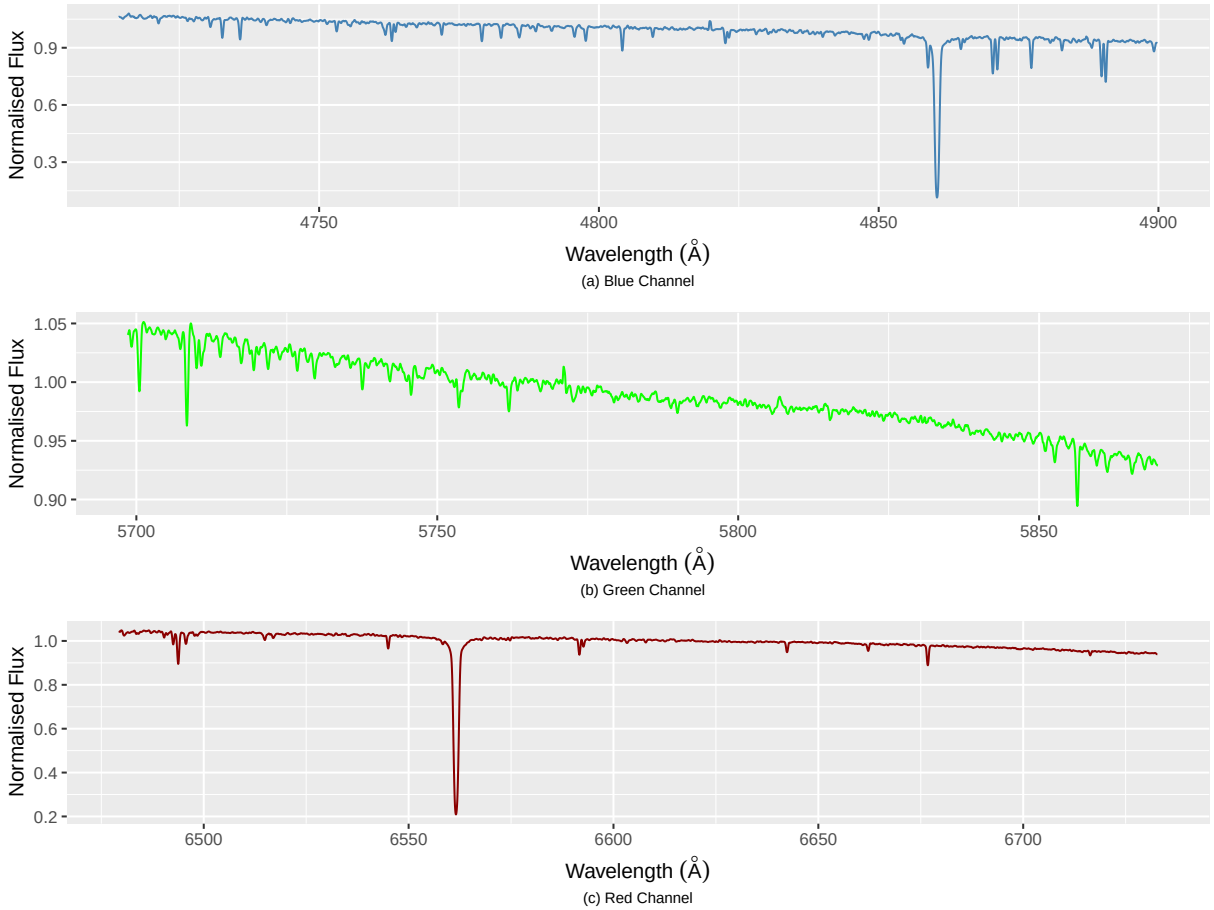


Figure 2.2: The Extremely Metal-Poor star HD 122563, with [Fe/H]=-3.15. As expected, the blue and red channel look relatively featureless except for the prominent Hα and Hβ lines.

stars, and are not claiming that such stars are truly Pop III stars.

2.1.5 Solar Spectra

The Sun-illuminated twilight sky was observed on the 5th April 2015 using HERMES. As HERMES has multiple fibres, where each fibre maps to one star in the sky, when observing the twilight sky each fibre can be taken to be one solar spectrum. This results in 356 Solar spectra, creating a larger sample than just a single Solar spectrum. These spectra are given the label "Solar" in the labelled dataset. An example is shown in Figure 2.3.

2.2 Processing the Spectra

There are several steps to process the sky-subtracted GALAH FITS files. These consist of:

1. Filtering and Clipping

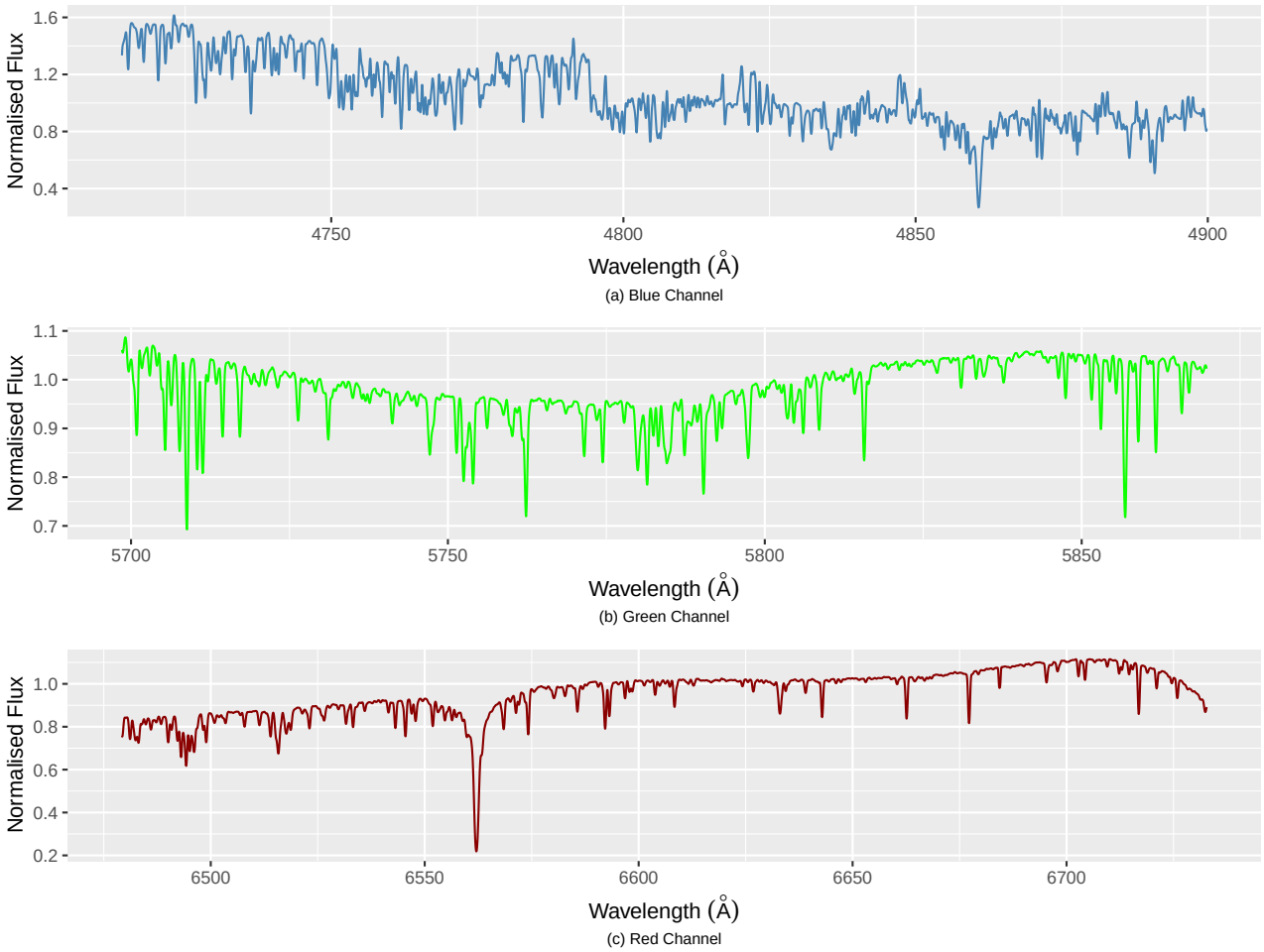


Figure 2.3: Solar Spectra obtained from twilight observations of the sky

2. De-Redshifting or de-blueshifting
3. Resampling and Normalising
4. Smoothing
5. Storing the Processed Spectra

As described in this section, the processing of the spectra is conducted in Python, as it has well-defined packages to handle FITS files and functions to deal with spectroscopic data.

2.2.1 Filtering and Sigma Clipping

Two possible filtering methods were considered: 1) Mean filter, and 2) Median filter. The first method was significantly impacted by outliers that appear as large spikes in the spectrum. The median ignores large spikes, thus the median filter was adopted.

By comparing the difference of the raw spectrum and the median-filtered spectrum over the error spectrum (σ), outlying flux values can be identified. Anything with a deviation greater than 10σ was chosen for rejection to allow for some emission lines and also remove significant outliers. This is a relatively aggressive clipping, as the presence of emission lines are not the focus of this thesis. The equation to implement sigma clipping is given here:

$$bad = \frac{|spectrum| - median}{\sigma} > 10.0 \tag{2.1}$$

2.2.2 Doppler Correction

The next step is to Doppler correct the spectra, to compensate for the star's motion relative to the Sun. To shift a spectrum, an estimate of its heliocentric radial velocity[2] was computed with GUESS (Lin, 2015). To implement Doppler correction, the dopplerShift algorithm in the PyAstronomy [3] library was used. An example of Doppler correction can be seen in Figure 2.4.

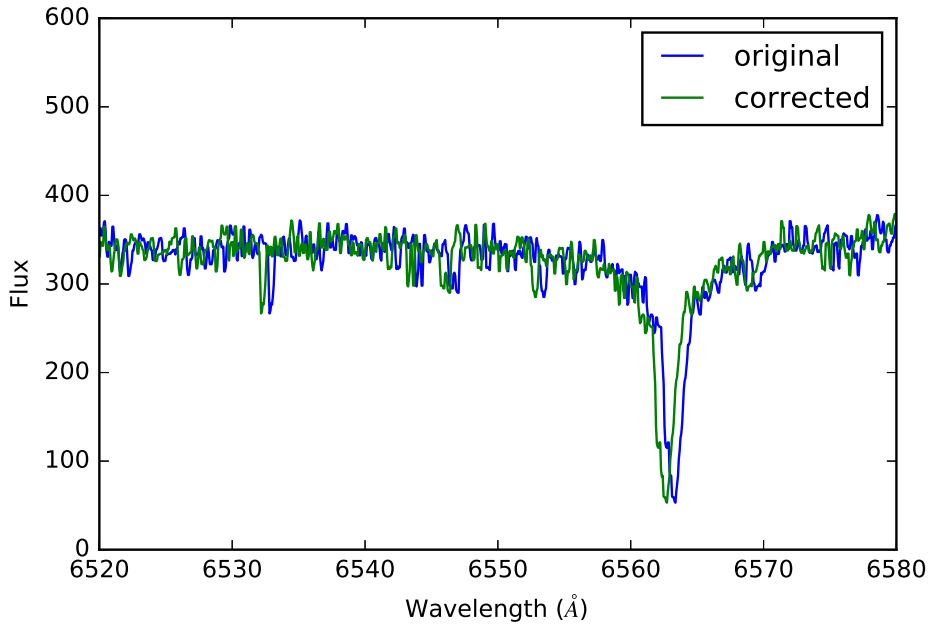


Figure 2.4: An example of a blueshifted and corrected spectrum

2.2.3 Resampling and Normalising

Resampling of the spectra was required to standardise each spectrum to the same number of pixels/wavelength values. Each spectrum was resampled to 4096 wavelength values using a spline of order 3 to interpolate onto the new wavelength range. The number of output wavelength values is identical to the input number, so that as much spectral information as possible was retained.

After resampling, the final stage in processing the spectra was to normalise the spectrum to 1 by dividing out the continuum, so that the spectral features were the dominant features in the data. Four possible normalisation options were considered:

1. Mean
2. Polynomial fit
3. Chebychev fit
4. Median

The first option normalised the spectrum by dividing each flux value by the mean of the spectrum. The mean normalisation was a simple initial method that removed differences in the continuum due to, for example, distance (and hence relative brightness) of the stars. It worked well for emission lines but introduced bad continuum features such as random absorption lines or non-linear continuum's.

[2]Three of the extremely metal-poor stars had incorrect radial velocities that had to be manually corrected.
[3]https://github.com/sczesla/PyAstronomy

The next two methods attempted to fit a polynomial through the continuum, which worked better than the mean as the polynomial fits better accounted for wavelength sensitivity differences. However, these methods tended to over-fit the strong spectral features observed. The last method, median, which divided each flux value by the median of the spectrum was ultimately the chosen method, as it neither over-fitted nor completely removed the spectral features. The median was computed from all wavelengths aside from 200 wavelength data points on either end of each spectral channel. This ensured the median value was not skewed by the ends of the spectra, which can be distorted after Doppler correction.

2.2.4 Smoothing

After the above corrections, some spectra were noticeably noisy. Noise has the effect of introducing an added complexity by causing similar stars to look different and ultimately making the t-SNE map uninformative. In an attempt to remove some of this noise, a boxcar filter of order 8 was applied to the spectra. A boxcar filter works by convolving the spectra with a box-shaped pulse of $2M + 1$ values all equal to $1/(2M + 1)$, where M is the order number. The impact of noise on the t-SNE map is discussed in 3.1.

2.2.5 Storing the Processed Spectra

To deal with the entire 1TB-sized GALAH dataset, initially the processed spectra were stored in a SQLlite database in two tables using a database key from fileID, starID, night and channel. One table was for the parameters (log g, t_{eff}, [Fe/H], radial velocity) and the other for the wavelength, measurement (flux, error, fit) and value. Storing the processed spectra in a SQLlite database structure was attempted in order to have customisable, quick and simple way of selecting observations of interest. For example, in such a structure it is trivial to select all stars from a specific night's observation or select all stars with [Fe/H] < -3. SQL clustered indexes were constructed in order to optimise commonly used SQL queries.

