DEPARTAMENTO DE ASTROFISICA

Universidad de La Laguna

Automated algorithms for spectroscopic classification of stars and applications to APOGEE

Memoria que presenta
D. Rafael Augusto Garcia Dias
para optar al grado de
Doctor por la Universidad de La Laguna.

Directores:
Dr. Carlos Allende Prieto y Dr. Jorge Sánchez Almeida

INSTITUTO Đ ASTROFISICA Đ CANARIAS
Diciembre de 2018

Examination date: x, 2018
Thesis supervisors: Dr. Carlos Allende Prieto and Dr. Jorge Sánchez Almeida

ISBN: xx-xxx-xxxx-x
Depósito legal: TF-xxxx/2018
Part of the material included in this document has been already published in a paper in the journal *Astronomy & Astrophysics*.

Agradecimientos

...

Resumen

El gran volumen de datos generados por los surveys astronómicos modernas ofrece una oportunidad interesante para la aplicación de *machine learning*. Es esencial explorar todas las herramientas existentes y determinar cuáles son adecuadas para generar conocimiento científico a partir de la observación disponible.

El objetivo principal de esta tesis es explorar la aplicación de algoritmos de aprendizaje automático para el estudio Apache Point Galactic Evolution Experiment (APOGEE). A pesar de nuestro enfoque en APOGEE, esta tesis proporciona una guía para la aplicación de los mismos algoritmos a otros conjuntos de datos similares.

A lo largo de la tesis, utilizamos dos problemas astrofísicos para probar diferentes algoritmos de aprendizaje automático. Primero, abordamos la clasificación espectral de más de 150,000 estrellas con K-means. Proporcionamos un análisis detallado de las virtudes y limitaciones del algoritmo, y una descripción extensa de las clases generadas con K-means. En segundo lugar, nos centramos en el análisis de abundancias químicas probando ocho algoritmos de agrupación para explorar la viabilidad de una búsqueda a ciegas de poblaciones estelares en el espacio químico. Usando pruebas estadísticas, demostramos que algunos cúmulos de estrellas son indistinguibles entre sí en el espacio de abundancia de APOGEE. Con este resultado, ampliamos la noción de etiquetado químico a la búsqueda de poblaciones estelares a un nivel más allá de los cúmulos estelares, a la cual llamamos *familias de estrellas*. Finalmente, utilizamos el algoritmo t-distributed stochastic neighbor embedding (t-SNE) para proyectar el espacio de 13 dimensiones de las abundancias químicas de APOGEE en 2 dimensiones. Clasificamos las estrellas en esta proyección usando Density-Based Spatial Clustering of Applications with Noise (DBSCAN) y mostramos que la clasificación puede identificar familias de estrellas relacionadas con poblaciones estelares conocidas.

Esta tesis ofrece una visión general del conjunto de datos APOGEE y también una guía para la aplicación de algunos algoritmos de agrupación no supervisados.

Abstract

The vast volume of data generated by modern astronomical surveys offers an interesting opportunity for the application of *machine learning*. It is essential to explore all the existing tools and determine which ones are suitable to generate scientific knowledge from the available observation.

The primary objective of this thesis is to explore the application of machine learning algorithms to the Apache Point Galactic Evolution Experiment (APOGEE) survey. Despite our focus on APOGEE survey, this thesis provides a guide for the application of the same algorithms to other similar datasets.

Throughout the thesis, we use two astrophysical problems to test different machine learning algorithms. First, we address the spectral classification of more than 150,000 stars with K-means. We provide a detailed analysis of the virtues and limitations of the algorithm, and an extensive description of the classes generated with K-means. Second, we focus on the analysis of chemical abundances using eight different clustering algorithms to explore the feasibility of a blind search for stellar populations in chemical space. Using statistical tests, we demonstrate that of some known stellar clusters are indistinguishable from each other in the space of abundances from APOGEE. With this result in hand, we expand the notion of chemical tagging to the search for stellar populations at a level beyond stellar clusters, which we call *families of stars*. Finally, we use t-distributed stochastic neighbor embedding (t-SNE) to project the 13-dimensional space of chemical abundances from APOGEE on 2 dimensions. We classify the stars in this projection using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and show that the classification can identify families of stars related to known stellar populations.

This thesis provides an overview of the APOGEE dataset and also a guide to the application of unsupervised clustering algorithms to similar problems.

Contents

1

Introduction

Understanding the formation and evolution of galaxies is an active field of research in modern astrophysics. The Milky Way (often called the Galaxy) represents a Rosseta-stone in these studies since it allow us to examine with unique details common fundamental processes unobservable in other galaxies. Early on, it became clear that to study the Galaxy we need to survey a large number of stars. In the 18th century William Herschel's survey explored the sky positions of more than two thousand stars, trying to determine the spatial structure of the Galaxy. In modern astronomy this idea was expanded not only by measuring the sky positions of more stars with better precision, but also exploring other aspects of the stellar distribution, such as measurements of distance, ages and chemical composition throughout the Milky Way. The Apache Point Galactic Evolution Experiment (APOGEE) (Majewski et al., 2017) provides chemical compositions of more than 300 thousand stars in throughout the Galaxy. Figure 1.1 shows the map created by William Herschel in comparison with our current knowledge of the Galaxy.

Although the data generated by surveys like APOGEE is fundamental to understand the formation and evolution of our Galaxy, the overwhelming data volume presents a challenge. The future of astronomy depends on the development of efficient algorithms capable of making the best of all available data. In some cases, the data acquisition itself depends on the algorithms that judge the relevance of the data, being the only way to analysis in many practical cases (Gaia Collaboration et al., 2016; Gressler et al., 2014). In this Thesis, we will consider the applications of pattern recognition techniques to large astronomical surveys.

1.1 Spectroscopy

Even before Newton's experiments with prisms and solar light in the 17th century humanity was intrigued by the study of matter-light interaction, but only with the findings of Fraunhofer, Wheatstone, and Kirchhoff in the 19th century the rela-

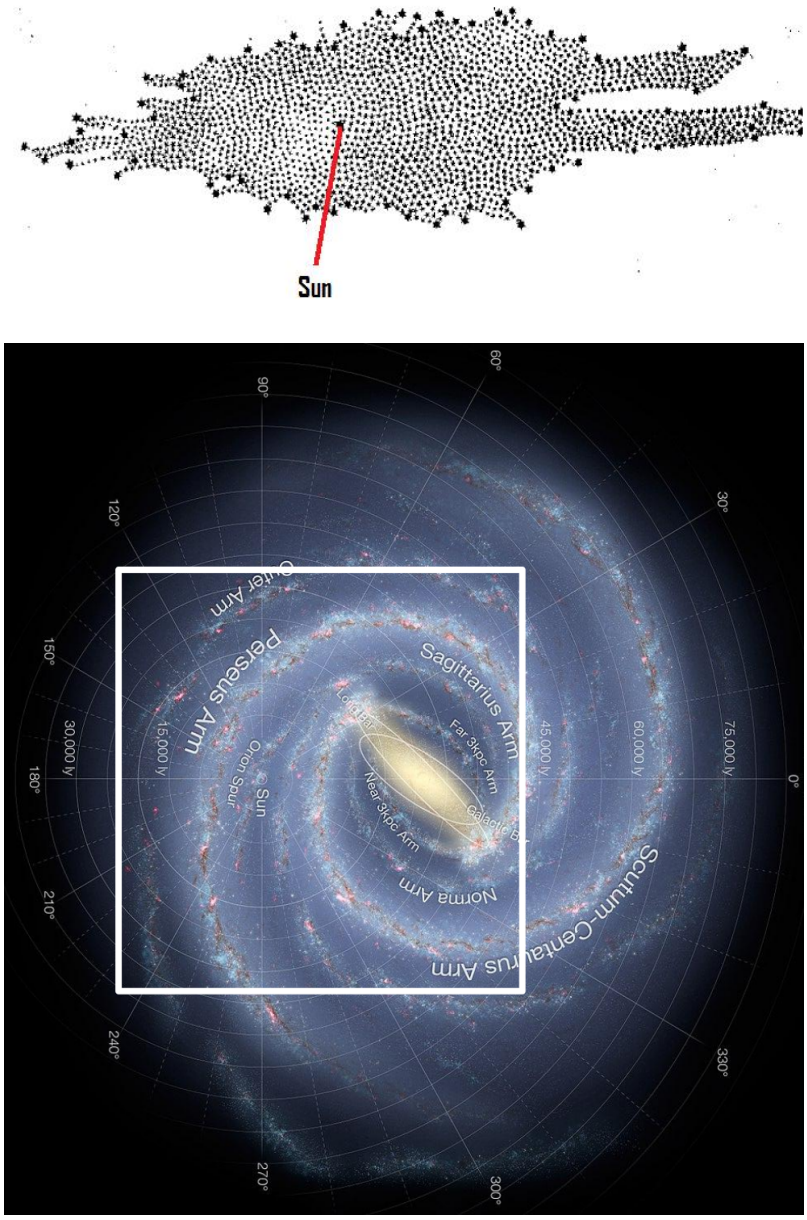

FIGURE 1.1— In the top panel we show the drawing of William Hershel's survey of stars, and in the bottom panel we show a illustration the current knowledge about Milky Way's structure, indicating the main spiral arms of the Galaxy. A white square shows the region covered by APOGEE. Source: https://www.eso.org/public/images/eso1339e/

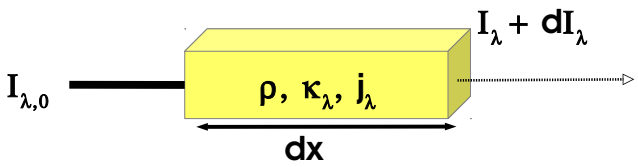

FIGURE 1.2— The diagram illustrates a beam of light traveling through a volume of gas. The mathematical symbols are defined throughout the text.

tion between spectral lines and chemical elements became clear. The Kirchhoff laws of spectroscopy stated that:

- An incandescent solid, liquid or gas under high pressure emits a continuous spectrum.
- A hot gas under low pressure emits a "bright-line" or emission-line spectrum.
- A continuous spectrum source viewed through a cool, low-density gas produces an absorption-line spectrum.

Kirchhoff demonstrated experimentally that different gases generate a distinct pattern of spectral absorption lines in a continuous spectrum. Figure 1.2 presents a diagram for a monochromatic beam of light with initial specific intensity $I_{\lambda,0}$ [erg s^{-1} cm^{-2} rad^{-2} Å^{-1}] and wavelength λ, passing through a volume of gas with density ρ [g cm^{-3}]. When the beam passes through the gas, it suffers an attenuation due to the absorption of photons by the atoms in the gas. Each atom can absorb light in a specific series of wavelengths according to its electronic structure. Also, some emission may be produced by the gas, as states the second law of Kirchhoff, and also by scattered light in the direction of the beam. These processes are described by the specific absorption and emission coefficients (κ_λ [cm^2 g^{-1}], j_λ [erg s^{-1} rad^{-2} Å^{-1} g^{-1}]). We can write the variation of the intensity of the beam, after traveling a distance dx through the gas, as

$$dI_\lambda = j_\lambda \rho dx - \kappa_\lambda \rho I_\lambda dx. \tag{1.1}$$

If we define the optical depth as the absorption experienced by the beam passing through a layer of length L,

$$d\tau_\lambda = \kappa_\lambda \rho dx \rightarrow \tau_\lambda = \int_0^L \kappa_\lambda \rho dx, \tag{1.2}$$

we can rewrite equation 1.1 as

$$dI_\lambda = \kappa_\lambda \rho \left(\frac{j_\lambda}{\kappa_\lambda} - I_\lambda \right) dx = \left(\frac{j_\lambda}{\kappa_\lambda} - I_\lambda \right) d\tau_\lambda. \tag{1.3}$$

In local thermodynamic equilibrium the ratio j_λ/κ_λ is given by the Plank function,

$$\left[\frac{j_\lambda}{\kappa_\lambda} \right]_{LTE} = B_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1}. \tag{1.4}$$

This makes it possible to rewrite equation 1.3 as

$$dI_\lambda(\tau_\lambda) = \left(\frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1} - I_\lambda \right) d\tau_\lambda. \tag{1.5}$$

The proper solution of this equation needs to take into account many essential aspects of radiative transfer such as geometry approximations, among many other important theoretical and practical issues. However, it suffices to get some intuition of how spectral lines are formed in stellar atmospheres and how lines relate to the physical parameters of a star. A rigorous approach to this problem can be found in books such as Gray (2008) or Hubeny & Mihalas (2014). Modeling the atmospheres of stars to determine how different phenomena affect line formation is an active research field. Numerical calculations provide astronomers with model atmospheres[1] that can be used to infer the physical properties and chemical composition of the stars. For example, works such as Castelli & Kurucz (2004) or Gustafsson et al. (2008) provide libraries of model atmospheres widely used in the literature.

In Figure 1.3 we show the spectra of six stars with very similar surface gravity ($\log g$) and effective temperature ($T_{\rm eff}$), but with different iron abundances ([Fe/H])[2]. We see that the intensity of the spectral lines increases with the iron abundance. That is to say, with a higher concentration of iron atoms in the atmosphere of the star, there is more absorption at wavelengths coinciding with the iron atomic transitions, generating deeper absorption lines in the spectra. As we will show in detail in Chapter 2, the shape of the spectra also depends on the $T_{\rm eff}$ and $\log g$ of the star, but here we have selected spectra with very similar atmospheric parameters to highlight the effect of chemical abundance in the spectra.

1.2 Chemical evolution of the Galaxy

The Big Bang formed mainly H and He in the so called primordial nucleosynthesis, and basically all other elements in the periodic table are generated by stellar

[1] Model atmospheres provides the stratification of physical parameters (temperature and density) from which one derives I_λ. Depending of the degree of sophistication, they can comprise only the photosphere or photosphere and the outermost layers. They can be 1D or in a few cases 3D.

[2] The brackets notation for two given elements X and Y, [X/Y] is defined as:

$$[\mathrm{X/Y}] = \log_{10} \left(\frac{N_\mathrm{X}}{N_\mathrm{Y}} \right)_{\rm star} - \log_{10} \left(\frac{N_\mathrm{X}}{N_\mathrm{Y}} \right)_{\odot}, \tag{1.6}$$

where N_X and N_Y are the number of X and Y nuclei per unit volume, respectively.

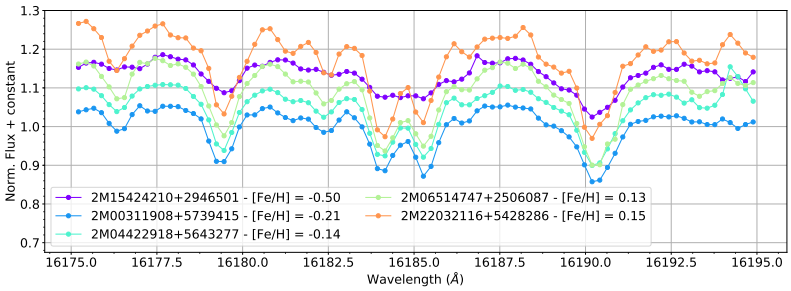

FIGURE 1.3— The figure shows part of the APOGEE spectra for six stars with same surface gravity ($\log g = 2.5$) and same effective temperature ($T_{\rm eff} = 4875$ K), but with different iron abundances ([Fe/H], using the bracket notation as defined in Equation 1.6).

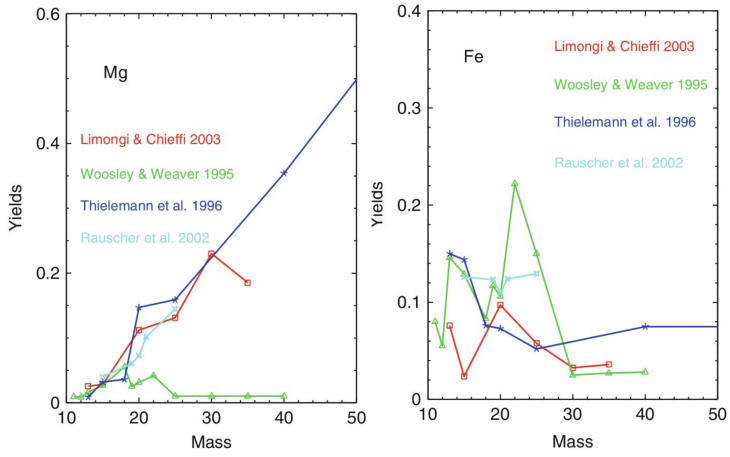

FIGURE 1.4— The Figure shows how the mass of the progenitors of type II supernova (core-collapse) are related to their production and posterior contamination of the interstellar medium with different chemical elements. The left panel shows the yields (the mass of metals eventually ejected to the ISM per unit mass locked into stars) of magnesium in function of the mass of the stars in solar masses, while the right panel presents the same quantity for the iron element. Figure from Matteucci (2012).

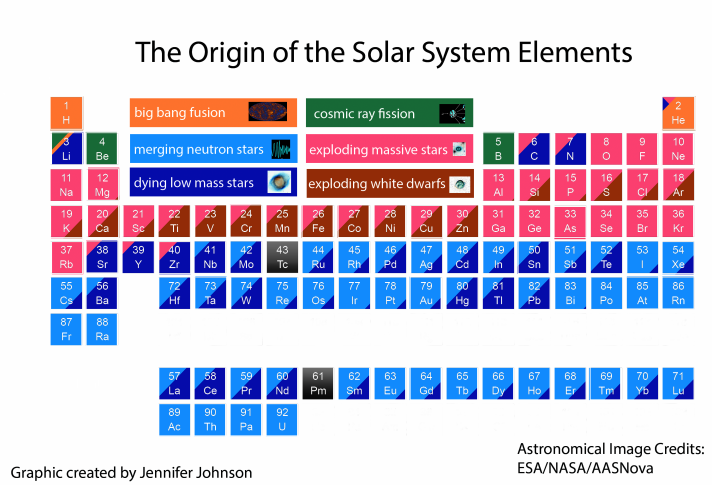

FIGURE 1.5— The periodic table of the elements colored by their cosmic origin.

nucleosynthesis. Some elements have more complicated origins, for instance, some Li was also formed in the Big Bang, it is destroyed in some stars, and 20-30% of the Li in the solar system was formed in the interstellar medium (ISM) involving high energy events with Galactic cosmic rays (Reeves et al., 1970). Be is also mainly formed in the ISM by the same mechanism. Stars can produce different chemical elements depending on its mass, Figure 1.4 shows how stars with different masses can produce and return to the ISM different quantities of Mg and Fe. Violent events like supernova are the main nucleosynthesis mechanism for many heavy elements. Figure 1.5 summarizes the production mechanisms for all the elements in the periodic table.

As mentioned in Section 1.1, we can measure the chemical composition of stellar atmospheres from the absorption lines in stellar spectra. The atmosphere of a star may retain the original composition of the ISM gas from which the star was formed, or suffer contamination in elements such as N and C from its core. Then, knowing the composition of the atmospheres of stars throughout the Milky Way can shed light on relevant aspects of Galaxy formation and evolution.

APOGEE aims to use the chemical composition a large number of stars ($\approx 10^5$) to disentangle the multiple processes that led to the current state of the Galaxy (e.g., Freeman & Bland-Hawthorn 2002; Hayden et al. 2015). Since APOGEE is arguably the most detailed survey of chemical abundances throughout the Galaxy, we use its data in the Thesis to search for patterns that can give us insights on the formation and evolution of the Galaxy.

1.3 APOGEE data: spectra and abundances

APOGEE makes use of a novel fiber-fed high-resolution $H-$band spectrograph to obtain simultaneously up to 300 stellar spectra (Wilson et al., 2012). The APOGEE spectrograph is usually coupled to the Sloan Foundation 2.5-m telescope at Apache Point Observatory, but it has also been linked to the New Mexico State University 1-m telescope at the same location. The project has already obtained spectra for more than 300,000 stars in the Milky Way, focusing on red giants and therefore covering a broad range of galactocentric distances, from tens of parsecs to more than 20.000 parsecs. Working in the near-IR, between 1.5 and 1.7 μm, APOGEE can access regions of the Galaxy heavily obscured by dust, such as the midplane of the Galaxy, the bulge, and the Galactic bar near the center (Majewski et al., 2017). The resolving power of the APOGEE data is $R \equiv \lambda/\delta\lambda \simeq 22,500$, and the typical signal-to-noise ratio exceeds 100 per half a resolution element.

APOGEE spectra are processed by a custom-made data pipeline that extracts the spectra, calibrates them, and corrects telluric absorption and sky emission lines before measuring radial velocities (Nidever et al., 2015). The Stellar Parameters and Chemical Abundances Pipeline (ASPCAP, García Pérez et al. 2016) performs an automated analysis based on model atmospheres, delivering atmospheric parameters and chemical abundances for the majority of the observed stars.

The APOGEE pipelines are in constant evolution, and the dataset continues to grow. In the first part of this thesis, presented in Chapter 2, we have adopted the data publicly available in DR12[3], which is the final data release from SDSS-III (Alam et al. 2015; Holtzman et al. 2015), and contains over 150,000 stars observed between 2011 and 2014. This was the latest stable version of the APOGEE dataset when we started the Thesis. In Chapters 3 and 4, we adopted the results in Data Release (DR14) (Abolfathi et al., 2018), which includes observations of more than 260,000 stars.