However, it was found that the structure of the SQLlite database with the indexes caused the size of the SQLlite file to grow to approximately 1TB for ~200,000 stars. This led to query times so long that they would often time-out and fail. To alleviate this issue, the parameters were stored in a single .CSV and the spectra were stored in three .CSV files by camera. Although the ability to write specific queries to select certain observations was no longer possible, the size of each spectral .CSV was approximately 10GB and that of the parameters .CSV was 30MB, considerably smaller than the single SQLlite database file.

2.3 Analysis

The following section outlines the pipeline used to analyse the processed spectra, to search for extremely metal-poor stars, solar twins and "zero" metallicity stars. First, an exploratory analysis of a known sample of spectra (labelled dataset) was undertaken to understand the impact different parameters in t-SNE and clustering had on the final results. The outcomes of the exploratory analysis were used to inform the appropriate set of parameters adopted in t-SNE for the full run on the unlabelled sample, which is described in Chapter 3. The analysis was run using the statistical software R on a 28-core Intel(R) Xeon(R) E5-2695 2.30GHz CPU server with 384GB of RAM maintained by the Faculty of Science and Engineering at Macquarie University .

2.3.1 Applying the t-SNE Algorithm

As described in Chapter 1, t-Distributed Stochastic Neighbour Embedding (t-SNE) by Maaten & Hinton (2008) is a popular technique in the Machine Learning community. t-SNE is well suited for the visualisation of high-dimensional datasets as it visually disentangles datasets comprised of a large number of dimensions into just to two or three dimensions. The algorithm was implemented using a multi-core parallelised R package which is actually a wrapper around the C++ implementation of Barnes-Hut TSNE, developed by RGLAB [4] Whilst experimenting with the algorithm, the 28 cores available were used and various combinations of the two primary t-SNE parameters, perplexity and iterations, were tested. These results, as well as computational statistics, are described in Chapter 3.

2.3.2 Applying the HDBSCAN Algorithm

Having projected the high dimensional set of spectra on to the 2-dimensional plane with t-SNE, HDBSCAN was then run on the resulting t-SNE space to identify dense regions. The aim of HDBSCAN is to quantitatively group stars of a particular class into a cluster. The HDBSCAN Python code was executed from within R. In the exploratory stage described below in Chapter 3, a variety of values for the two main parameters (min_cluster_size and min_samples) were looped through, to assess how effective HDBSCAN is at identifying the same cluster distribution as the labelled points.

2.3.3 Interactive t-SNE Visualiser

To enable a more user friendly approach to understanding the t-SNE maps, a new application was developed. This GUI interface (see Figure 2.5) enables the user to select specific stellar classes, zoom in on a region of the t-SNE map and also select a star and plot the corresponding spectrum, providing a more detailed exploration of a t-SNE projection. This tool may be accessed via a URL in the future.

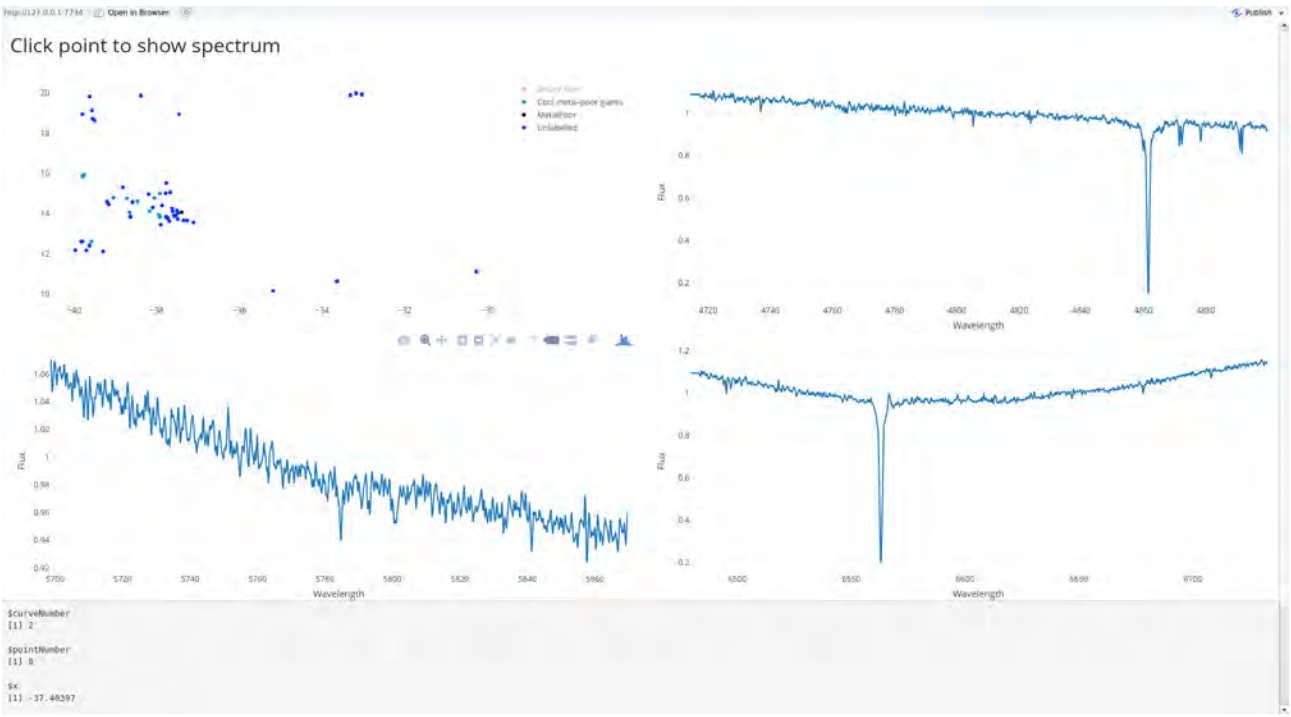


Figure 2.5: The t-SNE Visualiser provides a novel method to explore a t-SNE projection plot, allowing the user to remove points that are not of interest, zoom in to specific regions (in the upper left window) and plot the spectra for a clicked point in the blue, green and red channels of the HERMES spectrograph.

[4] https://github.com/RGLab/Rtsne.multicore

When theory and experiment agree, that is the time to be especially suspicious.

Niels Bohr

3 Results

This chapter demonstrates how to target objects of interest in a large spectroscopic dataset using t-SNE and HDBSCAN. We identify interesting candidates in the unlabelled dataset by examining regions in the resulting t-SNE projection map near where known interesting objects end up.

The first section explores how each parameter of the t-SNE algorithm affects the dimensional map. t-SNE is highly flexible but can be difficult to interpret. Thus by understanding how t-SNE behaves under different parameterisations, it becomes possible to develop an intuition as to what is happening behind the map. By using a labelled dataset and changing the t-SNE parameters, sample size and perplexity, we show how t-SNE distributes similar observations on the 2-dimensional projection. With a given t-SNE projection we then explore the impact of the HDBSCAN parameters, min_samples and min_points, in identifying groups around known interesting objects. The effectiveness of HDBSCAN will determine the ease of automated 'star of interest' detection. A new tool is developed to assist identifying candidates when the number of input known interesting objects is too low to generate an "island" in a t-SNE projection.

The conclusions drawn from testing t-SNE and HDBSCAN are then used to inform how to search for "zero"-metallicity stars, solar twins and extremely metal-poor stars in the final unlabelled dataset.

3.1 Learning from a Labelled Dataset

To understand how to best target rare but interesting objects in an unlabelled dataset using dimensionality reduction, we test the various parameterisation of t-SNE and HDBSCAN on a known labelled stellar sample. There are five parameters that vary in t-SNE & HDBSCAN; perplexity, sample size and the number of iterations in the t-SNE step and min_points and min_samples at the HDBSCAN stage.

The most important t-SNE parameter is perplexity. Selecting the optimal perplexity has no metric to date, therefore we are following the recommendations in Wattenberg et al. (2016) which uses visual inspection to determine the best perplexity to use for a given sample. To express the meaning of perplexity using statistical formalisms (see more details in Section 1.3.2.3), the t-SNE algorithm is a unique dimensionality reduction method, as the algorithm alters the idea of distance to be areas of differing density in the data set. This results in dense clusters being expanded and sparse clusters contracted in an attempt to smooth out cluster sizes. This idea of density equalisation is a construct of

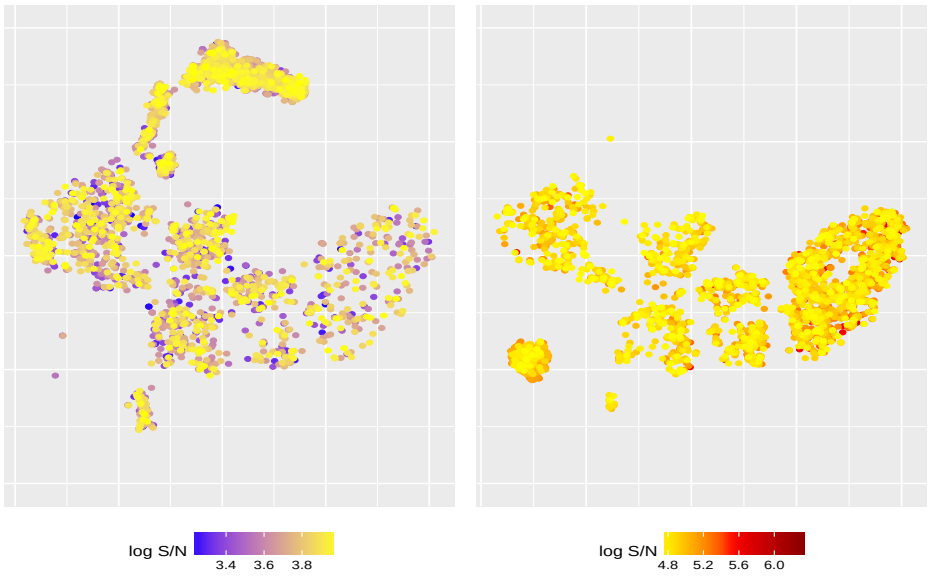
the algorithm, and means that the relative sizes of clusters are not detectable. The global geometry or the distances between clusters may not have meaning (unlike the distances between individual points). As this is determined by the perplexity, which is a global parameter, there may not be one perplexity value that will capture distances across all clusters. This is a potential area for further development.

The additional parameter we consider here is the wavelength range that is fed to t-SNE. Unfortunately, the well-known metallicity line series, the Calcium triplet at $\sim$ 8600Å, commonly used to identify metal-poor stars (e.g., Matijevič et al., 2017) is outside the wavelength range of GALAH. Therefore alternative wavelength ranges were examined to assess whether metal-poor stars may be found using different spectral features. The 4 different wavelength ranges that were considered in this thesis are:

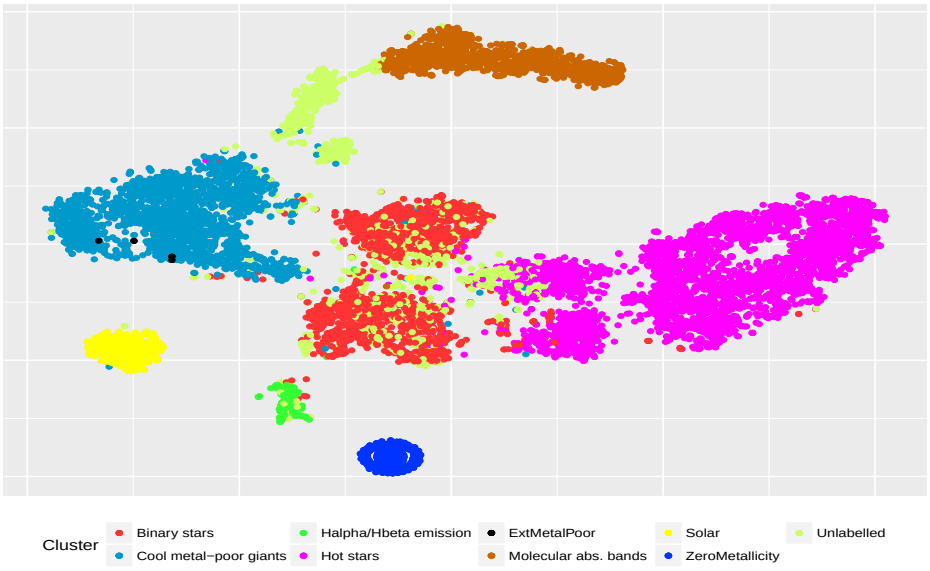
	Wavelength Range
A	4850 Å - 4880 Å, 5750 Å - 5780 Å, 6550 Å - 6580 Å
B	4841 Å - 4881 Å, 6543 Å - 6583 Å, 6706 Å - 6710 Å
C	4841 Å - 4881 Å, 5750 Å - 5780 Å, 6543 Å - 6583 Å, 6706 Å - 6710 Å
D	4714 Å - 4900 Å, 5650 Å - 5870 Å, 6479 Å - 6733 Å

Table 3.1: The four wavelength ranges tested for distinct island groups on the t-SNE projection.

Wavelength range A was chosen to capture the Hα and Hβ lines, and other diagnostic spectral features (as discussed in (Traven et al., 2017)), which should be sufficient to distinguish between hot and cool stars. Strong Hα is typically indicative of hot stars and weak Hα of cool stars. Wavelength range B was defined in order to test a wider window around Hα and Hβ as well as including a lithium spectral feature. Lithium was included as it can be used to further constrain Population II stars through the Spite Lithium plateau (Ryan et al., 1999). The combination of wavelength range A & B was expanded around Hα and Hβ and additional spectral lines were used in wavelength range C. Lastly, when computationally feasible the full GALAH wavelength range, defined by wavelength range D, was considered. This was to assess whether, despite noise, if t-SNE could identify subtle spectral features that are common between similar stars.