The main process of abundance determination is very similar in the two data releases. First, a grid of models suitable for the specifications of the APOGEE survey is established. The grid is based on the ATLAS9 models (Castelli & Kurucz, 2004) and is described in detail in the work of Mészáros et al. (2012). It has six or seven free parameters (depending in the version of the ASPCAP), $T_{\rm eff}$, $\log g$, metallicity ([M/H]), α abundance ([α/M]), [C/H], and [N/H], assuming solar abundances from Asplund et al. (2005). Metallicity is a measure of all the chemical elements heavier than He, assuming they vary in the same proportions with respect to the solar values. Analogously, [α/M] is a measure of all α-elements (O, Ne, Mg, Si, S, Ar, Ca, and Ti) assuming they vary in union. The microturbulence (χ_v) is either a free parameter, in DR14, or it is inferred from a linear relation with $\log g$, in DR12. Each model atmosphere describes the variations of optical depth (τ), temperature (T), gas pressure (P_g) and the electron density (N_e) in the stellar atmosphere. The ASSϵt code (Koesterke et al., 2008; Koesterke, 2009) integrates the equation

[3]The catalog is available at http://data.sdss3.org/sas/dr12/apogee/spectro/redux/r5/allStar-v603.fits.

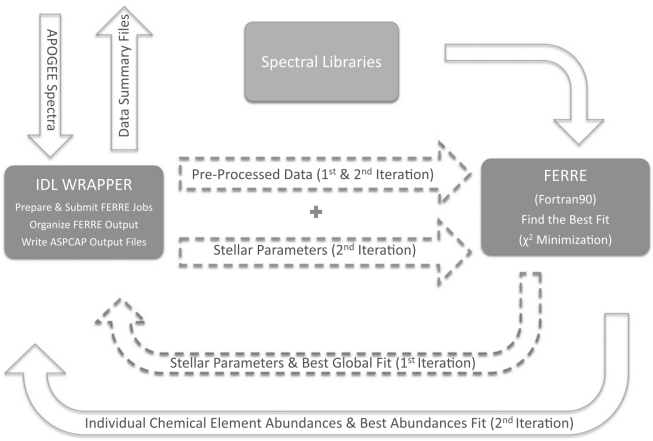

FIGURE 1.6— The flow chart describes the main steps of the abundance determinations in the ASPCAP. Dashed lines represent the first phase of the fitting and solid lines represent the second. Adapted from García Pérez et al. (2016)

of radiative transfer (eq. 1.3) through the model atmospheres, with the spectral line list described in Shetrone et al. (2015), to generate a grid of synthetic spectra. Finally, the grid of synthetic spectra is compared with the spectra of the stars to determine the model atmosphere that provides the best fit to observations through a χ^2 optimization,

$$\chi^2 = \sum_{\lambda} \frac{(O_\lambda - F_\lambda)^2}{\sigma_\lambda^2}. \tag{1.7}$$

O_λ is the observed flux, F_λ is the flux of the synthetic spectra and σ_λ are weights based on the flux uncertainties measured during the data reduction. Continuum-normalized spectra are used in this comparison.

The search for the best solution is done with FERRE[4] (Allende Prieto et al. 2004, 2006), varying the models by interpolation in the grid of synthetic spectra. The solution is determined in two phases. First, the code searches for the best solution considering the entire APOGEE spectral range and determining the six or seven free parameters ($T_{\rm eff}$, $\log g$, [M/H], [α/M], [C/H], [N/H], and χ_v if its is the case), and then the second round of iterations is performed to determine the individual chemical abundances of 26 element species, C, CI, N, O, Na, Mg, Al, Si, P, S, K, Ca, Ti, TiII, V, Cr, Mn, Fe, Co, Ni, Cu, Ge, Ce, Rb, Y, and Nd. The second phase is carried out by using specific spectral windows for each element. In this phase, five parameters remain fixed while only one parameter is varied to search

[4]https://github.com/callendeprieto/ferre

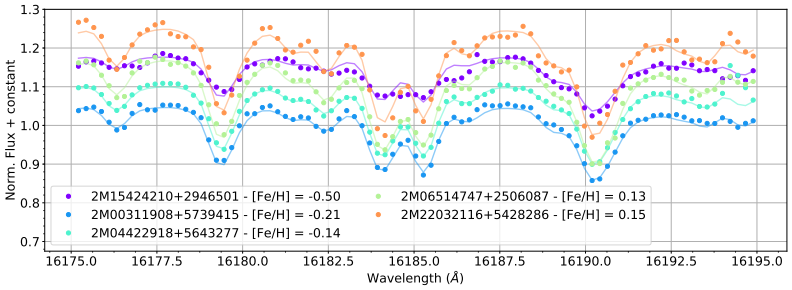

FIGURE 1.7— The figure shows part of the APOGEE spectra and their best fit for six stars with same surface gravity ($\log g = 2.5$) and same effective temperature ($T_{\rm eff} = 4875$ K), but with different iron abundances. The observed spectra are plotted with dots, while the best fit is plotted with solid lines with the same color. To help visualization we have artificially shifted each spectra in the vertical direction.

for the best solution. In the case of carbon and nitrogen [C/M] and [N/M] are used to determine the best solution in the windows of this elements. All the α-elements are determined by optimizing [α/M], and the other elements are determined based on the value of [M/H]. The windows used in each element are determined based on the partial derivatives of the flux in a grid synthetic of spectra with fixed atmospheric parameters and varying elemental abundances (i.e., computing the change of flux produced by a change in the abundance of each element, and selecting windows where this change is significant.). Figure 1.6 illustrates the two main phases we have described here. In Figure 1.7 we show the result of this process to the five stars in figure 1.3, dotted lines show the best fit for each spectra.

More details about the data used in each part of the work are given in the following chapters.

1.4 Machine learning algorithms

Statistical tools capable of finding patterns and generalizing rules from data, without being explicitly programmed to do so, are often called machine learning algorithms. There are two main kinds of machine learning algorithms, supervised and unsupervised ones. Supervised algorithms are those which use a set of input data and their corresponding desired output to learn how to map new input data into outputs. For example, in Ness et al. (2015) they use APOGEE spectra as input data and the chemical abundances and stellar parameters from ASPCAP as output data to train an algorithm capable of predicting the parameters and stellar abundances of unlabeled APOGEE spectra. On the oder hand, the unsupervised approach does not require the output examples. It searches for patterns in the input data and defines the most prominent features in the data. In this work, we are interested in exploring unlabeled data and search for unusual objects. Thus, we will focus on

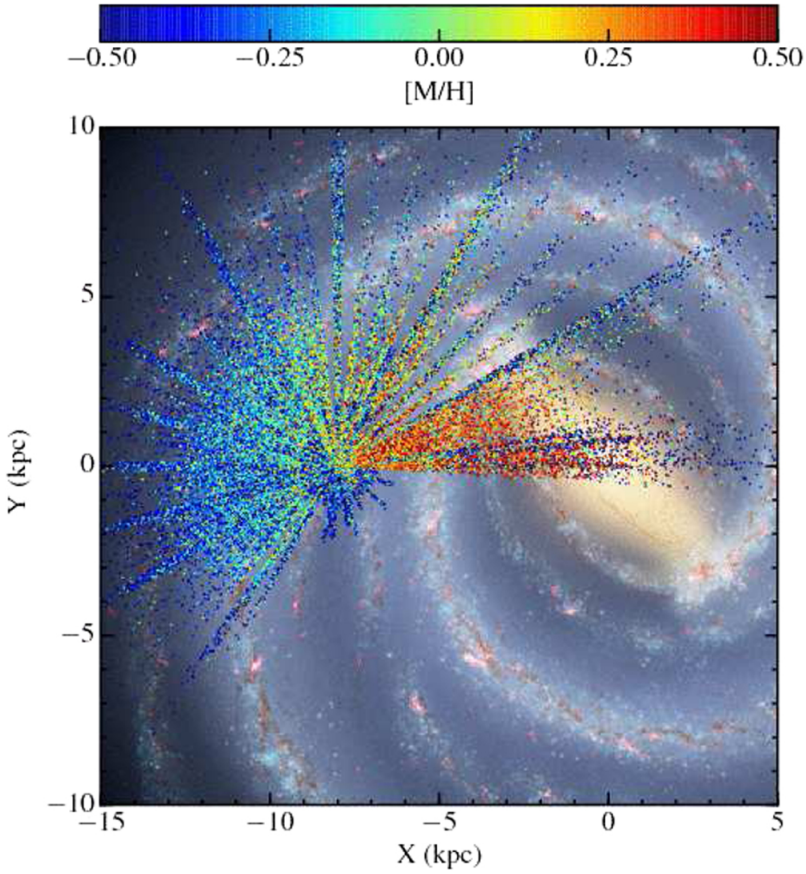

FIGURE 1.8— The figure shows the metallicity distribution of the stars in DR12 throughout the Galaxy. The image overlaps the data from DR12 to the region inside of the white square in Figure 1.1. Figure from Majewski et al. (2017).

unsupervised algorithms.

Cluster analysis is a particular kind of unsupervised algorithm which aims at organizing a collection of objects into classes based on a similarity criterion, such that objects in the same class are more alike than objects in different classes. There is a numerous set of cluster algorithms available in the literature (e.g., Everitt et al. 2011), but in general, all involve the following main steps: (1) feature selection, the identification of the features that better represent the objects in the dataset; (2) choosing a similarity metric, the figure of merit that optimally defines the similarity between objects in the dataset; (3) establishing the grouping criterion - meaning the clustering algorithm itself, and (4) cluster validation, an evaluation of the output quality. They are explained in some detail in the forthcoming subsections.

1.4.1 Feature selection

Let S be a set of n objects, $S = \{\vec{x}_0, \vec{x}_1, ..., \vec{x}_n\}$, each defined for a series of N parameters, $\vec{x_i} = (a_1, a_2, ..., a_N)$. Feature selection aims at defining a subset of the parameters of $\vec{x}$ which are the most relevant for a particular task, $(a_i, a_j, ...) \subseteq (a_1, a_2, ..., a_N)$. This can be done with a priori knowledge of the field, or through some feature selection algorithms. One example of feature selection is the choice of chemical abundances to identify stellar populations formed in a single burst. Although we have H-band apparent magnitudes for all the stars in APOGEE, grouping the stars through their apparent magnitude would not map populations sharing a common origin, since their apparent magnitudes depend on their distance, radius, and temperature. None of these properties are constrained in groups of stars that were formed together. On the other hand, we know that stars in stellar clusters originate from the same molecular cloud and so share similar chemical properties. Thus, working with chemical abundances and not with apparent magnitudes to identify stellar populations is a form of feature selection.

1.4.2 Similarity metric

The similarity metric is the measure of how alike two objects are. The most straightforward similarity metric among objects defined by numerical quantities is the Euclidean distance, defined as:

$$d(\vec{x_i}, \vec{x_j}) = \sqrt{(a_{i1} - a_{j1})^2 + (a_{i2} - a_{j2})^2 + \cdots + (a_{in} - a_{jn})^2}. \quad (1.8)$$

In comparison with other metrics, the Euclidean distance has the advantage of being easy to interpret, simple to implement, and having a low computational cost.

1.4.3 Grouping criteria

The grouping criterion is the way one assigns each object to a particular cluster and is how groups are constructed. For example, groups can be selected as a single partition, that is to say, all clusters are simple groups, hierarchically equivalent samples. Otherwise, they would be hierarchical clusters that have a structure with

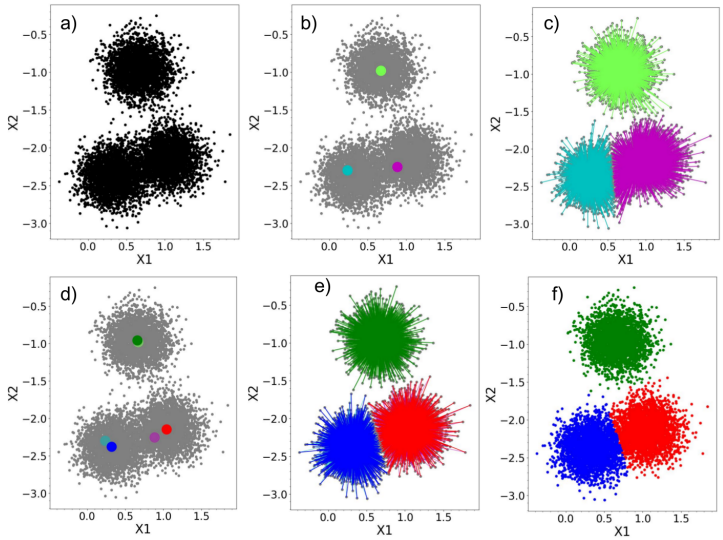

FIGURE 1.9— This figure illustrates the steps in K-means cluster. Panel (a) shows the unlabeled data, panel (b) shows the choice of the cluster centers guess, panel (c) illustrate the Euclidean distance measurements from the centers to each point, panel (d) shows how the centers are updated, panel (e) represents the iteration of panels (c) and (d), and panel (f) shows the result of the algorithm converged.

clusters and sub-clusters. Furthermore, clustering is said to be hard if it assigns each object to a single cluster, as opposed to soft clustering, where objects may be given a non-zero probability of belonging to more than one cluster.

For example, K-means (MacQueen et al., 1967) is a partitional hard clustering algorithm. It is one of the most popular clustering algorithms, mainly because it is easy to implement and its computational cost scales linearly with the number of objects to be classified. The fundamental steps in K-means are (1) to choose the number of clusters K; (2) define K initial cluster centers; (3) assign each object in the sample to the closest cluster; (4) recompute cluster centers as the centroid of the objects assigned to each cluster; and (5) repeat steps 3 and 4 until a convergence criterion is met. Usually, the convergence criterion is either a decrease in the within-cluster variance under a threshold or a minimal re-assignation between two consecutive iterations. In the Thesis we adopt the rule of having less than one percent of re-assignation between two consecutive iterations.

Figure 1.9 illustrates these steps. In panel (a) we show one thousand points of three bidimensional random Gaussian distributions. In panel (b) we represent the initial guess of the cluster centers by randomly choosing three objects in the whole

sample. In the panel (c) we present the measure of the distances between each point and each center, painting each point with the same color as the closest center. In panel (d) of the same figure we show how the cluster centers are updated with the mean of all the objects assigned to each cluster. Panel (e) represents the iteration of steps (3) and (4) until convergence, while panel (f) shows the final result for the application of K-means to this dataset.

1.5 Organization of the Thesis

We are presenting a Thesis which is exploratory in nature. The era of Big Data analysis is in its childhood, and the best tools to carry out particular astronomical tasks have not been identified yet. Probably, they have not been proposed yet. In the thesis we use a huge dataset of stellar spectra, and try to extract astrophysically relevant information in a way that has not been tried yet. Thus the final organization of the chapters in the thesis is mainly dictated by the APOGEE spectra and the algorithms at hand. In Chapter 2 we apply K-means to perform clustering analysis of APOGEE spectra. In Chapter 3 we work with K-means and other seven clustering algorithms that will be presented in the following chapters. The final chapter of the thesis contains a summary of the findings and conclusions.

2

Automated classification of stellar spectra[1]

2.1 Introduction

Classification is the first step in any automated analysis. It can be used to identify and discard noisy data, or to group similar objects to follow a common interpretation pipeline. It is certainly needed when exploring new types of data and it is also an invaluable tool to identify rare objects, usually the most telling from a scientific point of view. In spectroscopic surveys such as APOGEE, there are two main approaches to this problem. It is possible to use the stellar spectra to find patterns or classes in the stellar populations or, instead, it is possible to work with the atmospheric parameters and chemical abundances.

At present, all the available models have limitations. Either they are limited by the lack of knowledge on the input physical data (e.g., oscillation strengths), or they are limited by the approximations made to accelerate or simplify the calculations, such as plane-parallel geometry and local thermodynamic equilibrium. In this sense, the legacy of the MK classification (Morgan et al., 1943) to stellar astrophysics is undeniable. The classification is based on spectral features easily identifiable by visual inspection in medium-resolution spectra. It benefits from being a method independent of model atmospheres but has the downside of being heavily supervised.

Numerous works have addressed the automatic classification of spectra in the MK system (see, e.g., Bailer-Jones et al. 1998; Singh et al. 1998; Bailer-Jones 2001; Rodríguez et al. 2004; Giridhar et al. 2006; Manteiga et al. 2009; Daniel et al. 2011; Navarro et al. 2012). The primary approach followed in these works was to apply supervised learning training on labeled data. In this chapter, we focus on an unsupervised strategy that does not aim to reproduce the MK classification.

[1]This chapter is based on the published article Garcia-Dias et al. (2018).

Among all unsupervised classification methods, K-means (e.g., MacQueen et al. 1967, Everitt et al. 2011, Jain 2010) is a flexible clustering algorithm that has been extensively used in the literature. K-means was applied in several spectroscopic studies, including the identification of similar targets to average and reduce noise (Sánchez Almeida et al., 2009), the classification of one million galaxy spectra representative of the local universe (Sánchez Almeida et al., 2010), a systematic search for rare extremely metal-poor galaxies (Morales-Luis et al., 2011; Sánchez Almeida et al., 2016), or the classification of the large stellar spectra dataset available from the Sloan Digital Sky Survey, in particular data from the Sloan Extension for Galactic Understanding and Exploration (SEGUE; Sánchez Almeida & Allende Prieto 2013).

In this chapter we show the virtues and limitations of K-means in this context, making a first step in the search for alternatives to spectral classification. We examine whether or not the massive APOGEE dataset is amenable to a sensible unsupervised classification scheme based on K-means.

2.2 K-means clustering

We have introduced K-means in Sect. 1.4.3. Here we give more details about the implementation adopted in this chapter. The main difference with respect to the algorithm described in Sect. 1.4.3 is the choice of the initial centers, which can be done in different ways. The simplest approach is to randomly choose objects in the entire sample, but if the dataset has an over-abundance of a particular kind of object, the clusters will over-sample those objects. In order to avoid this, we initialize in an iterative fashion; we carry out a couple of K-means iterations with $K = 10$, with random initializations, then we randomly choose an object in the most abundant cluster as the initial center, discard all objects in this cluster and repeat the process until the desired number of initial cluster centers is reached. During the process, if more than 95 percent of the objects are discarded, we select the remaining cluster centers randomly in the whole sample. In this work we have translated the algorithm presented by Sánchez Almeida et al. (2010) from IDL[2] to Python [3]. Besides serial and parallel performance optimization, no major modifications were made. Using Python we achieved a simpler and faster code, which also has the advantage of being available in an open-source platform.

We have compared our results with those obtained using scipy[4] and scikit learn[5] algorithms. The results are qualitatively equivalent. The advantage of using our own code is that we are coherent with our previous works in the literature (Sánchez Almeida et al., 2009, 2010; Morales-Luis et al., 2011; Sánchez Almeida & Allende Prieto, 2013; Sánchez Almeida et al., 2016).

[2] http://www.harrisgeospatial.com/ProductsandSolutions/GeospatialProducts/
[3] www.python.org
[4] www.scipy.org
[5] http://scikit-learn.org

2.3 Spectroscopic data to be classified

As described in Sect. 1.3, for the spectral classification we use the spectra in the APOGEE DR12 dataset. The APOGEE spectra cover the near-IR, between 1.5 and 1.7 μm, in 8575 pixels with a resolving power $R \equiv \lambda/\delta\lambda \simeq 22,500$. For the vast majority of APOGEE observations, 35 fibers are devoted to observe warm stars, measuring telluric absorption, 35 fibers to observe the sky, pointing them to *blank* regions in the sky, and 230 fibers to acquire science spectra. Each star is affected by telluric absorption and sky emissions in different wavelengths depending on its radial velocity. The standard K-means algorithms are designed such that all input objects must have the same dimensions, and therefore we have to consider the same wavelengths in all spectra. To discard the pixels more affected by sky emission and telluric absorption, we have taken the average of the normalized sky and the telluric spectra for all fields in APOGEE DR12, and used them to identify all pixels for which the mean sky count is above 1 percent of the maximum mean normalized sky counts. We have also excluded all pixels at which the mean normalized telluric spectrum falls more than 5% below the continuum.

Figure 2.1 shows the mean sky and telluric spectra (blue and red lines, respectively), the cuts applied (black horizontal lines), and gray shaded areas highlighting the regions excluded from the spectra used in the K-means classification. In this figure we have displaced vertically the mean sky spectrum for clarity. Since stars have different heliocentric velocities, the spectra were corrected for Doppler shifts, and therefore they get affected by sky emission and telluric absorption at different wavelengths for different stars. This can be seen in Figure 2.1 from the width of the mean normalized telluric lines and sky emissions lines. From the 8575 original wavelength pixels, we kept 4838 pixels, or 56 percent of the APOGEE spectral coverage. All the spectra were also normalized using a 4th degree polynomial regression for each of the 3 chips in the APOGEE spectrograph. We have also removed values inthe normalized flux higher than 1.02 (i.e. two percent above the pseudo continuum level), avoiding any remaining problem with sky emission lines.

2.3.1 ASPCAP data with the stellar parameters

In order to analyze in physical terms the result of the classification based on observed spectra, we will employ the physical parameters that ASPCAP assigns to each individual star. We have used quality and target flags[6], and the uncalibrated parameters derived by ASPCAP[7] in order to evaluate the result of the K-means classification. Beside sky coordinates, atmospheric parameters ($T_{\rm eff}$, $\log g$ and $\xi_{\rm v}$), and [M/H], the dataset includes metallicities, α-element abundance, and individual chemical abundances for 15 elements[8]. As described in Holtzman et al. (2015), the

[6] More details about the flags are available in the documentations of TARGFLAGS, STARFLAGS, ANDFLAGS and ASPCAPFLAGS, at http://www.sdss.org/dr13/algorithms/bitmasks/

[7] These parameters are accessible through the objects FPARAM and FELEM, see allStar data model documentation website, https://data.sdss.org/datamodel.

[8] Al, C, Ca, Fe, K, Mg, Mn, N, Na, Ni, O, S, Si, Ti and V.