(a)



(b)

Figure 3.1: The left panel shows the lowest quartile of $\log_{10}$ signal-to-noise (S/N) stars; The right panel shows the highest quartile of $\log_{10}$ S/N stars; (b) The known labelled stars distribution for all SN. Comparing the left and right panel, the arm structures are noticeably dominated by stars with low S/N whereas, the dense circular cluster shapes have a higher signal to noise. Using this comparison and looking at plot (b) the stars with the highest S/N are the hot stars and the Solar spectra. The stars with the lowest S/N are the molecular absorption band stars that are completely absent in the right panel.

Before illustrating the effect of different parametrisations on the t-SNE projection, Figure 3.1 shows the results of a test to determine whether noise influences the analysis. It has been shown by Kos et al. (2017) that noise in a data set has an effect on how points are distributed on a t-SNE map. They find a high signal-to-noise ratio (S/N) leads to more well-rounded and defined circular clustering, whereas low S/N points tend to form strings or tendril-like arms. It should be noted that the evidence of tendril-like arms may not solely indicate noise, as it may also be indicative of a higher dimensional structure being lost. Kos et al. (2017) demonstrate this by projecting a 3-dimensional helix on to a 2-dimensional plane using t-SNE, and show that the tendrils on the 2-D projection are reflective of

the higher dimensional substructure of the helix in 3-D. This motivates the strict criteria of selecting stars with an overall S/N > 25 adopted in the following analysis.

3.1.1 Effect of t-SNE parameters

In understanding how the t-SNE map is projected there are three parameters that need to be explored: perplexity, sample size, and the number of iterations. Perplexity is the primary variable in t-SNE and can be interpreted as a measure of the number of neighbours a given point will be compared to in the high dimensional space in order to construct the 2-D projection.

t-SNE iteratively reduces the Kullback-Liebler divergence between the high-dimensional space and the 2-D projection. As the 2-D projection approaches that of the high-dimensional space the Kullback-Liebler divergence becomes smaller. By testing the number of iterations it is possible to assess whether a higher value leads to a more distinct separation between "islands" on the t-SNE map.

Since this thesis outlines an methodology to find rare objects in a large dataset, we also consider how various input interesting object sample sizes (relative to the number of objects being searched). This will show how effective t-SNE is at distinguishing small clusters within an overall larger dataset. In addition, testing input sample size may highlight whether perplexity and the number of iterations need to be of a higher value as the overall size of the dataset to be searched increases.

The following subsections illustrates the effect of changing perplexity, sample size and the number of iterations using wavelength range C.

3.1.2 Effect of Perplexity

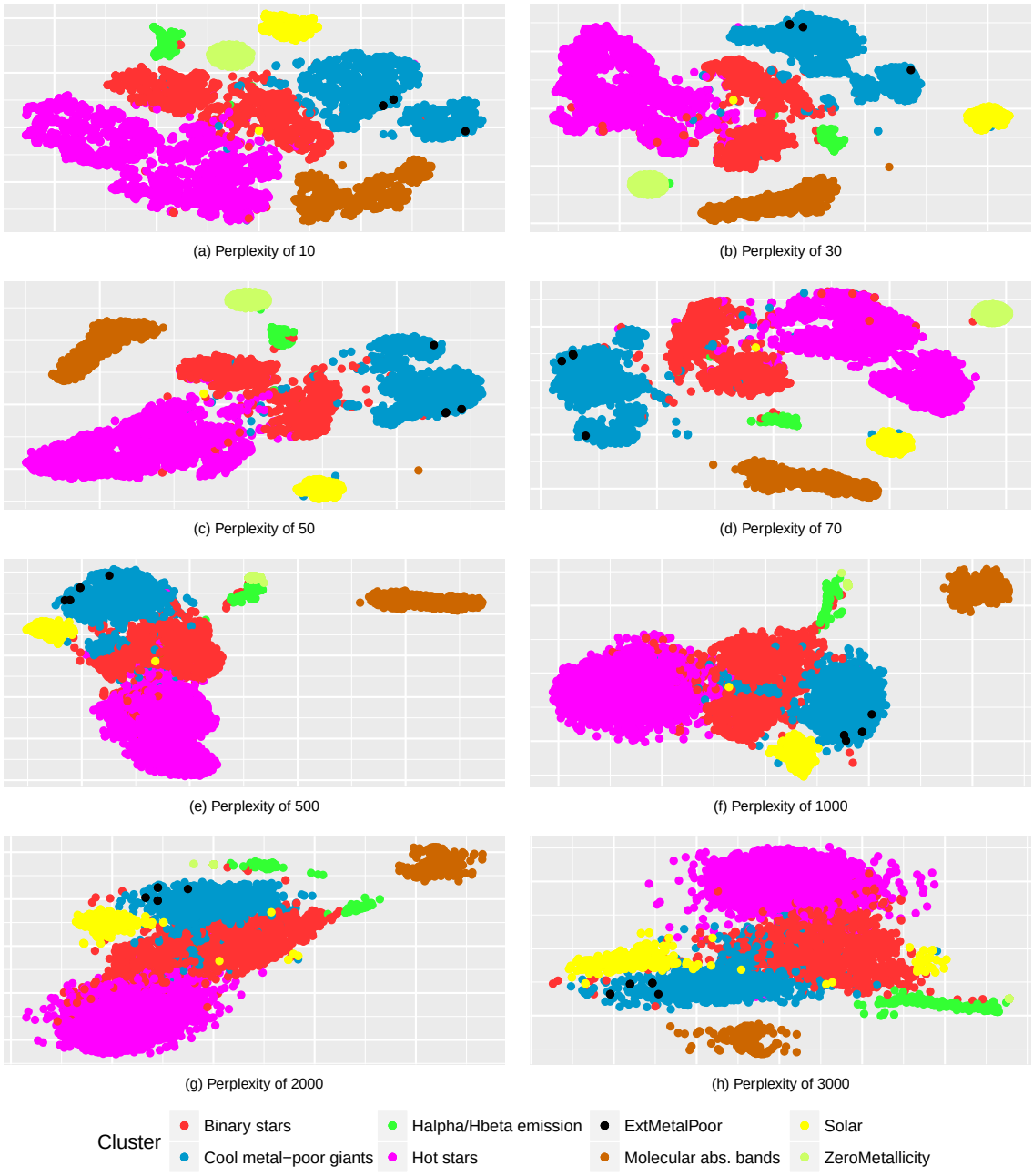


Figure 3.2: The effect of altering perplexity on the t-SNE map, using the labelled sample and the number of iterations at 10000 with wavelength range C. As the perplexity value increases, the distribution of points become more grouped and overlapped.

Maaten & Hinton (2008) state that typical perplexity values for t-SNE are in the range of 5-50, as values outside this range alter the map in 'strange ways'. For very large sample sizes, however, the value of perplexity may exceed 50. Iterating perplexity over wavelength range C as shown in Figure 3.2 indicates that, as perplexity increases, the clusters become less distinct. This is evident when tracing the position of either the Solar cluster or the ZeroMetallicity cluster through each plot. The optimal cluster separation for the labelled sample is between 50 and 70. Despite being outside the range discussed by Maaten & Hinton (2008), the value of 70 is selected as the perplexity value for the remainder of this chapter based on the Author's intuition.

3.1.3 Effect of Sample Size

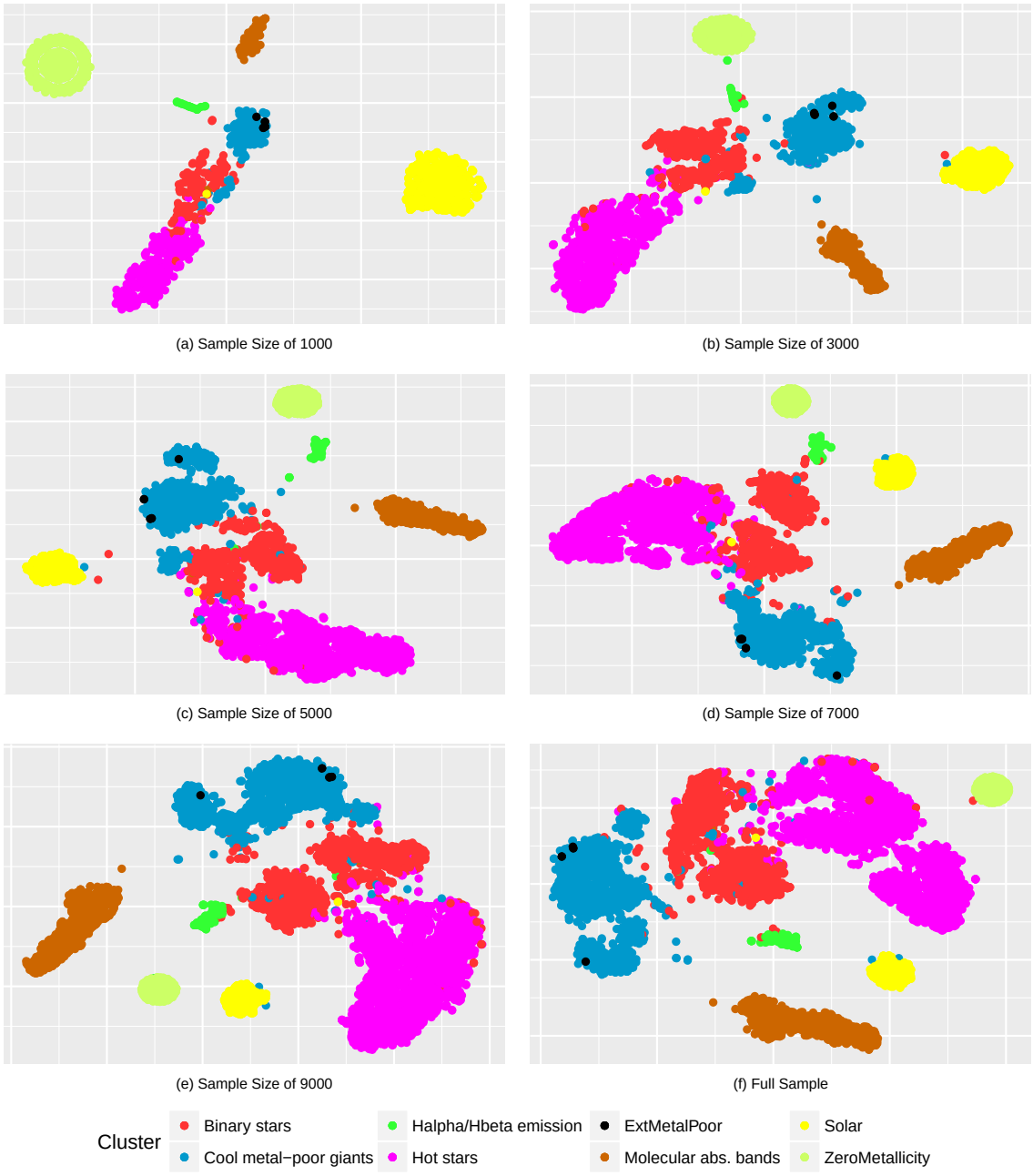


Figure 3.3: The effect of changing sample size on the t-SNE map, holding perplexity fixed at 70 and the number of iterations at 10000 with wavelength range C. Each sample was randomly generated but was built to always contain all the extremely metal-poor stars, Solar spectra and the synthetic "zero"-metallicity spectra. The addition of more stars did not impact the distribution of the extremely metal-poor stars, but did result in a denser clustering of the Solar and zero-metallicity spectra

.

Figure 3.3 suggests that different sample sizes do not strongly impact the distribution of the input stars of interest. This likely is a reflection of the fact that the input are just as rare in all of the sample sizes considered for the larger dataset. In addition, looking at (a), t-SNE tries to find structure in data where there is none, evident in the faint ring like structure around the ZeroMetallicity cluster, which are identical spectra. This indicates the perplexity value of 70 is too high for a sample size of 1000. Therefore there is likely a co-dependence of the optimal value of perplexity and the size of the dataset.

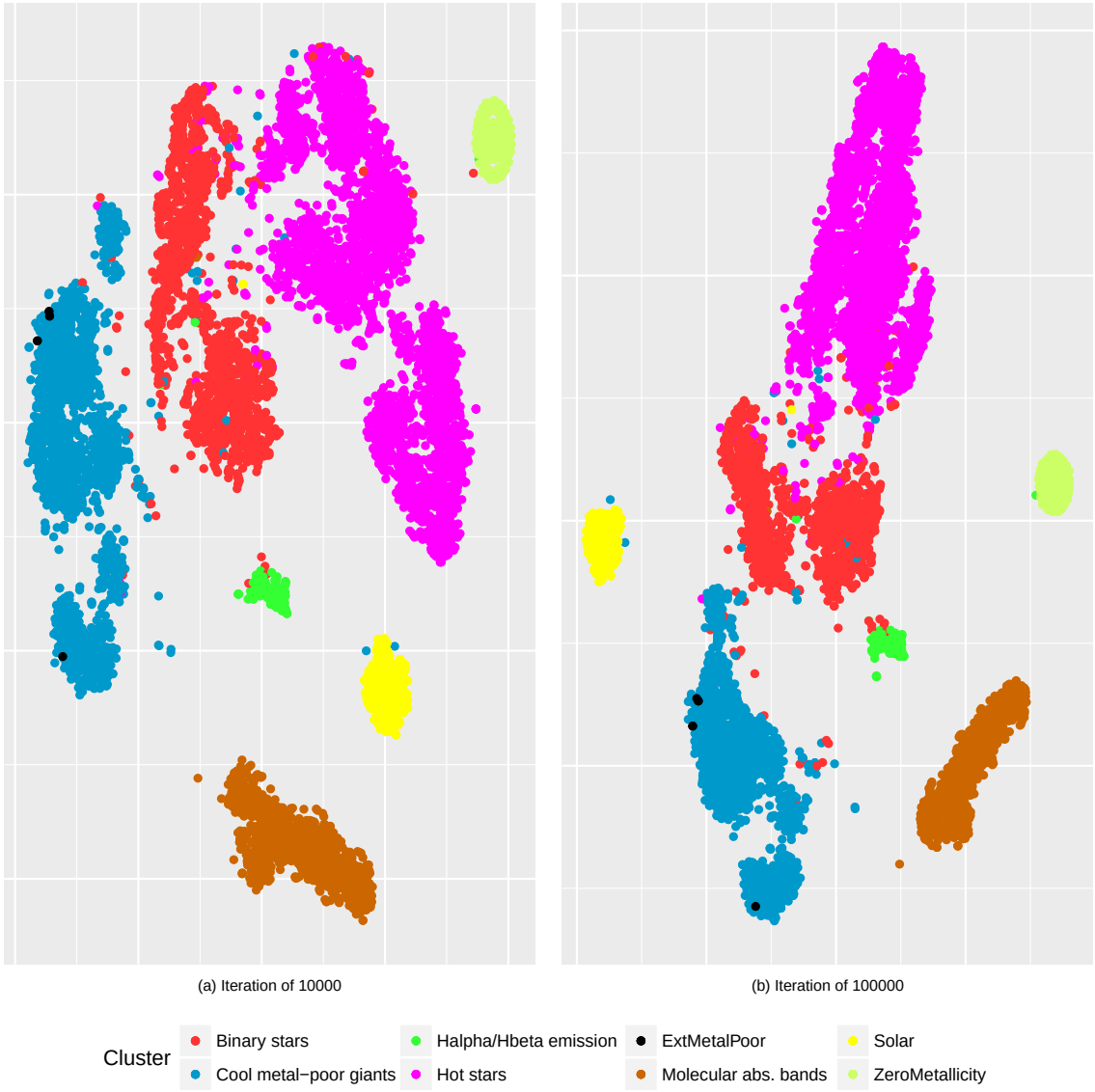


Figure 3.4: The effect of changing the number of iterations on the t-SNE map, using the labelled sample and the perplexity at 70 with Wavelength Range C. Comparing (a) with (b), the cluster separation is similar but overall the density of the clustering is higher in (b). This may simply be a result of this particular run of t-SNE but in general the higher the number of iterations, the more likely it is that the cost function converges with a smaller error.

3.1.4 Effect of Iterations

As expected, increasing the number of iterations in Figure 3.4 allows t-SNE to approach a more stable solution (i.e. the Kullback-Leibler divergence approaches stability). This results in t-SNE converging to a more optimal solution (i.e., a smaller error in the approximation of the cost function). Although not immediately apparent, the cluster of similar objects in (b) become more tightly grouped. Furthermore, there is a lower likelihood of similar stars clustering in (a) than (b). This result suggests that 10,000 iterations is too low in this scenario. Theoretically, the higher the number of iterations, the more likely it is that t-SNE will approach the optimal 2-D projection.

3.1.5 Effect of Wavelength Ranges

The following section considers the effect of different wavelength ranges on the t-SNE map, with perplexity at 70 and the number of iterations at 100,000. Figure 3.5 shows that, given the GALAH dataset, using the defined different wavelength ranges does not affect the clustering differentiation on the t-SNE projection, as for each parametrisation the clusters are distributed similarly. The Solar observations and the mock "zero"-metallicity spectra are clustered uniquely in a majority of the

configurations. For wavelength range D, however, the extremely metal-poor stars are more widely dispersed and there is some clustering overlap between stellar types. This higher dispersion may be due to other features not accounted for, indicating substructure within the extremely metal-poor classification. For the purposes of this analysis, developing a more detailed classification labelling of the extremely metal-poor category (as there are many types of metal-poor stars, from binaries to carbon-enhanced metal-poor) is ignored, as it is not the focus of this thesis.

3.1.6 Computational Statistics

Iteration	Wavelength Range A	Wavelength Range B	Wavelength Range C	Wavelength D
10000	6.27	6.34	5.96	46.2
100000	56.38	57.25	58.35	96

Table 3.2: Iteration run time in minutes, given the different wavelength ranges.

Table 3.2 shows that increasing from a limited to full wavelength range results in a large increase in computational, as expected with from the large increase in dimensionality. Figure 3.6 shows the computational time for each of the different parametrisations considered. In summary, holding the other variables fixed, increasing the value of a parameter – perplexity, sample size, iteration step or wavelength range – roughly linearly increases the time to complete.

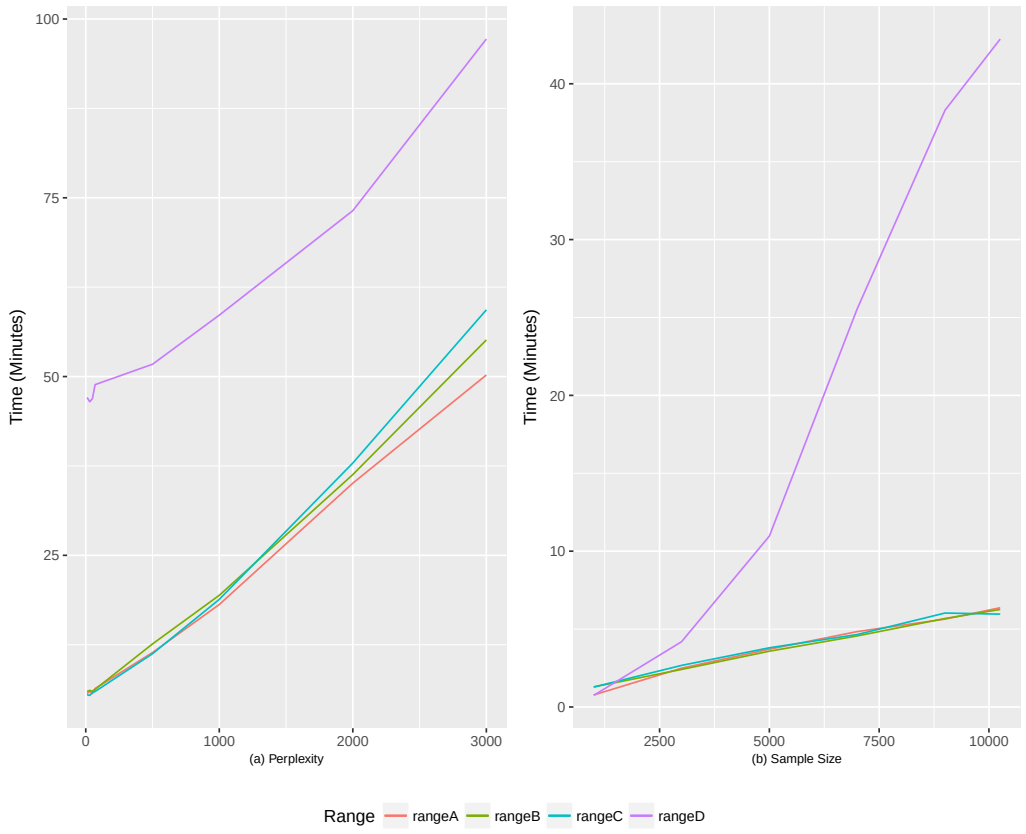


Figure 3.6: Computational run times for the different wavelength ranges. Given the dimensionality of the full wavelength range is higher than the other ranges, it is to be expected that the run time will be considerably higher. It is interesting to note the almost exponential run-time increase for Range D in (b) in comparison to the other ranges' roughly linear run-time increase.

3.1.7 HDBSCAN

t-SNE is used to collapse the stellar spectra into two dimensions. Section 3.1.1 shows that the optimal specification for t-SNE includes a high number of iterations and a perplexity parameter that is between

30-70 which may be influenced by sample size. In order to determine cluster membership, on a t-SNE projection, HDBSCAN or a visual inspection may be applied. The following section will take the best t-SNE map representation, a map that has a distinct visual cluster separation, and apply HDBSCAN in an attempt to obtain the same clustering as in the labelled setting seen in plot (c) of Figure 3.5.

From Section 2.3.2, the two parameters in the implementation of HDBSCAN are min_samples and min_cluster. As a reminder, the definitions given in Section 1.3.3.2 for min_samples and min_cluster are 'the threshold for a point to become a core point' and 'the criteria for determining which clusters are kept', respectively.

Figure 3.7 depicts the chosen t-SNE representation: wavelength range C with a perplexity of 70; using the full labelled sample size; and the HDBSCAN parameter min_samples set to 2. The parameter min_samples affects the restrictions on cluster membership: the higher the value provided, the more restrictive the clustering, as more points will be declared as noise and the clusters will be constrained to progressively denser regions. Various values of min_samples were tested, but the value of 2 was chosen, as it neither over- nor under-restricted the groupings. Larger values of min_samples caused the map to be defined as 1-3 clusters, not allowing for more intricate structures.

Figure 3.7 is divided into 4 plots by min_cluster size. This parameter is relatively intuitive to select, as it defines the minimum number of members of a cluster - set it to be the size of the smallest cluster group you wish to consider as a cluster.

When min_cluster size is set to be a low value, such as 10, HDBSCAN finds many dense regions with over 100 clusters; this result is uninformative when displayed, and shows more detail than the defined stellar classification labels. This indicates that HDSBCAN is identifying sub-structures within the defined stellar classes, suggesting the classification labels can be broken into sub-categories. By visually inspecting the detected clusters in Figure 3.7, run using values of min_cluster size between 50 and 500, the cluster of Solar spectra (the cluster labelled 1) is easily recognised by HDBSCAN. However, the clustering never approaches the exact distribution as in Section 3.1.4. HDBSCAN is unable to identify the 4 extremely metal-poor stars because of the low density group that their positions generate. This suggests that to search for a labelled object with a low sample count requires a much finer and more interactive detection method.

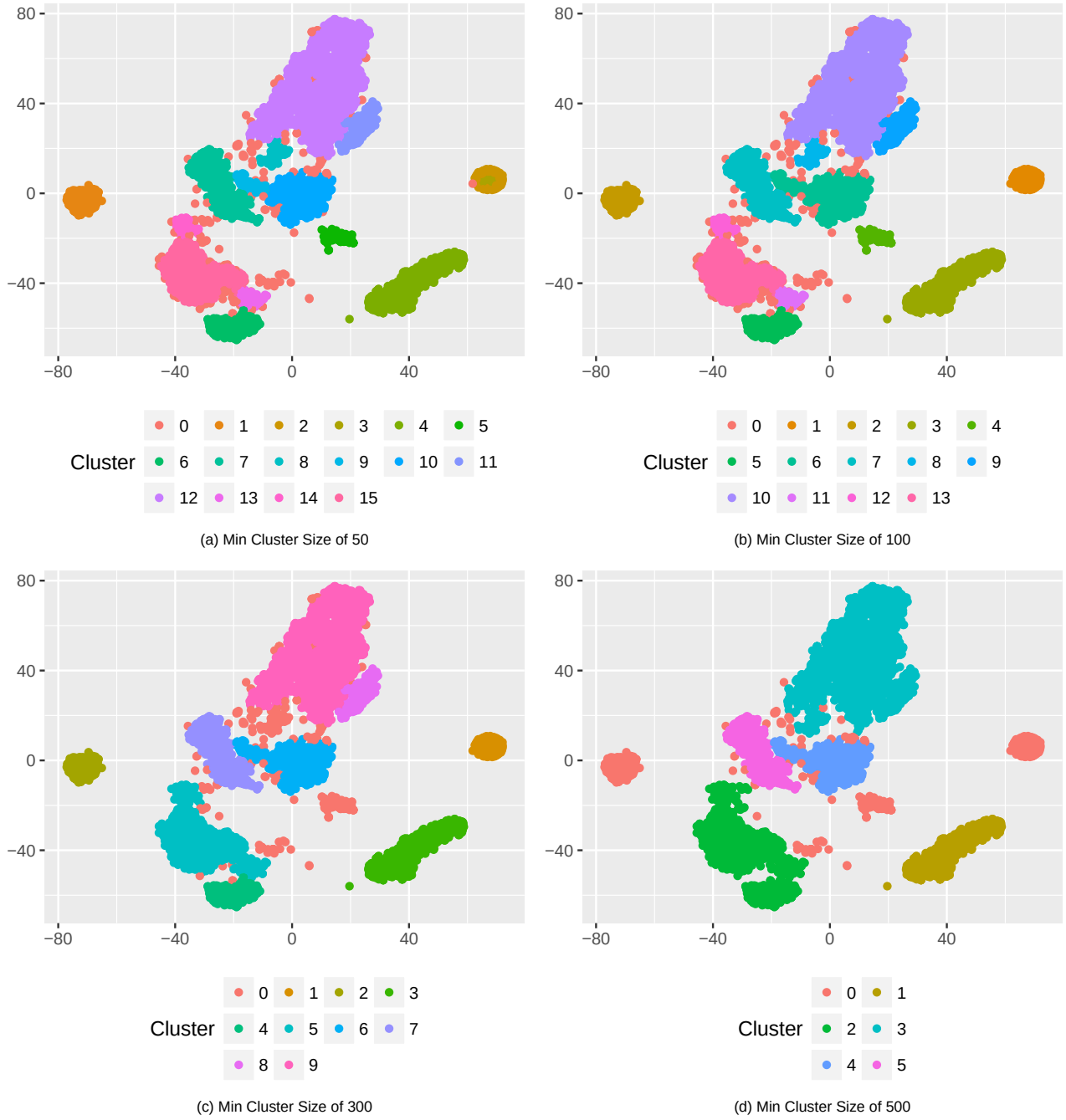
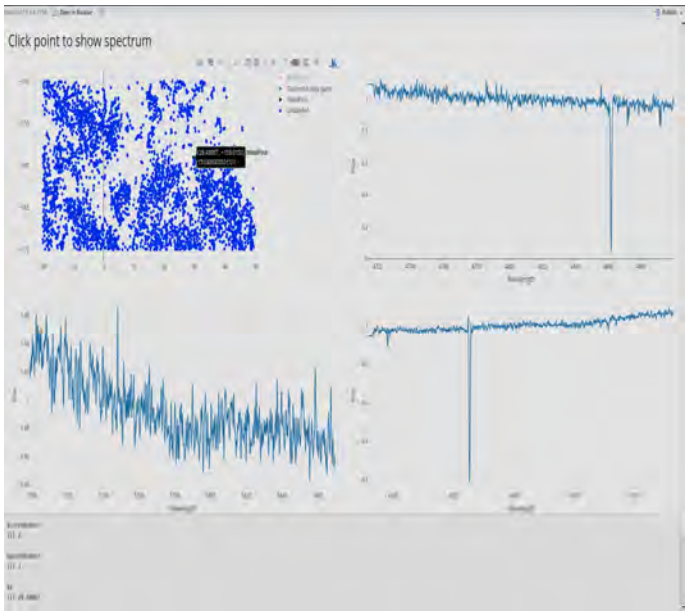


Figure 3.7: Different min_cluster sizes were iterated to identify whether HDBSCAN could render the same clustering as the labelled sample in panel (b) of Figure 3.5. Starting at low values of min_cluster size, the dense Solar and "zero"-metallicity cluster is identified. Increasing the min_cluster size sees the "zero"-metallicity cluster disappear. In every tested case, HDBSCAN fails to identify the extremely metal-poor cluster, as the cluster is too sparse (**CL: I think this is simply because of their small number?**); hence an alternative method is required to locate extremely metal-poor stars.