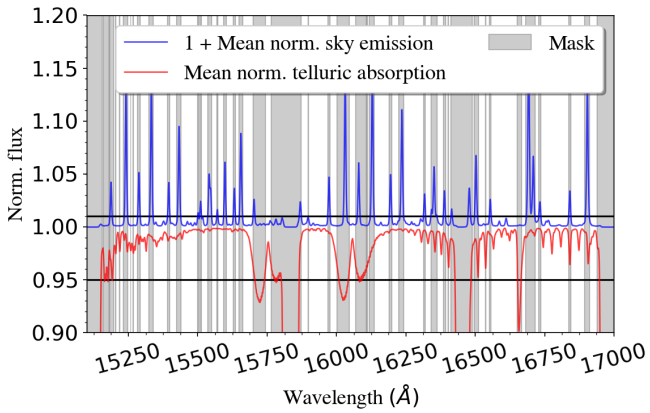

FIGURE 2.1— Mean sky normalized emissions (blue line) and telluric absorption (red line) spectra for the 153,847 spectra in the sample. Mean sky normalized emissions fluxes are displaced by one unit to help visualization. Black horizontal lines define the cut applied to each spectrum. Gray shades highlight the areas excluded from the K-means classification.

DR12 results were calibrated using star clusters in order to eliminate abundance trends with temperature and systematic differences with respect to the values in the literature. Since calibrated parameters are not available for all stars in DR12[9], we chose to use the uncalibrated parameters and chemical abundances. This choice should not affect the interpretation of our results; we are not interested in absolute values for each object, but in relative differences among spectra with intrinsically different shapes. In addition, using the uncalibrated data we can understand some bias and uncertainties of the ASPCAP. Given a set of stars with very similar spectrum, the dispersion of physical assigned parameters informs on the errors of the ASPCAP procedure.

2.4 Defining the number of clusters

Choosing the optimal number of clusters is a critical step in K-means classification. There is no universal criterion to do it, although many heuristic criteria have being developed over the last fifty years (Tibshirani et al. 2001 present some examples). In an attempt to select the best criteria for our problem, we built a testbed dataset with 6900 synthetic spectra spread over 69 well-defined clusters in surface gravity ($\log g$, in cgs units), $T_{\rm eff}$, α abundance ([α/M]) and [M/H], as shown in Figure 2.2. Metallicity is a measure of all the chemical elements heavier than

[9] Approximately 93% of the spectra in APOGEE DR12 have uncalibrated atmospheric parameters. The calibrated values are given for ≈ 63 percent of the spectra.

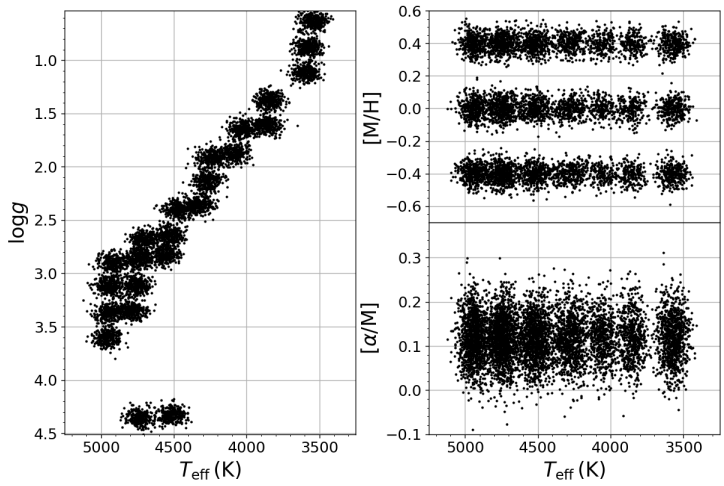

FIGURE 2.2— Atmospheric parameters for the synthetic dataset used to determine the optimum number of clusters. The left panel shows the effective temperature and the surface gravity for the synthetic spectra. The top right panel presents the projection of the clusters in the $T_{\rm eff}$ − [M/H] plane, while the bottom right panel shows the plane $T_{\rm eff}$ − [α/M].

He, assuming they vary in the same proportions with respect to the solar values. Analogously, [α/M] is a measure of all α-elements (O, Ne, Mg, Si, S, Ar, Ca and Ti) assuming they vary in union. The centers of the clusters were chosen based on the most dense regions in the distribution of $T_{\rm eff}$ and $\log g$ of the empirical dataset with uncalibrated parameters from DR12 for all available stars. The parameters for each spectrum were randomly chosen around each cluster center, following a normal distribution with a standard deviation of $\sigma_{T_{\rm eff}} = 50$K and $\sigma_{\log g} = \sigma_{\rm [M/H]} = \sigma_{\rm [\alpha/M]} = 0.05$. The synthetic spectra were built with FERRE[10], interpolating in a grid of theoretical models (Allende Prieto et al. 2004, 2006; Zamora et al. 2015). We use a grid with seven parameters per spectrum, microturbulence velocity ($\xi_{\rm v}$), carbon abundance ([C/M]), nitrogen abundance ([N/M]), [α/M], [M/H], $\log g$, and $T_{\rm eff}$. But the parameters $\xi_{\rm v}$, [C/M], [N/M] were fixed to the mean values of the stars in DR12 sample for all spectra[11]. In order to explore the best-case scenario we have not added any noise to the spectra.

We applied K-means on the simulated dataset 10 times, with K varying from 5 to 100. We then applied four different statistical criteria trying to recover the optimal number of clusters, knowing that the actual number is 69. We tried the KL index (Krzanowski & Lai, 1988), the gap statistic (Tibshirani et al., 2001), the CH index

[10] https://github.com/callendeprieto/ferre
[11] $\langle \xi_{\rm v} \rangle = 0.169\, km\, s^{-1}$, $\langle$[C/M]$\rangle = 0.122$, $\langle [N/M] \rangle = 0.227$

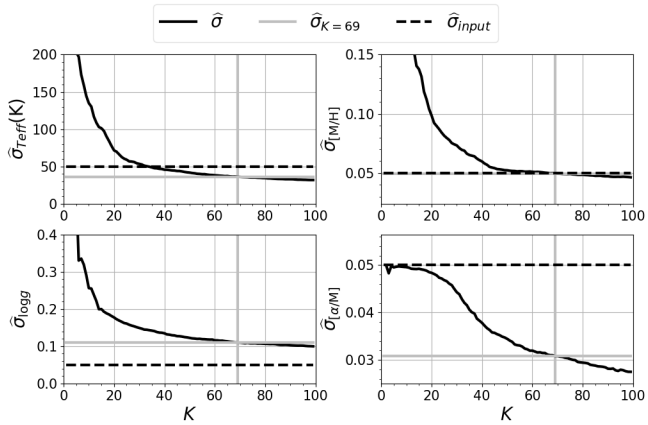

FIGURE 2.3— Variation of the median of the standard deviation as function of the number of clusters K in the synthetic dataset. The solid black lines represent the median of the standard deviation for all the classes in a run. The solid horizontal gray lines show the median of the standard deviation at $K = 69$, which is the true value. Dashed black lines show the input standard deviation. The top left panel refers to effective temperature $T_{\rm eff}$, the top right to metallicity [M/H], the bottom left to logarithmic surface gravity $\log g$, and the bottom right to α abundances [α/M].

(Caliński & Harabasz, 1974) and the silhouette index (Rousseeuw & Kaufman, 1990). These indexes were selected for being the most widely and successfully used in the literature. Their definitions are given in Appendix A.1. None of the chosen criteria was able to identify the right number of clusters, with the CH index being the only capable of giving consistent results over different initializations, finding $K = 9 \pm 1.8$, far from the *true* value 69. The other methods found a standard deviation in the number of clusters $\sigma_K > 12$ over the 10 different runs, while randomly selecting 10 numbers in this range would result in $\sigma_K \approx 25$. A possible explanation for this failure is that, despite the fact that the clusters are well defined in parameters space, the classification is made in flux space, where the separation between classes seems to be more subtle.

In the absence of better criteria, we have chosen the numbers of clusters based on the within-class standard deviation of the atmospheric parameters and chemical abundances. We use the notation $\widehat{X}$ meaning the median of X. In what follows, it is important to use medians instead of means in order to minimize the impact of the classes gathering a few faulty and unusual spectra. This is particularly important when the inferred criteria is applied to the observed dataset, with classes having few spectra (< 30) and a large dispersion in atmospheric parameters and chemical abundances. Figure 2.3 shows the variation of the median of the within-class standard deviation ($\widehat{\sigma}_X$) for each of the four main input parameters. We see a decrease

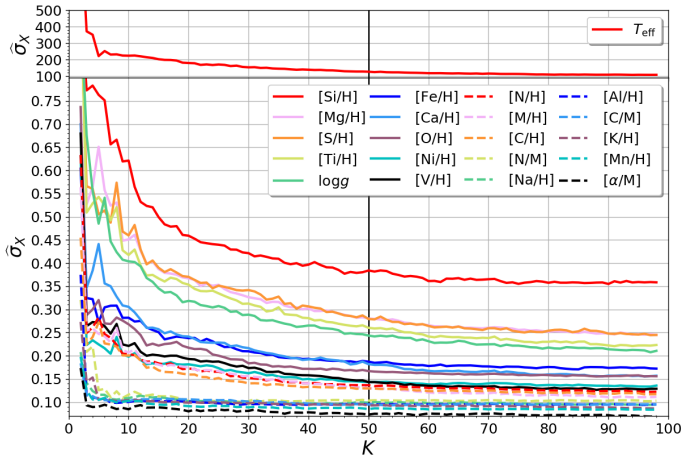

FIGURE 2.4— Variation of median standard deviation as function of the number of clusters K in the real dataset. Top panel refers to effective temperature $T_{\rm eff}$, while the bottom panel to the variation of median standard deviation for the other 20 parameters available in DR12 as indicated in the legend box.

in $\widehat{\sigma}_X$ as K grows for all quantities. This means that dividing the spectra in flux space into more classes results also in more fine partitions in atmospheric parameters/abundances space. Therefore, we can choose K based on a threshold value for $\widehat{\sigma}_X$. The extreme case would be to increase K until having one star per class, reaching the minimum variation. However, since the computational cost scales with K and we also lose generality when increasing K, we choose K in a trade-off among accuracy, agility and generality.

We know $\widehat{\sigma}_X$ and K for the synthetic dataset. Therefore we can verify how much we can trust the median standard deviation for the choice of K. Figure 2.3 shows that when $K = 69$ we have exactly the input metallicity dispersion, $\widehat{\sigma}_{\log(g)}$ is highly above the input value, while $\widehat{\sigma}_{\alpha}$ and $\widehat{\sigma}_{Teff}$ are both below the input value. The figure also shows how the slope of the curves (i.e., $|\partial\widehat{\sigma}_X/\partial K|$) decreases rapidly for $K \gtrsim 50$. Therefore, increasing K does not produce a significant change in $\widehat{\sigma}_{T_{\rm eff}}$ and $\widehat{\sigma}_{\log g}$ for $K \gtrsim 50$. We adopt this value as an optimum value for K in the simulation.