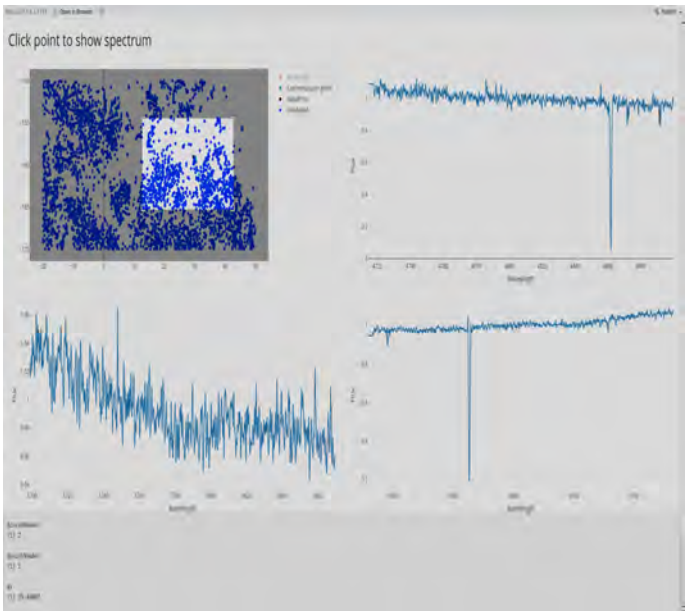
3.1.8 t-SNE Visualiser

As HDBSCAN was ineffective at finding low-density clusters (e.g., metal-poor stars that were distributed over the map), the t-SNE visualiser was developed to allow a user to interactively explore the t-SNE space and focus on regions where the density of stars is low. Having selected a region, the tool allows one to remove known stellar classification labels, leaving clusters that contain stars of interest. The user is able to select a point on the t-SNE space which then generates a plot for each HERMES

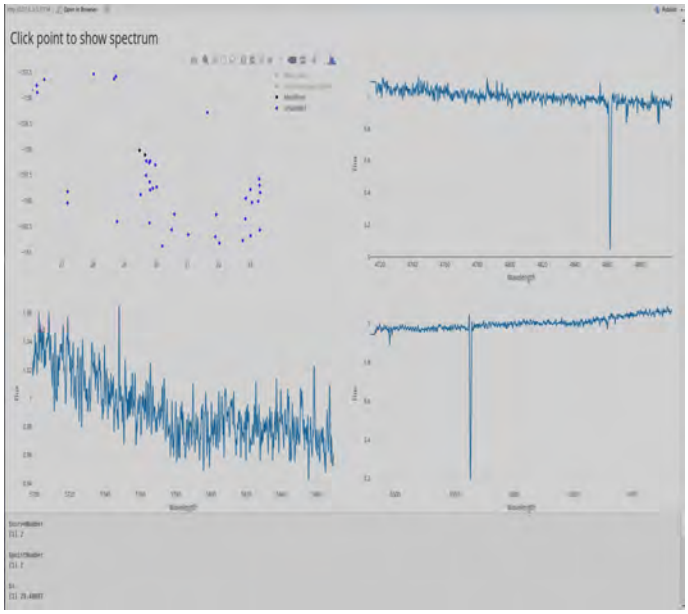
wavelength channel. This enables the user to interactively classify unknown points.



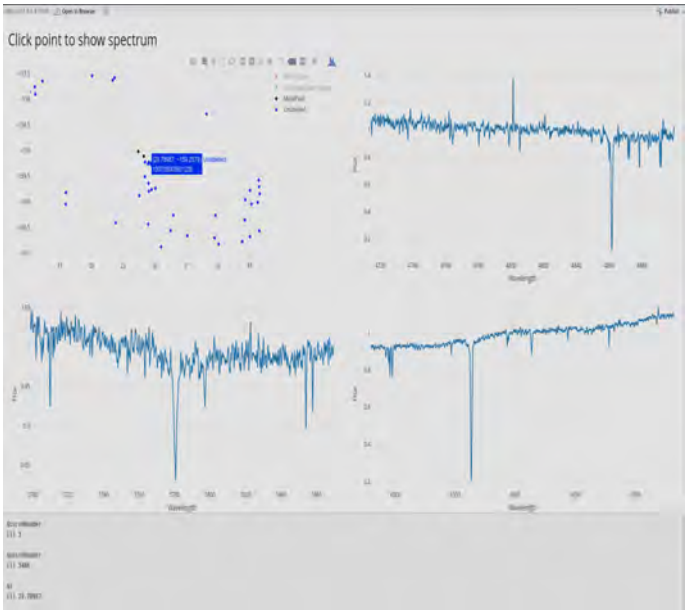
(a) Locate the position of the object of interest and plot its spectrum in each HERMES channel



(b) Use the zooming tool to restrict the area of interest



(c) Clearer zoomed in region, with redundant objects turned off



(d) Click an unknown point around the object of interest to plot the spectra

Figure 3.8: An application of identifying unknown stars around an extremely metal-poor star. Stepping through each panel from (a)-(d) shows how to use the t-SNE visualising tool to identify unknown stars around a star of interest.

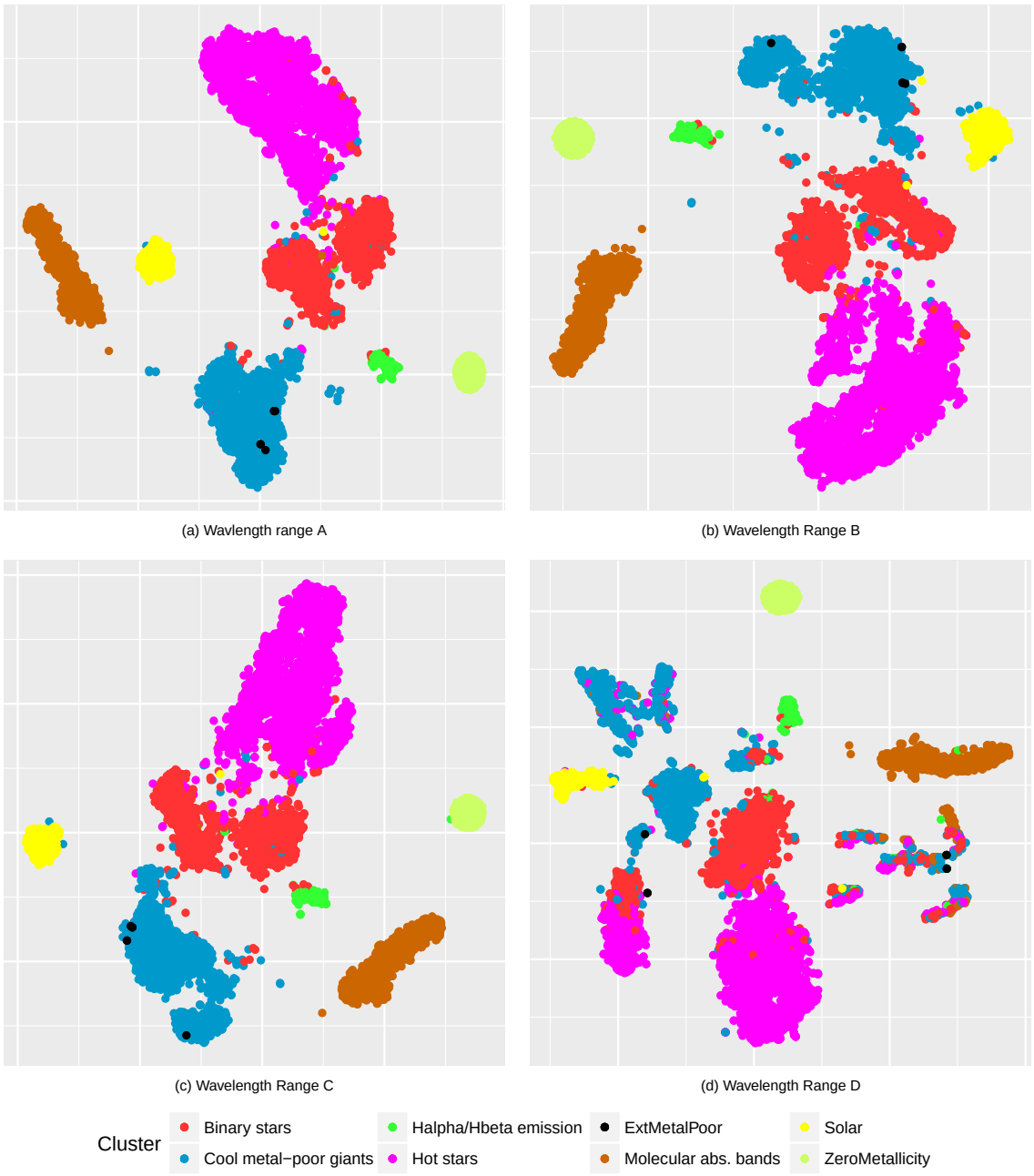


Figure 3.5: Different wavelength ranges with a perplexity of 70 and iterations at 100,000. There are no obvious cluster separation differences between the selected wavelength ranges (A-C), indicating that any of the defined wavelength ranges may be used. The full wavelength range (D), however, has some overlap between stellar types.

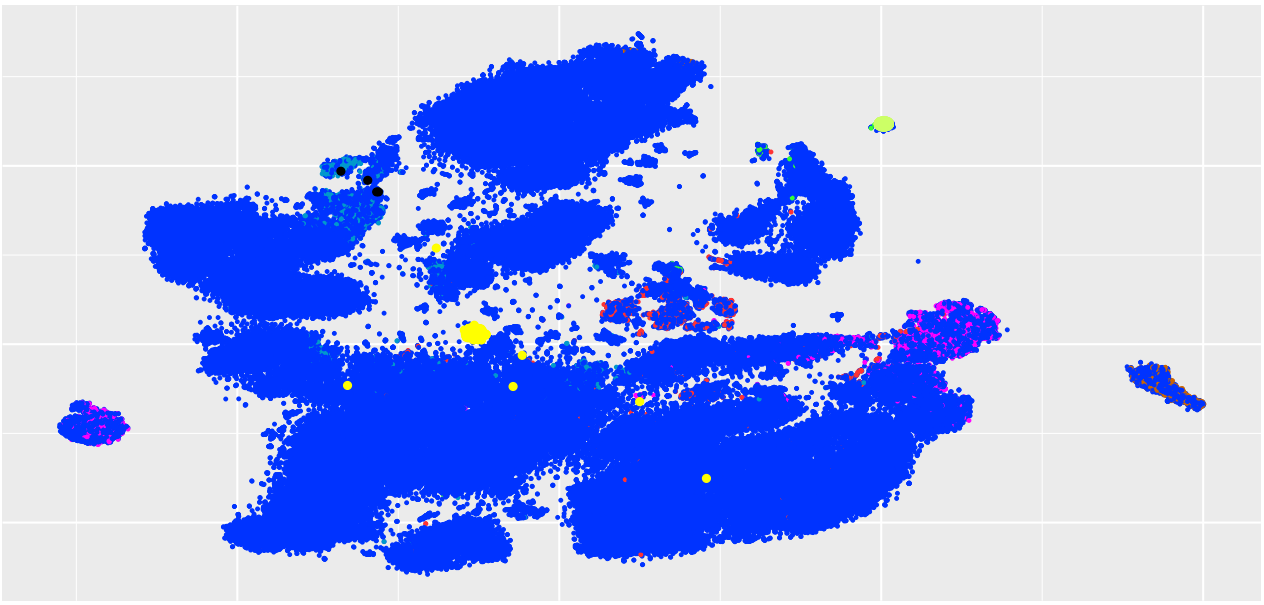
3.2 Targeting Metal-Poor Stars and Solar Twins in the Unlabelled Dataset

The following section is an outline and illustration of the methodology used to find extremely metal-poor stars and Solar twins, given the exploration conducted on the labelled data in Section 3.1.

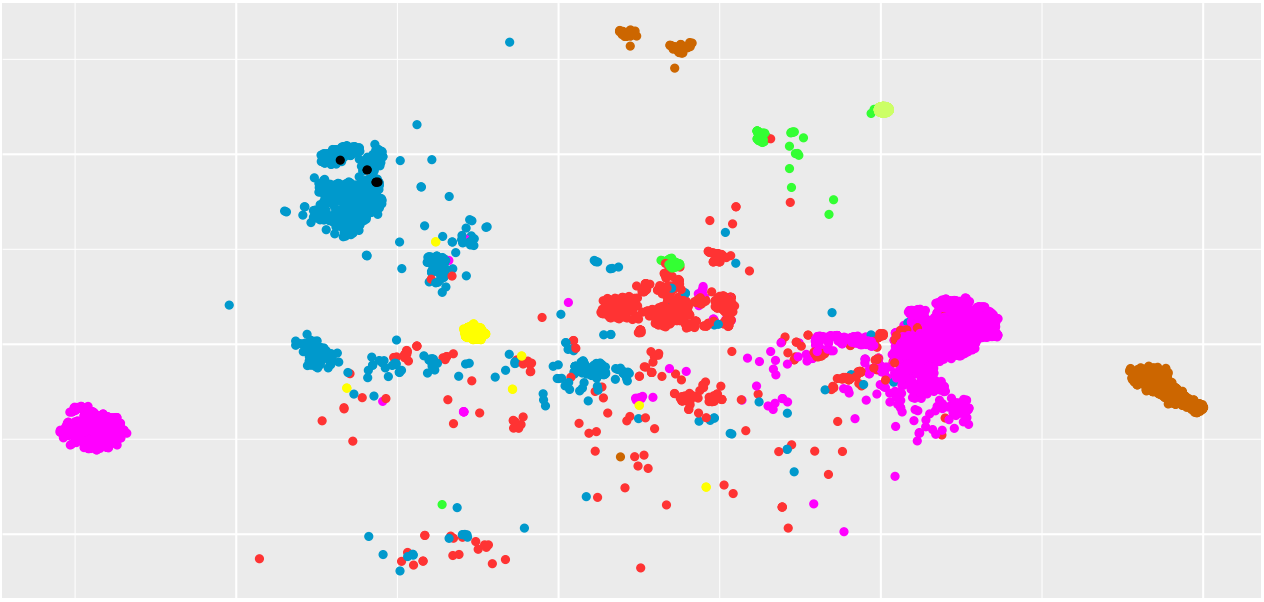
From Section 3.1, the most significant parameter in achieving a well-distributed t-SNE map was the number of t-SNE iterations: the higher the better. This results in a lower error in the convergence criteria, leading to more distinct, separated and compact clusters. With the t-SNE parameter perplexity, however, at low values the points within a given cluster become too sparse, and if the perplexity is set too high the distinction between clusters is lost. This suggests that, to achieve a well-distributed t-SNE map, the number of iterations should be set as high as possible and the value of the perplexity parameter should be between 30 and 70.

The above parametrisation was sample size dependent, as a smaller sample size would not require the number of iterations or the perplexity (See Section 3.1.3) to be as high. To find metal-poor stars and solar twins efficiently, rather than passing the entire observed wavelength range for every star's spectra, certain wavelength ranges that contain specific spectral features were tested. From Section 3.1 it is noticeable that the clustering was not affected when comparing the full wavelength range to the limited wavelength ranges. Hence for computational efficiency, the full wavelength range can be omitted, and indeed including the full wavelength range may introduce spurious substructure, due to issues such as poor flux normalisation at the edges of each channel. To maximise information we adopted wavelength range C as defined in Section 3.1. Wavelength range C includes numerous spectral features that may differentiate between subtle structures in the larger unlabelled dataset and hence assist with forming distinct clusters. Importantly, wavelength range C also retains a similar clustering to the alternative wavelength ranges tried on the labelled data, as seen in Figure 3.5.

Figure 3.9 shows the final chosen t-SNE map projection from a run on the entire GALAH dataset. To achieve this selected projection, t-SNE was run on wavelength range C using 200000 iterations and a perplexity of 70. Figure 3.9 panel (a) shows the general structure of the map using all the points, panel (b) of the same figure, however, shows the underlying labelled structure. The ideal clustering for the underlying structure would be perfectly shaped circular-like clusters around the SIMBAD classification groups. The overlaying of unlabelled points would thus result in slightly larger clusters with similar shape. However, the figure shows this clearly did not happen. This does not mean the projection is invalid, but rather that t-SNE is able to detect subtleties in the substructure that the broad SIMBAD stellar classification labels do not recognise. This would be useful for creating a finer labelling classification system, but as the focus of this thesis is on how to find targeted objects in a large dataset this aspect has not been explored.



(a) Full sample with unlabelled stars



(b) Full sample without unlabelled stars

Figure 3.9: t-SNE map on the full sample. (a) shows the the t-SNE projection of the unlabelled and labelled sample, with many different distinct islands. The extremely metal-poor and solar stars are located in their own similar regions but the "zero" metallicity stars are in well defined small clusters. It is interesting to see that there are not many unlabelled stars like the solar cluster. (b) depicts the underlying structure of the labelled stars, with the previously well defined hot stars clusters broken apart into smaller sub clusters. This dispersion of the labelled data is indicative of a more detailed classification that the currently defined broad labels cannot identify.

Different HDBSCAN configurations were run on the t-SNE projection, however it was deemed more effective to manually inspect the few points around the extremely metal-poor, Solar twin and "zero"-metallicity clusters using the t-SNE visualiser, as only the Solar cluster was easily identified by HDBSCAN.

Using the t-SNE visualiser resulted in 66 possible metal-poor candidates and 20 possible solar twin candidates. 14 objects were identified near the "zero"-metallicity cluster but these will require further analysis, as they displayed distinctly diverse spectra. A random subset of each of these are shown in Figs. 3.10, 3.11 and 3.13

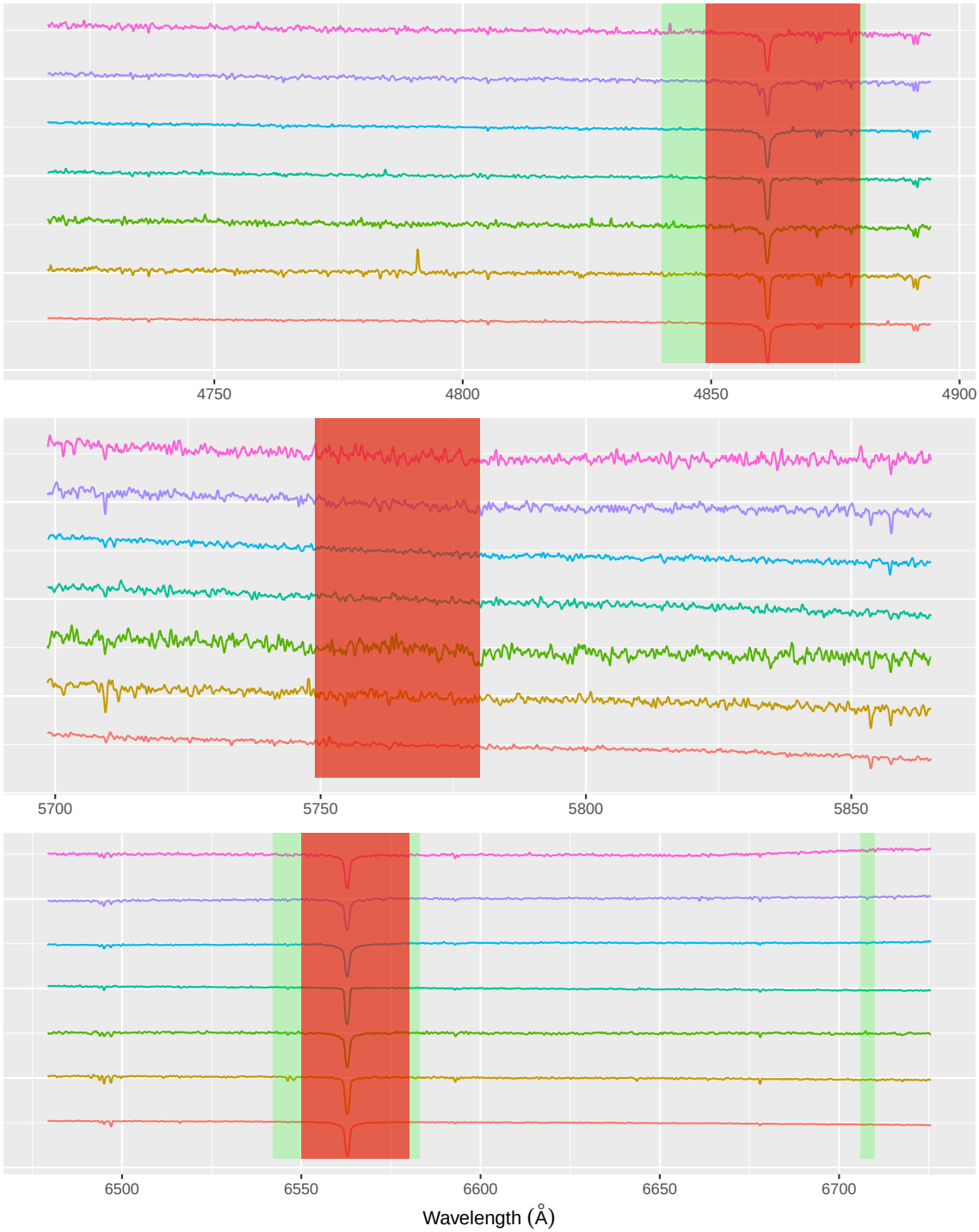


Figure 3.10: Seven randomly selected extremely metal-poor candidates with the different wavelength ranges shaded. The red regions are Wavelength Range A, the green regions are Wavelength Range B, and both the wider green regions and the red regions are Wavelength Range C. To identify these metal-poor candidates Wavelength Range C was considered. Many of the spectra look similar to or smoother than Figure 2.2, but require further investigation.

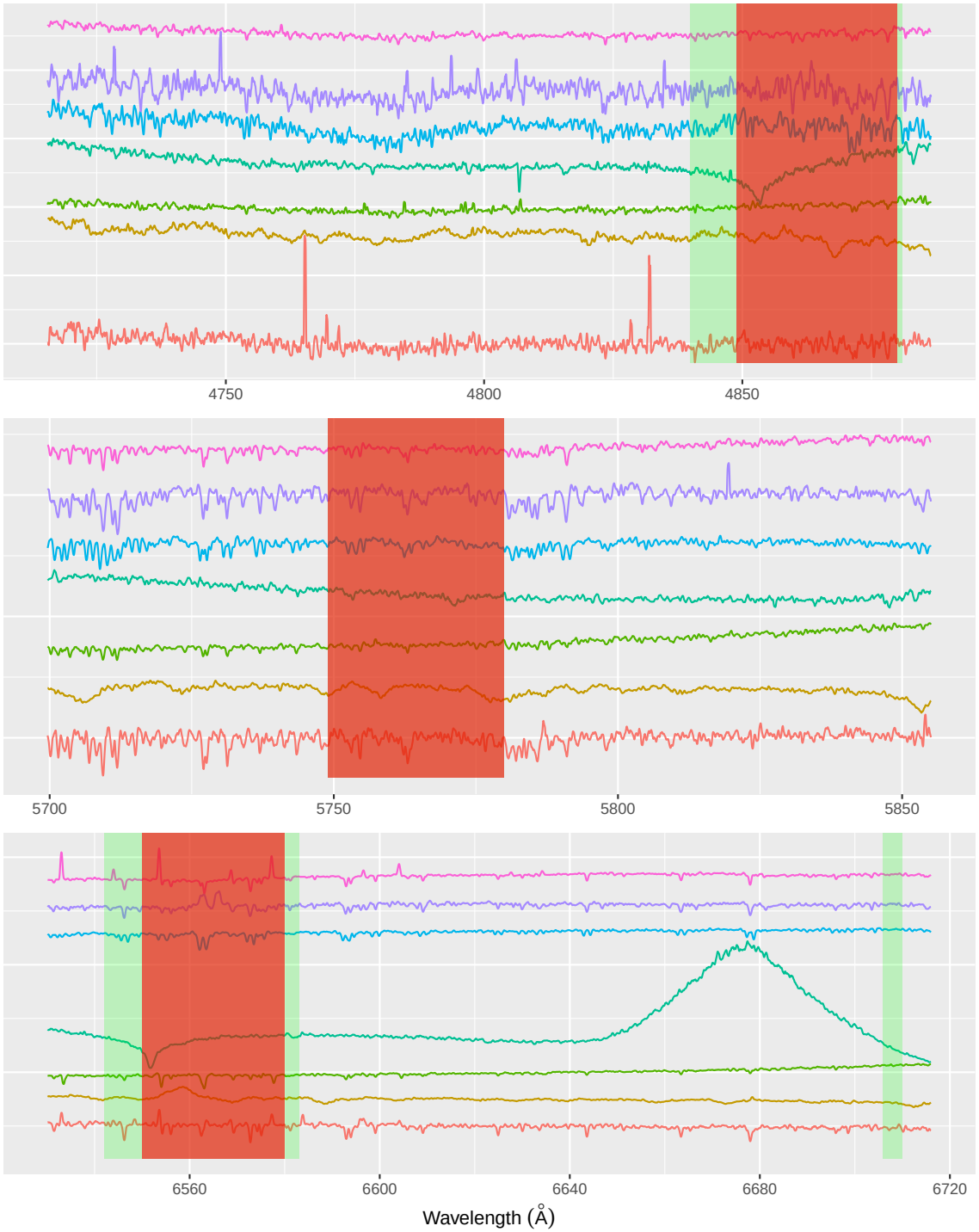


Figure 3.11: Seven randomly selected "zero"-metallicity candidates with the different wavelength regions as in Figure 3.10 shaded. One star has an almost perfectly flat spectra, with minimal hydrogen features. One selected star has significant emission spikes outside the wavelength range used in t-SNE. All stars, however, are significantly different and diverse, which suggests that the simulated flat line spectra may be effective at identifying odd outlying spectra. This is a line of investigation to explore in the future.