The actual APOGEE dataset behaves in a way similar to the simulation. Figure 2.4 shows how K affects the median of the standard deviation of $T_{\rm eff}$, $\log g$, [M/H], [C/M], [N/M], [α/M], and for the abundances of the chemical elements Al, Ca, C, Fe, K, Mg, Mn, Na, Ni, N, O, Si, S, Ti and V. From these plots we have chosen $K = 50$ as the number of clusters to be used throughout the study presented in this chapter, since increasing K does not reduce significantly the within-cluster dispersion.

2.4.1 Repeatability of the classification

The randomized nature of the initialization of K-means implies that different runs generate slightly different results. In order to evaluate the repeatability of the process we define a coincidence index ε, which measures the ratio of coincidence between two different classifications based on the number of spectra in equivalent classes, as described in Sánchez Almeida et al. (2010). We note that the label assigned to a class can vary over the classifications, even when the class remains with essentially the same objects. Therefore, when comparing two different classifications, we first need to cross identify the classes. For example, let S be a set of N objects, $S = \{\vec{x}_0, \vec{x}_1, ..., \vec{x}_N\}$ classified in K clusters, with two different initializations. Each initialization generates a set of clusters, say $\Omega = \{\omega_0, \omega_1, ..., \omega_K\}$ in one classification and $\Gamma = \{\gamma_0, \gamma_1, ..., \gamma_K\}$ in a second classification. In each classification, we order the classes ensuring that the number of objects in the ith class (n_i) is larger than the number of objects in the ith+1 class, i.e., $n_i \geq n_{i+1}$. TO compare the two classifications, we define a coincidence matrix $\mathbf{\Psi}_{K,K}$, with the elements $\psi_{i,j}$ being the number of objects in cluster ω_i that are also in cluster γ_j.

$$\psi_{i,j} = \sum_{\iota \epsilon \omega_i} \delta_\iota^j, \text{ where } \delta_\iota^j = \begin{cases} 1, & \text{if } \vec{x}_\iota \text{ is in cluster } \gamma_j \\ 0, & \text{if it is not.} \end{cases} \tag{2.1}$$

Thus, we match the jth cluster in Γ to the cluster in Ω having the maximum number of coincidences with it, $j_{match} = argmax\{\psi_{0,j}, \psi_{1,j}, ..., \psi_{i,j}\}$, always ensuring that no cluster in Ω is assigned to more than one cluster in Γ. Then we use the matches to transform the matrix $\mathbf{\Psi}$ into $\mathbf{\Psi'}$ permuting its columns to have their largest numbers in the diagonal. The elements of the diagonal of $\mathbf{\Psi'}$ ($\psi'_{i,j}$, with $i = j$) give the number of coincidences between the two classifications, while the off-diagonal elements ($\psi'_{i,j}$, with $i \neq j$) count the number of objects that differ in the two classifications. The trace of $\mathbf{\Psi'}$ divided by the total number of classified objects gives an estimate of the mean overall coincidence rate between the two classifications, $\bar{\varepsilon}_{total} = \mathrm{Tr}\{\mathbf{\Psi'}\}/N$. By defining the mean normalized coincidence matrix between a chosen classification ($\bar{\mathbf{\Psi}}'_{chosen}$) and a set of η classifications with the same K as $\bar{\mathbf{\Psi}}'_{chosen}$, the diagonal elements of the resulting matrix (Φ'_{chosen}) will give the mean coincidence ratio of each class over the η classifications, which is a measure of how stable the classes in the chosen classification are. Likewise, the elements out of the diagonal measure the mean confusion ratio between different classes. Here we present the evaluation of the repeatability for the synthetic dataset and the DR12 dataset, and in Sect. 2.4.2 we present the $\bar{\mathbf{\Psi}}'_{chosen}$ for the classification that we have chosen as reference.

Synthetic dataset repeatability

We performed a series of classifications for the synthetic dataset varying the number of clusters from $K = 5$ to 100. For each value of K we initialized the classification with ten different random seeds, the same ten seeds for all values of K. In order to avoid some possible bias caused by choosing a particular reference, the coincidence

ratio was measured for every pair of classifications having the same K. For the expected number of clusters in the synthetic dataset ($K = 69$) the mean coincidence ratio is $\bar{\varepsilon}_{total}(K = 69) = 74.7 \pm 6.2$ percent. The mean coincidence ratio computed for all runs with $K = 5$ to 100 for the synthetic dataset is 75.1 ± 8.4 percent.

DR12 dataset repeatability

Under equivalent conditions, that is to say, comparing all combinations of the ten classifications per value of K, with K from 5 to 100, the DR12 dataset had a mean coincidence ratio of $\bar{\varepsilon}_{total} = 77.9 \pm 7.8$ percent. When we consider $K = 50$, and 100 classifications with different random initialization, the mean coincidence ratio is found to be $\bar{\varepsilon}_{total}(K = 50) = 79.6 \pm 2.6$ percent. The classification used for reference in this estimate is described in Sect. 2.4.2.

To understand what a mean coincidence ratio of 79.6 percent means, we measured the mean difference between the paired classes over the 100 classifications, and compared this with the mean within-cluster dispersion of the chosen classification (see Appendix A.2 for more details). We found that the variations of the position of the class centroid over the 100 classifications amount to 6.4 ± 3.3 percent of the average mean within-class standard deviation of its corresponding class in the chosen classification. That is to say, even for runs where about 25 percent of the spectra do not agree on the class (coincidence of 75 percent), the main classes end up having their centers displaced by about 6 percent of the internal dispersion of its class in the 4838-dimensional classification space. As we show in Sect. 2.4.2, the confusion occurs mainly between classes sharing borders in the space $T_{\rm eff} - \log g -$ [M/H]. Except for some outlier classes, the shapes of the classes are very similar over different classifications.

2.4.2 Choosing the best classification

After running K-means a hundred times with $K = 50$, we select one of them to be used to analyze the result of the classification procedure in astrophysical terms. We chose the classification with the lowest sum of squared error (SSE). As we are working with the Euclidean metric, the SSE is computed as

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\iota \epsilon \omega_i} ||\vec{x_\iota} - \vec{\mu_i}||^2, \text{ where } \vec{\mu_i} = \frac{1}{n_i} \sum_{\iota \epsilon \omega_i} \vec{x}_\iota, \tag{2.2}$$

x_ι is the ιth spectrum in cluster ω_i and $\vec{\mu_i}$ the centroid of the class i. The chosen classification has an SSE 9 percent smaller than the average SSE over all classifications. As mentioned in 2.4.1, the coincidence ratio is measured by the number of spectra sharing the same class over two distinct classifications. Comparing the chosen classification with the other 99 runs, the average coincidence ratio is 79.6 ± 2.6 percent, which can be considered a high repeatability rate. Also the mean variation of the centers of the most popular classes, containing 99 percent of the objects, is

≈ 2.4 percent of the mean within-cluster variation of the classes in the chosen classification. Again, this is a comparison between the standard deviation of the centroids over the 100 classification with the internal standard deviation of the main classes in the chosen classification. In this case the number falls from 6.4 to 2.4% taking into account the classes containing 99% of the spectra in the sample, which in the case of the chosen classification corresponds to classes from 0 to 31.

Figure 2.5 shows $\mathbf{\Psi'}_{chosen}$, comparing the chosen classification with the other 99 classifications. The elements of $\mathbf{\Psi'}_{chosen}$ are represented by a color scale in a 2D histogram; the bottom panel in this figure shows a histogram with the main diagonal of $\mathbf{\Psi'}_{chosen}$. This plot will be useful in Sect. 2.5, where we will describe each group of classes and comment on the stability of each class. From now on, we will refer to the elements in the main diagonal of $\mathbf{\Psi_{chosen}}'$ as *coincidence rates* and the off-diagonal as the *confusion rates*.

In Figure 2.6 we compare the root mean squared distance of each spectrum from its best fit spectrum with the distance to the centroid of its class. The plot shows how the centroids are five times closer to the class centroid than to its best fit. The large distance between the spectra and the models is due to systematic differences between the synthetic spectra based on model atmospheres and real spectra.[12]

Table 2.1 shows a comparison between the standard deviation within clusters ($\widehat{\sigma}$) and the overall standard deviation ($\sigma_{overall}$), the standard deviation for all stars in the dataset. For example, $T_{\rm eff}$ and $\log g$ have a $\widehat{\sigma}$ about 3.6 and 4.2 times smaller than their corresponding $\sigma_{overall}$, respectively. This means that the algorithm is especially sensitive to $T_{\rm eff}$ and $\log g$. In Table 2.1 we also highlight the parameters that present $\widehat{\sigma}$ at least two times smaller than its $\sigma_{overall}$. They are $T_{\rm eff}$, $\log g$, [M/H], [Ca/H], [C/H], [Mg/H], [N/H], [Si/H], [S/H] and [Ti/H]. Since these are the most sensitive parameters to K-means, we will focus mainly on them when interpreting the classes in the appendix A.3.

2.5 Results

After visual inspection of the classes main spectra we divided all the classes into nine groups sharing similar properties. Here we briefly describe the results and present the main characteristics of the classes. Table 2.2 summarizes the main properties of the classes and provide hyperlinks for the description of the classes in the appendix A.3 and also for online supplementary material for the visualization of the classes. Figure 2.7 presents contour plots in the $T_{\rm eff}$ − [M/H] plane for the 32 most populated classes. We highlight regions enclosing 15, 30, 45, and 68.3 percent of the stars in each class, with the color shades varying from strong to light, respectively. Class 21 is too concentrated to have clear contours at this scale, so it is represented by purple dots in the figure. In some cases the separation of the contours is too tight and only three contours are visible. The figure is divided into

[12] This can also be seen in panel F of the summarized plots in the appendix A.3. For instance, https://garciadias.github.io/APOGEE/group0/class2/index.html present these systematic differences near 16205 Å and 16215 Å in class 02.