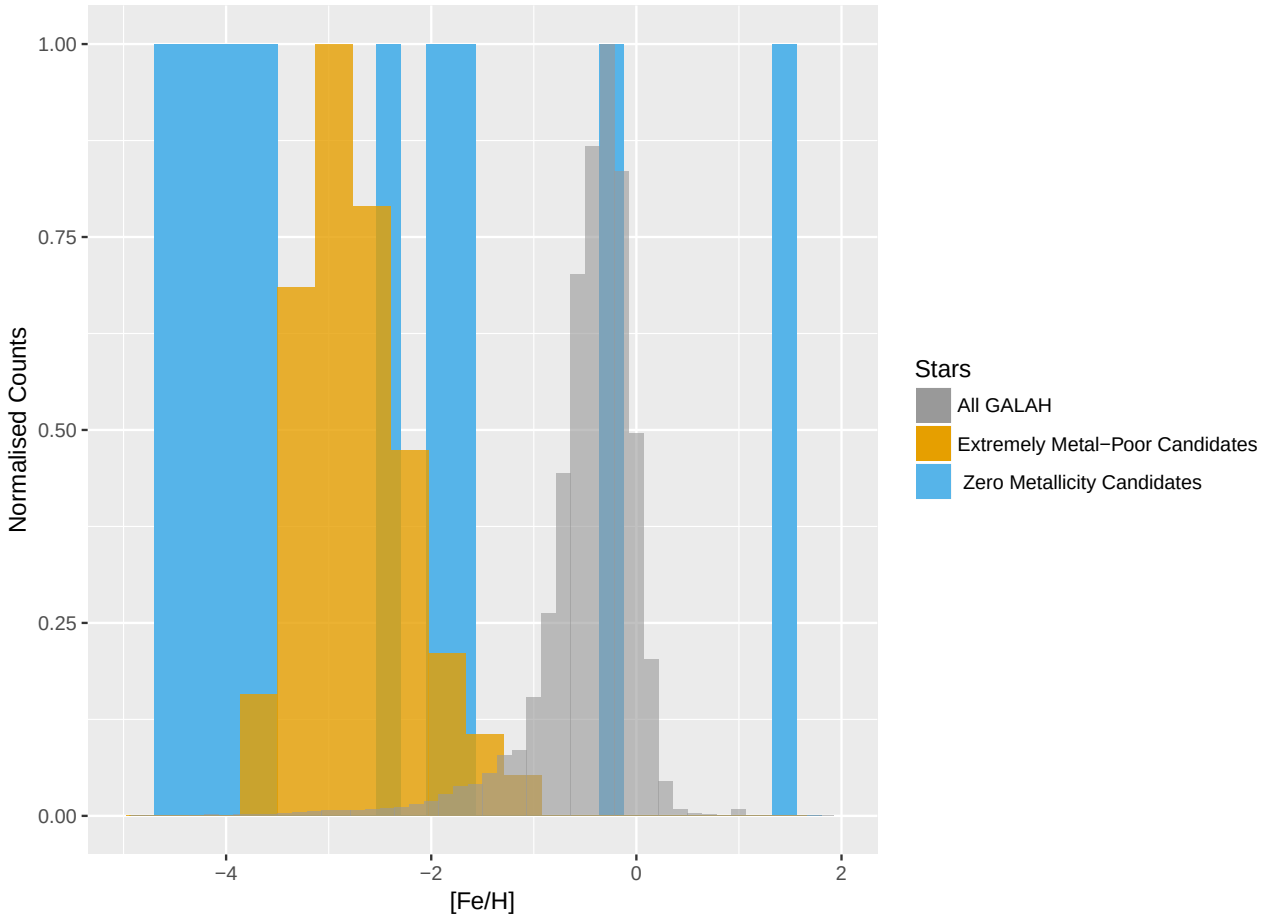


Figure 3.12: Metallicity distribution of the Extremely Metal-Poor and the 'Zero'-Metallicity candidates compared to all stars in the GALAH survey. The [Fe/H] values are those derived with GUESS. An important feature to note about GUESS is it utilises a fixed grid of templates to estimate the stellar parameters, which may lead to incorrect parameters. There is a clear distinction between the All GALAH and the Extremely Metal-Poor distributions. The randomness of the 'Zero'-Metallicity candidates may indicate that GUESS is failing to calculate accurate metallicities or reflects the 'Zero'-Metallicity cluster attracting random and diverse objects. This problem requires further investigation.

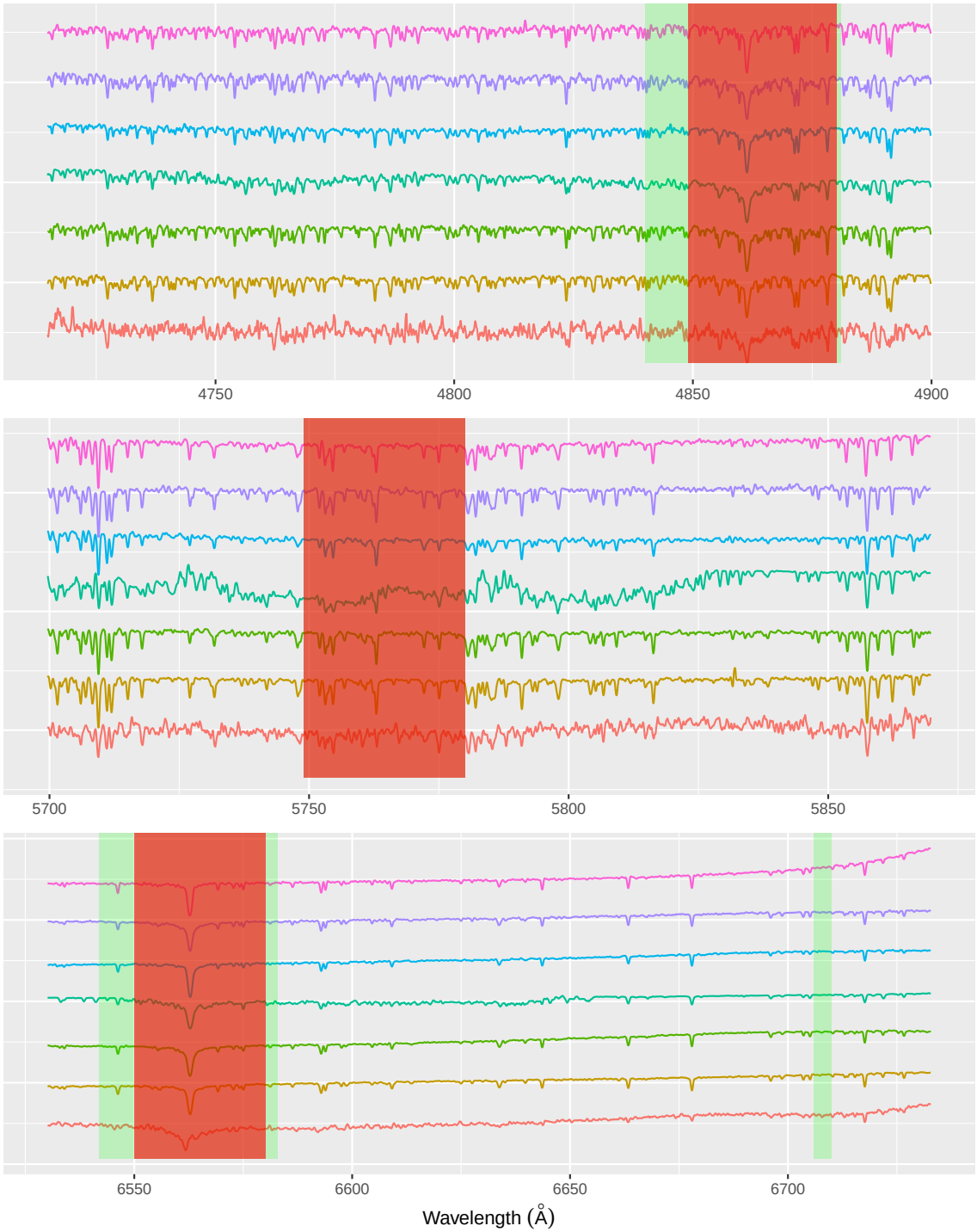


Figure 3.13: Seven randomly selected Solar twin candidates with the different wavelength regions as in Figure 3.10 shaded. See Figure 2.3 for an example Solar spectrum.

The initial exploration of the position, and stellar parameters of the Solar twins has produced some interesting results that will require a further inquiry. Figure 3.14 displays the stellar parameters of the Solar twins. The effective temperature and surface gravity plot ((b) & (c)) agree with the Suns parameters well. Analysing the metallicity plot ((d)) indicates the mean [Fe/H] is approximately Solar. The radial velocity panel in Figure 3.14 shows a very disperse range of velocities.

One hope for this type of work is to search for a long-disrupted star cluster in which the Sun formed. Figure 3.15 plots the positions of the solar twins relative to the most recent version of the GALAH observing footprint (S.Martell 2017, private communication, 30 August 2017)

It is tempting to start speculating that these candidates might have some sort of relationship with the Sun, including the possibility that they originate from the same primordial cloud. Given the distinct non-zero relative velocity compared to the Sun, another possibility is that these stars originated from a different primordial cloud with approximately Solar abundances that formed relatively recently (or not). This is an entirely cursory exploration, and more rigorous future work is required.

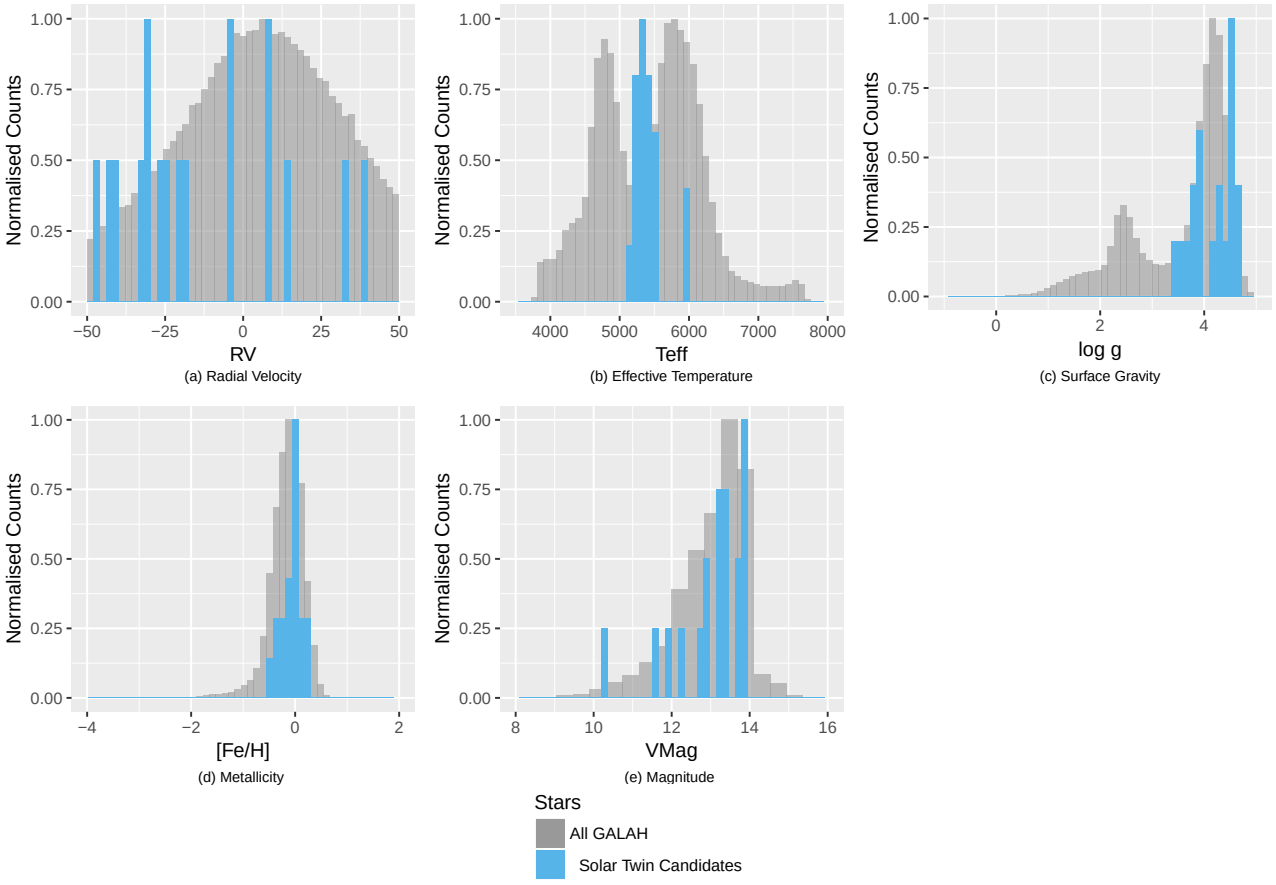


Figure 3.14: The stellar parameters of the solar twin candidates against all the entire observed GALAH stars. (a) is the distribution of radial velocity (km/s), (b) shows the effective temperature (K), (c) is the range of surface gravity (cm/s^2), (d) shows metallicity and (e) displays the magnitude in V band. The radial velocity distribution is scattered. The T_eff and log g show general agreement with the Solar values. The metallicity distribution shows a relatively narrow distribution, centred around 0. The rough agreement with solar parameters is perhaps indicating that the wavelength coverage might need to be expanded to include more metal lines (or that the GUESS parameters are only approximate).

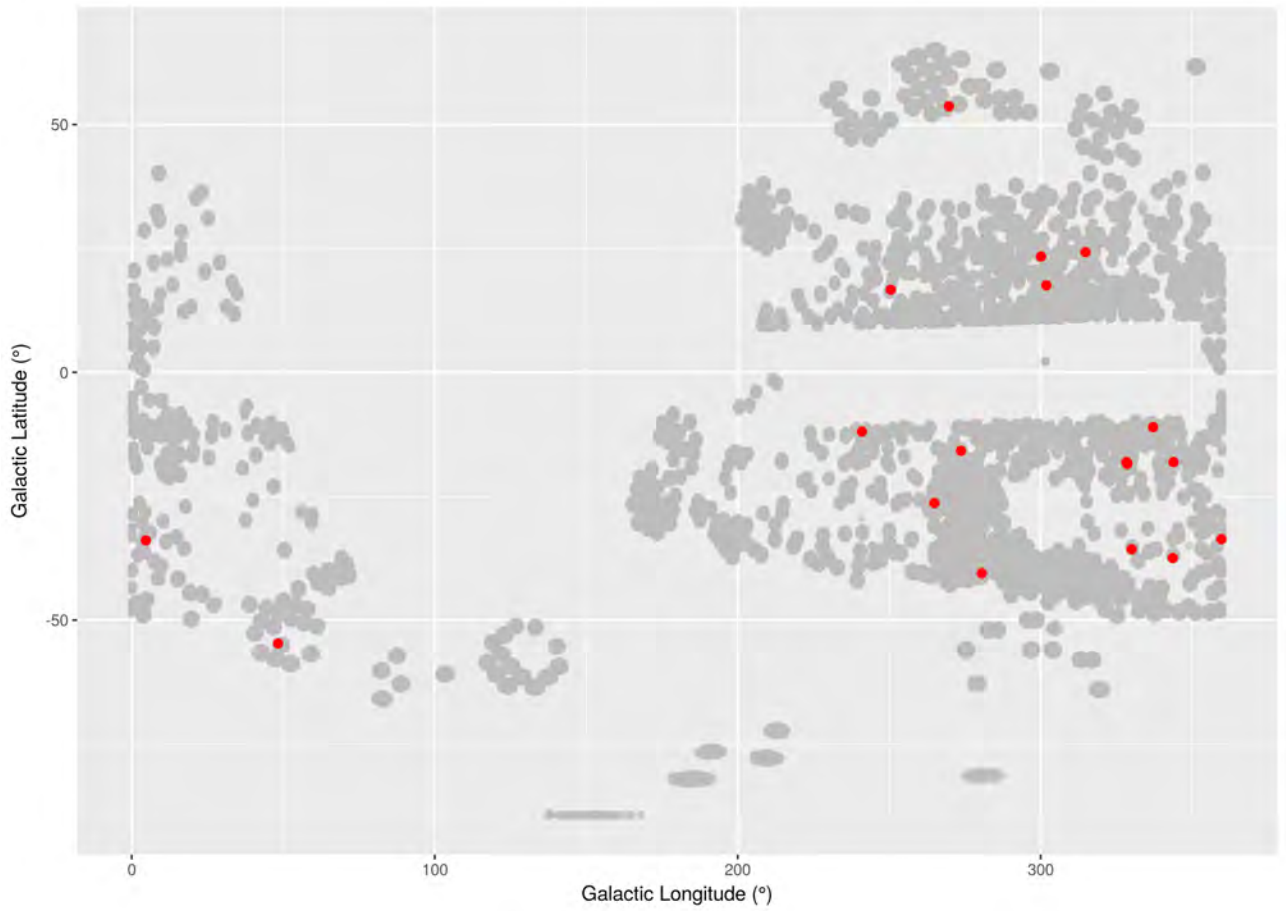


Figure 3.15: Distribution of possible Solar twin candidates. The grey circles represent the positions of all the observed GALAH stars, whereas the red circles represent the distribution of the possible Solar twins.

.

If cats looked like frogs we'd realize what nasty, cruel little bastards they are. Style. That's what people remember.

Terry Pratchett

4 Conclusion

4.1 Summary

This thesis presents a new methodology to search for rare, but interesting objects in the GALAH survey. The technique is based on the proposition that stars with similar spectra will lie in a similar space on a 2-D projection of the dimensionality reduction space produced by t-SNE.

A summary of the steps entailed in this methodology is presented below:

1. Resample spectra to a common wavelength range, normalise and smooth;
2. Dimensionally reduce the dataset using t-SNE to 2-dimensions to be able to visualise each spectra on a plane;
3. Choose the method by which to identify objects of interest in the projected dataset; this depends on how many of the interesting stars are provided as labelled data:
    - HDBSCAN - used when there is a high number of known interesting stars, as they are able to form dense clusters easily identifiable by HDBSCAN
    - t-SNE visualiser - used when there is a low number of known stars of interest, as the t-SNE visualiser can interactively plot the spectra for neighbouring stars on the t-SNE map.

Previous work using t-SNE focused on grouping similar stars, to estimate stellar parameters, or to determine spectral outliers. In these cases the t-SNE projection would be plotted by a stellar parameter, This would result in a more general understanding of a dataset, and is effective at classifying abundant stellar types within a dataset. The new method described in this thesis approaches classification from a different angle, by using a few known stellar classification labels to find regions of similar stars in a larger unlabelled dataset,to identify rare and unique stars of interest. t-SNE was the preferred dimensionality technique compared to other dimensionality reduction methods, like LLE, because of its superior ability to distinguish the finer structure in a dataset. HDBSCAN was the trialled clustering method because of its ability to identify clusters of a variety of shapes and densities, structures that a t-SNE projection can produce. The power of the methodology presented is its flexibility, simplicity and its ability to detect rare objects in data from large surveys.

This flexibility and ability is key, as it does not have to be used to just search for extremely metal-poor stars and solar twins specifically in the GALAH survey; in fact, the method may be adapted to searching for any type of astronomical object, as long as there are known examples in the data. Further examples of possible extensions of this methodology are discussed in Section 4.2 .

In addition to a new way to detect rare objects of interest in a large dataset, the application of this method uncovered 66 extremely metal-poor star candidates, 14 "zero"-metallicity star candidates and 20 solar twin candidates in the GALAH survey. As outlined in Chapter 1, finding extremely metal-poor stars will help to inform theories of galaxy formation and facilitate a better understanding of nucleosynthetic processes in the early Universe. The detection of Solar twins enables astronomers to better understand the Sun by providing analogues for comparison, as well as identifying potential Solar siblings from a long ago disrupted star cluster. These can furthermore be targeted for exosolar planet searches.

4.2 Future Work

The most immediate step is to follow up on the potential extremely metal-poor star and solar twin candidates, by further exploring their chemical abundances, metallicity and kinematics. These calculations can be later incorporated into the object detection or screening process, and integrating the functionality of accurate stellar abundance estimations into the t-SNE visualiser would be an invaluable extension.

When applying this methodology to future GALAH releases, a clearer specification of which stars to include or exclude – particularly with regards to S/N – will be required. The value of 25 was somewhat arbitrarily chosen for this work to keep a significant number of the labelled sample, but as the known GALAH sample increases the S/N cut may be tightened.

As the stellar classes of interest are equally well represented in the labelled GALAH sample, when the labelled GALAH sample increases there are possible alternative statistical methods that may be considered to identify stars of interest. Such methods include Artificial Neural Networks and Support Vector Machines, which require a larger training sample than is currently available. Neural networks have proven to be a significant tool capable of extracting reliable information and patterns from large amounts of data – even in the absence of models describing the data – and thus may provide a more accurate way of determining stellar classes.

Another result from this work is that t-SNE appears to be effective at finding finer details within a dataset. This can be useful for deriving a more detailed stellar classification system.

Moving beyond applying this method to GALAH and large high resolution spectroscopic datasets, the low-resolution Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) (Luo et al., 2004) spectroscopic survey has produced significant results studying the kinematics and dynamics of the Milky Way. A further application of the t-SNE methodology would be to identify interesting stars in the dataset produced by this survey.

Lastly, with the upcoming second Gaia data release (scheduled for April 2018) that will contain the stellar parameters and kinematics of a billion stars in the Galaxy, the methodology described may be adapted to add precise positional and velocity information to the high dimensional spectroscopic datasets in GALAH and other large surveys. This would make it possible to determine whether similar stars are kinematically associated, and even whether they share a common birthplace.

References

Abel, T., Bryan, G. L., & Norman, M. L. 2000, The Astrophysical Journal, 540, 39

Albareti, F. D., Prieto, C. A., Almeida, A., et al. 2016, arXiv preprint arXiv:1608.02013

An, D., Beers, T. C., Johnson, J. A., et al. 2013, The Astrophysical Journal, 763, 65, arXiv: 1211.7073

Babu, G. J., & McDermott, J. P. 2002, Astronomical Data Analysis II, 4847, 228

Barnes, J., & Hut, P. 1986, Nature, 324, 446

Bayes, M., & Price, M. 1763, Philosophical Transactions, 53, 370

Benítez, N. 2000, The Astrophysical Journal, 536, 571

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, The Astrophysical Journal Supplement Series, 208, 20

Boroson, T. A., & Green, R. F. 1992, The Astrophysical Journal Supplement Series, 80, 109

Bromm, V., Coppi, P. S., & Larson, R. B. 1999, The Astrophysical Journal, 527, L5, arXiv: astro-ph/9910224

Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II, ed. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160, dOI: 10.1007/978-3-642-37456-2_14

Cayrel, R., Hill, V., Beers, T. C., et al. 2001, Nature, 409, 691

Cayrel de Strobel, G., Knowles, N., Hernandez, G., & Bentolila, C. 1981, The Astrophysical Journal, 94, 1

Cohen, J. G., McWilliam, A., Christlieb, N., et al. 2007, The Astrophysical Journal Letters, 659, L161

Collister, A. A., & Lahav, O. 2004, Publications of the Astronomical Society of the Pacific, 116, 345

Daniel, S. F., Connolly, A., Schneider, J., Vanderplas, J., & Xiong, L. 2011, The Astronomical Journal, 142, 203

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 2604

de Strobel, G. C. 1996, The Astronomy and Astrophysics Review, 7, 243

Deeming, T. J. 1964, Monthly Notices of the Royal Astronomical Society, 127, 493

Eatough, R. P., Molkenthin, N., Kramer, M., et al. 2010, Monthly Notices of the Royal Astronomical Society, 407, 2443, arXiv: 1005.5068

Efstathiou, G., Ellis, R. S., & Peterson, B. A. 1988, Monthly Notices of the Royal Astronomical Society, 232, 431

Einstein, A. 1916, Annalen der Physik, 354, 769

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (Portland, Oregon: AAAI Press), 226

Feigelson, E. D. 2009, arXiv preprint arXiv:0903.0416

Fisher, R. A. 1922, Philos Trans R Soc Lond A, 222, 309

Frebel, A. 2008, arXiv:0802.1924 [astro-ph], arXiv: 0802.1924

Gagné, J., Lafrenière, D., Doyon, R., Malo, L., & Artigau, Ã. 2014, The Astrophysical Journal, 783, 121

Gauss, C. F., & Stewart, G. 1995, Theory of the Combination of Observations Least Subject to Errors, Part One, Part Two, Supplement (SIAM)
Hardorp, J. 1978, The Astrophysical Journal, 63, 383
Houck, J. C., & Denicola, L. A. 2000, in Astronomical Society of the Pacific Conference Series, Vol. 216, Astronomical Data Analysis Software and Systems IX, ed. N. Manset, C. Veillet, & D. Crabtree, 591
Hubble, E. 1929, Proceedings of the National Academy of Sciences, 15, 168
Hughes, G. 1968, IEEE transactions on information theory, 14, 55
Keller, S. C., Bessell, M. S., Frebel, A., et al. 2014, Nature, 506, 463
Konijn, R. M., & Kowalczyk, W. 2012, in International Conference on Hybrid Artificial Intelligence Systems (Springer), 174
Kos, J., Bland-Hawthorn, J., Freeman, K., et al. 2017, arXiv:1709.00794 [astro-ph], arXiv: 1709.00794
Lahav, O. 1996, arXiv preprint astro-ph/9612096
Lahav, O., Nairn, A., Sodré, L., J., & Storrie-Lombardi, M. C. 1996, Monthly Notices of the Royal Astronomical Society, 283, 207
Larson, R. B. 1998, Monthly Notices of the Royal Astronomical Society, 301, 569
Lin, J. 2015, PhD thesis, The Australian National University
Lucy, L. B. 1974, The Astronomical Journal, 79, 745
Luo, A.-L., Zhang, Y.-X., Zhang, J.-N., & Zhao, Y.-H. 2004, , 178
Maaten, L. v. d., & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579
MacQueen, J. 1967, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (Berkeley, Calif.: University of California Press), 281
Mahdi, D., Soubiran, C., Blanco-Cuaresma, S., & Chemin, L. 2016, Astronomy & Astrophysics, 587, A131
Martell, S. L., Sharma, S., Buder, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 465, 3203
Matijevič, G., Chiappini, C., Grebel, E. K., et al. 2017, Astronomy & Astrophysics, 603, A19, arXiv: 1704.05695
Palla, F. 1983, Mem. Societa Astronomica Italiana, 54, 235
Park, T., Kashyap, V. L., Siemiginowska, A., et al. 2006, The Astrophysical Journal, 652, 610
Paulovich, F., Eler, D., Poco, J., et al. 2011, Comput. Graph. Forum, 30, 1091
Plackett, R. L. 1958, Biometrika, 45, 130
Planck, M. 1901, Ann. Phys, 4, 90
Porto de Mello, G. F., & da Silva, L. 1997, The Astrophysical Journal, 482, L89
Postman, M., Coe, D., BenÃŋtez, N., et al. 2012, The Astrophysical Journal Supplement Series, 199, 25
Ryan, S. G., Norris, J. E., & Beers, T. C. 1999, The Astrophysical Journal, 523, 654, arXiv: astro-ph/9903059
Szalay, A., & Gray, J. 2001, Science, 293, 2037
Trager, S. C., Worthey, G., Faber, S. M., Burstein, D., & GonzÃąlez, J. J. 1998, The Astrophysical Journal Supplement Series, 116, 1
Traven, G., Matijevič, G., Kos, J., et al. 2017, arXiv preprint arXiv:1612.02242
Trotta, R. 2008, Contemporary Physics, 49, 71
Tumlinson, J. 2010, The Astrophysical Journal, 708, 1398, arXiv: 0911.1786
Wang, Y., & Rekaya, R. 2010, Biomarker insights, 5, 69
Wattenberg, M., ViÃľgas, F., & Johnson, I. 2016, Distill
Yeche, C., Petitjean, P., Rich, J., et al. 2010, Astronomy & Astrophysics, 523, A14