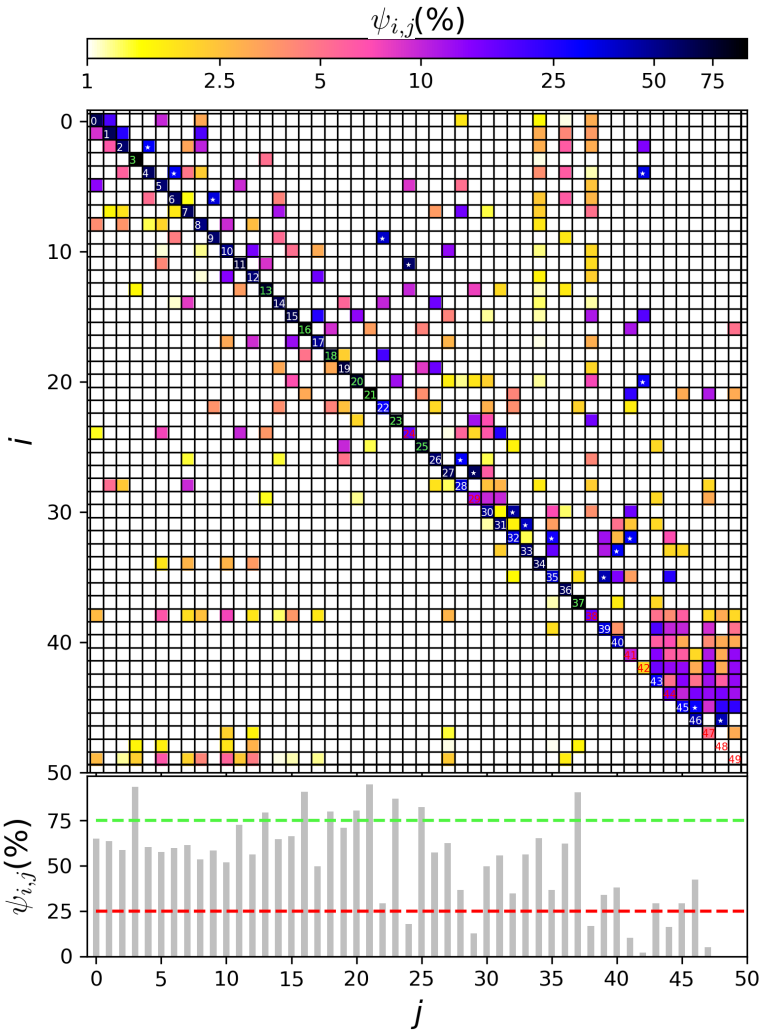

FIGURE 2.5— Top panel shows the mean coincidence matrix comparing the chosen classification with the other 99 classifications. The value of the different cells is color-coded according to the color bar given on top of the figure. Cells on the mean diagonal represent the coincidence ratio of a class and quantify the stability of the class upon the random initialization of the classification procedure. The cells in the diagonal are labeled with their corresponding class number and highlighted in green if the class has a coincidence ratio above 75 percent or in red if class has coincidence ratio below 25 percent. Off-diagonal cells can be interpreted as the confusion rate between two classes. We highlight confusion rates above 25 percent with white stars. The bottom panel presents a histogram of the coincidence ratios corresponding to the diagonal of the coincidence matrix. A green dashed line marks the 75 percent level, while a red line marks the 25 percent level.

TABLE 2.1— Comparison of the within cluster median standard deviation (third column) with the overall standard deviation (second column) for each parameter. The fourth column lists the ratio of these quantities. We highlight the parameters that have within cluster median standard deviation at least two times smaller than the overall standard deviation.

Parameter	$\sigma_{overall}$	$\widehat{\sigma}^{K=50}$	$\sigma_{overall}/\widehat{\sigma}^{K=50}$
$T_{\rm eff}$ (K)	553	152	**3.6**
$\log g$	1.17	0.28	**4.2**
[M/H]	0.35	0.17	**2.1**
[C/M]	0.12	0.11	1.1
[N/M]	0.18	0.12	1.5
[α/M]	0.10	0.08	1.3
[Al/H]	0.13	0.10	1.3
[Ca/H]	0.48	0.22	**2.2**
[C/H]	0.31	0.15	**2.1**
[Fe/H]	0.38	0.23	1.7
[K/H]	0.12	0.10	1.2
[Mg/H]	0.75	0.35	**2.1**
[Mn/H]	0.15	0.09	1.6
[Na/H]	0.15	0.10	1.5
[Ni/H]	0.29	0.18	1.6
[N/H]	0.32	0.16	**2.0**
[O/H]	0.39	0.21	1.9
[Si/H]	1.00	0.43	**2.3**
[S/H]	0.77	0.35	**2.2**
[Ti/H]	0.71	0.33	**2.1**
[V/H]	0.36	0.19	1.9

TABLE 2.2— Summary of the classes and complementary material.

Group	Class[a]	Stellar type[b]	Gal. component[c]	Comment
Metal-rich RC and RGB				
0	Class 02	K-Giants	Thin disk	Lowest [M/H] in the group, 31% RC.
0	Class 04	K-Giants	Thin disk	26% RC
0	Class 06	K-Giants	Thin disk	26% RC
0	Class 08	Sub Giants, K-Giants	Thin disk	Warmest in the group, 1% RC.
0	Class 09	K-Giants	Thin disk	[M/H] near to grid limits, 21% RC.
Metal-poor cool RGB				
1	Class 07	K-Giants	Disk	Thick disk.
1	Class 14	K-Giants	Disk	—
1	Class 19	K/M-Dwarfs	Disk	$T_{\rm eff}$ near to the grid limits.
1	Class 25	M-Giants	Disk	$T_{\rm eff}$ near to the grid limits.
1	Class 26	K-Giants	Disk	High $[\alpha/M]$ blob.
1	Class 28	K-Giants	Bulge/center	Most metal-poor stars.
Warm stars				
2	Class 03	Blue stars	Disk	Warmest telluric standards
2	Class 11	F/G-Dwarfs	High g. latitude	Warm, telluric standards.
2	Class 13	Blue stars	—	Warm fast rotation stars. Telluric standards.
Fast rotators				
3	Class 27	K/M-Dwarfs	—	Fast rotators.
3	Class 29	M-Dwarfs	—	Fast rotators.
Metal-rich cool RGB				
4	Class 16	K/M-Giants	Disk	$T_{\rm eff}$ near to the grid limits.
4	Class 18	K-Giants	Disk	—
4	Class 22	K-Giants	Thin disk	[M/H] near to the grid limits.
Metal-poor RC and RGB				
5	Class 00	K-Giants	Disk	Broad in atmospheric parameters.
5	Class 01	K-Giants	Disk	Whole RGB
5	Class 05	Sub Giants, G/K-Giants	Disk	Broad in atmospheric parameters.
Dwarf stars				
6	Class 10	G/K-Dwarfs	Thin disk	—
6	Class 12	K-Dwarfs	Thin disk	—
6	Class 15	K-Dwarfs	High g. latitude	—
6	Class 17	K-Dwarfs	Thin disk	—
6	Class 20	M-Dwarfs	High g. latitude	Atmospheric parameter near to the grid limits.
Sparse classes				
7	Class 21	M-Giants	Bulge/center/Disk	Atmospheric parameter near to the grid limits.
7	Class 23	M-Dwarfs	—	Atmospheric parameter near to the grid limits.
7	Class 24	Giants	Halo	High $[\alpha/M]$, metal-poor stars.
7	Class 30	—	—	Poor fit, M31 clusters, high g. latitude.
7	Class 31	Giants	High g. latitude	metal-poor high [α/M].

[a] Hyper-links to figures as described in Appendix A.3

[b] The stellar types here are inferred simply from the distribution of $T_{\rm eff}$ in the classes.

[c] Based on the mean distribution of the class on the galactic plane and on the [α/M]-[M/H] plane.

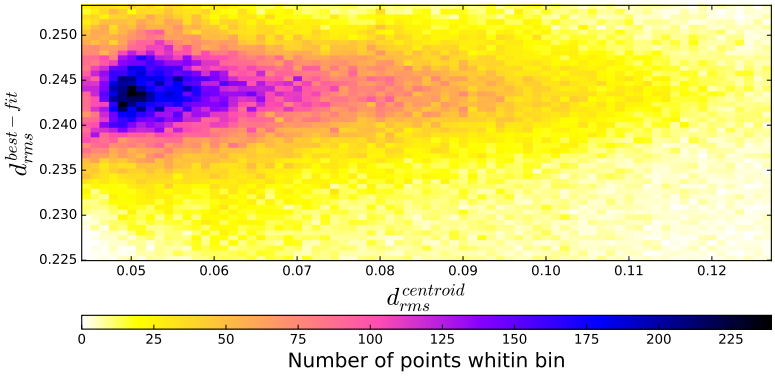

FIGURE 2.6— 2D histogram comparing distances of each spectra to its best fit with the distance of each spectra to its class centroid. Each pixel in the image is color coded according to the number of spectra in that region, as indicated by the color bar.

three panels, trying to minimize the superposition of classes. We use different colors to help identifying borders between classes. Some classes have the same color, but there is no overlap between classes with the same color. Classes are identified with labels in Figure 2.7. For the labels we use the abbreviations G for group and C for its associated classes. Classes in group 8 have few objects, which are sparsely distributed in the $T_{\rm eff} - [{\rm M/H}]$ plane, making this its contour plot very noisy and hard to read; we present a scatter plot for group 8 in Figure 2.8. Figure 2.9 shows the main distribution of the groups in the $T_{\rm eff} - \log g$ plane. Besides the differences found in the $T_{\rm eff} - [{\rm M/H}]$ space, we also found some other particularities in the classes and groups, some of them based in the spatial distribution (RA - DEC), global chemical abundances, or spectral fluxes.

Figure 2.10 presents the mean spectra, in a narrow spectral window, for all classes in groups 0 to 7. Each panel in this figure shows the mean spectrum of the classes in each group color-coded as in Figure 2.7. In order to offer the highest contrast between the mean spectra of the classes, we chose the spectral coverage which maximizes the cumulative variance over the first 32 classes in a 150-pixels-long window. The gray shades in the background of these plots highlight the masked pixels (those discarded from the classification, as discussed in Sect. 2.2).

Besides the description presented in this Section, we include complementary material with detailed plots for many of the DR12 available features in the supplementary online material described in the Appendix A.3, and linked in Table 2.2. This table gives a short description for each class and provides links to the online figures, which are described in A.3. With these figures the reader can find more details about the atmospheric parameters, spatial distributions, and chemical abun-

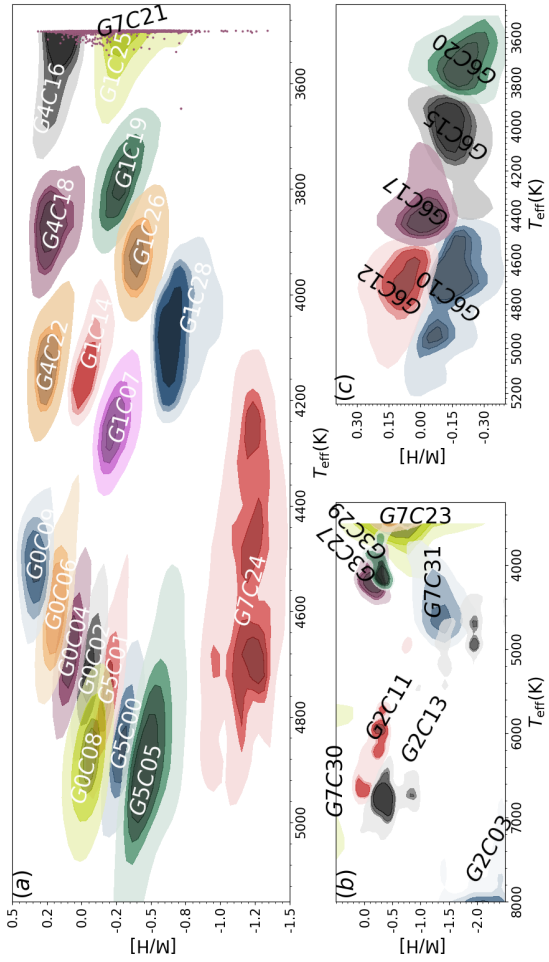

FIGURE 2.7— Contour diagrams in the $T_{\rm eff}-$[M/H] plane. Different colors are used to distinguish different classes. Each class is represented by four color shades, from dark to light, the shades enclose 15, 30, 45 and 68.3 percent of the data points in the class. The groups are separated in three panels minimizing the superposition of classes. Panel (*a*) shows groups 0, 1, 4, 5 and two classes of group 7. Panel (*b*) groups 2, 3 and three classes of group 7. Finally, panel panel (*c*) shows group 6. In these panels each class is flagged with a floating label in the form GxCxx, C referring to class and G to its group. Class 21 is represented as a scatter plot, since it is too concentrated to present visible contours on this scale.

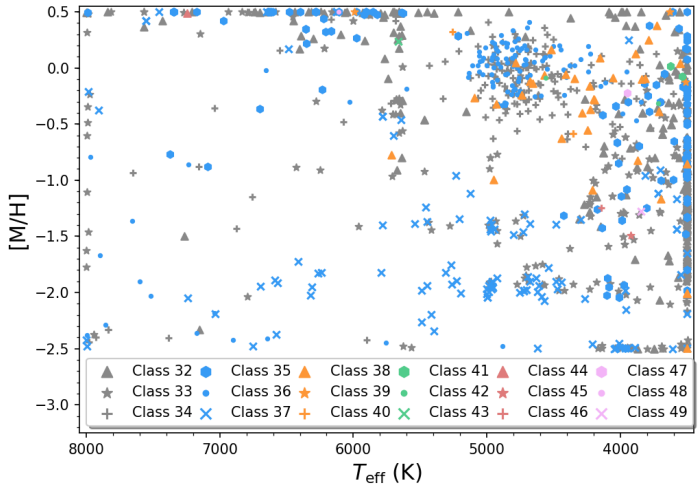

FIGURE 2.8— Scatter plot of $T_{\rm eff}$ against [M/H] for the classes in group 8. The classes are identified as shown in the legend. The stars in this group are scattered throughout the plane.

dances for each class presented. The complete description of the online material is found in appendix A.3. Table A.1 presents the classification of each spectra and Tables A.2 and A.3 shows mean and standard deviation for each pixel in the spectra of each class. The Tables A.4, A.5, A.6 and A.7 present the median values for the atmospheric parameters and all the individual chemical elements in each class. The upper and lower limits presented in the tables, as well as those shown in this chapter and its appendixes, were calculated by taking the interval around the median, which encloses 68.3 percent of the points in each class.

2.6 Summary and conclusion

2.6.1 Main results

In this chapter we have used K-means to perform an automated unsupervised classification of all the 153,847 APOGEE spectra included in DR12. We classified the spectra into 50 classes, which were afterwards sorted manually into nine major groups. By construction, each class collects spectra that are very similar. The resulting classes and groups are interpreted using the physical parameters inferred by the APOGEE Stellar Parameters and Chemical Abundances Pipeline (ASPCAP). We found that classes were divided mainly according to their $T_{\rm eff}$, $\log g$ and [M/H], and less strongly by other characteristics, such as elemental abundances or the quality of the spectra. Groups from 0 to 7 include 32 classes containing 99.3 percent of

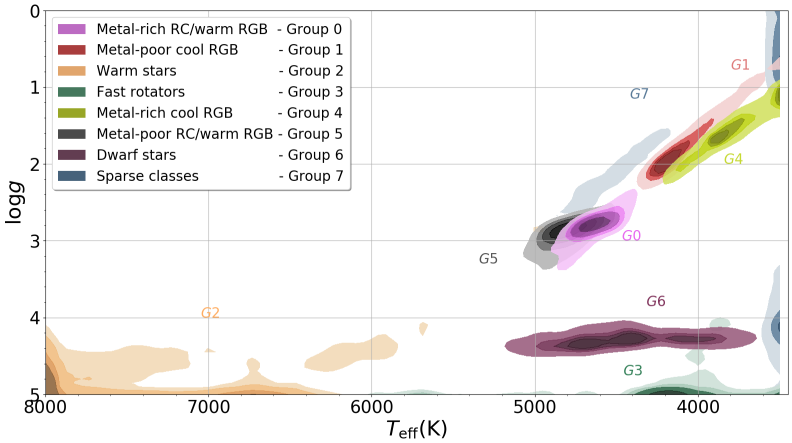

FIGURE 2.9— Contour diagram for the groups in the $T_{\rm eff} - \log g$ plane. Each group is represented by a different color. color shades enclose, from dark to light, 15, 30, 45 and 68.3 percent of the objects in each group. In shades of pink we show group 0, which contains metal-rich stars in the RC and in the war part of the RGB. In shades of red we show group 1 with metal-poor stars in the RGB. Orange contours present the distribution of group 3, which is mainly formed by warm stars. Group 3 of fast rotator stars is presented in shades of green. Group 4, made up of stars in the metal-rich cold tip of the RGB is presented in shades of yellow. The gray shades show the distribution of metal poor star in the RC and in the warm end of the RGB that form group 5. Dwarf stars are in group 6, represented by shades of purple. Shades of blue represent the group 7, with the most sparse classes.

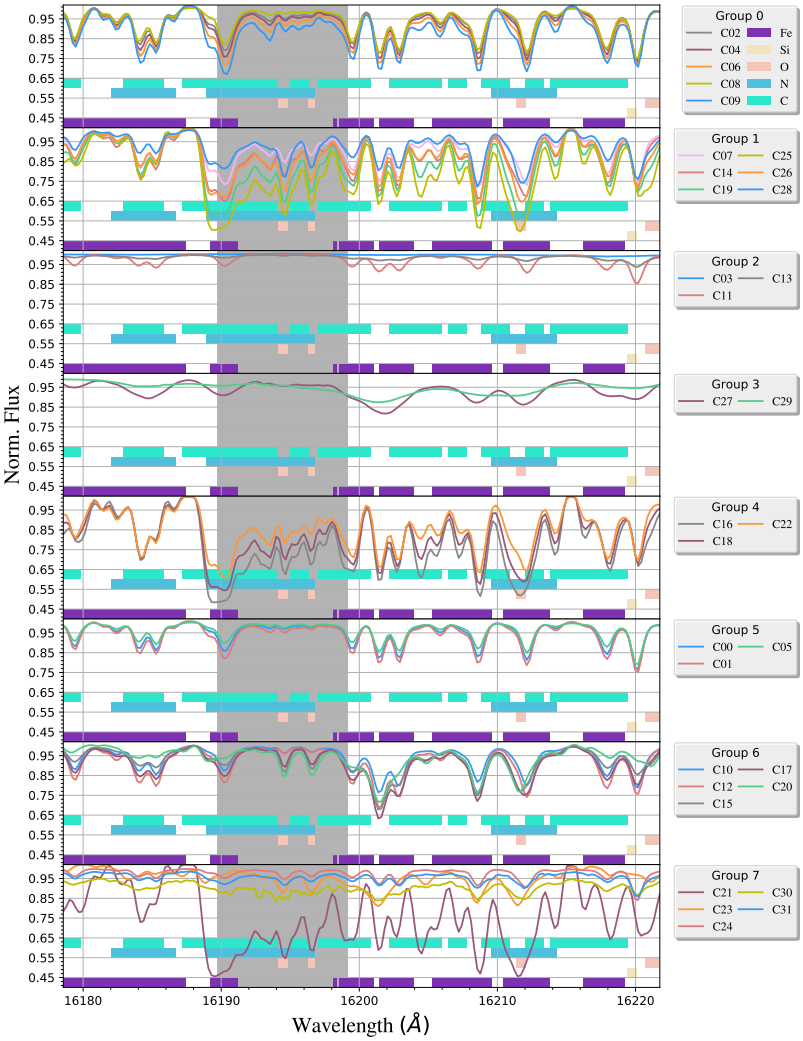

FIGURE 2.10— Mean spectra of the classes in the wavelength range from 16178 to 16222 Å, where the differences among classes are particularly enhanced. Top to bottom, panels show the mean spectra for classes belonging to groups from 0 to 7. Each mean spectrum is drawn with the same colors used in Figure 2.7. In all panels, the shaded gray region indicates the spectral windows masked in ASPCAP to determine the chemical abundances of stars. Each set of windows used to determine the abundances of each element, as described in 1.3, is presented as horizontal bars color coded as indicated in the legend of the first panel.

the spectra in DR12. The identified groups can be described as follows:

- **Group 0:** Includes five classes dominated by RC stars and the warmest end of RGB sequence with different chemical abundances.
- **Group 1:** Composed of six classes with stars from the RGB, cooler than those in group 0, and mainly separated from each other by their chemical abundances.
- **Group 2:** Made up of three classes mainly populated by warm dwarfs, warm subgiant stars, and some A- and B-type stars used for telluric correction.
- **Group 3:** Composed of two classes with fast rotating stars. Due to the strong line broadening, they are among the most poorly-fitted spectra in the survey.
- **Group 4:** Has two classes covering almost the same range of $T_{\rm eff}$ and $\log g$ as group 1. They are RGB stars, but with higher metallicities.
- **Group 5:** Contains three classes formed by stars from the RC and the warm end of RGB, with stellar populations from both the thin and thick disk.
- **Group 6:** Formed of five classes composed of dwarf stars with a wide range of temperatures.
- **Group 7:** Including five classes with peculiar stars.
- **Group 8:** Collects 18 classes with all the outliers of the classification, less than 1 percent of the spectra in SDSS DR12.

2.6.2 Uses of the classification

As with any classification, this work can be used to provide an overview of the APOGEE DR12 dataset, which simplifies the visualization and highlights some features of the survey. For example, we can easily see that class 3, composed of very warm stars with almost featureless spectra, has an unexpectedly well-behaved distribution of values for [C/M], [N/M], [α/M], [Mn/H] and [Na/H]. It also easily identifies strange behaviors such as the bimodality in [K/H] for class 15, the gaps in metallicity found in class 11, and the similarity in parameters of stars with very different spectra, as is the case for classes 20 and 27.

We provide extensive additional online material in order to encourage the search for features that may be interesting for specific purposes. For example, the catalog provides a set of standard spectral templates that could be applied in stellar populations synthesis for galaxies. The mean spectrum (centroids) of the classes are arguably more reliable templates than the traditional synthetic models of standard MK type stars. However, the application should be restricted to those classes with a high number of members and low internal dispersion. Moreover, calibration of the

atmospheric parameters and abundances is required, since the ones presented here are based on uncalibrated parameters.

The centroids of the classes are also useful to find substantial differences between the spectra and their best fit model found by ASPCAP. Since the classes are a collection of very similar spectra, the comparison between the class' mean and the mean of their best fit model can underline systematic differences between spectra and models.

Some classes have a different spatial distribution without an obvious reason, for example, classes in group 2 differ in their spatial distribution, something unexpected since the main difference among them is the $T_{\rm eff}$. Class 31 has an especially peculiar distribution, occupying mainly the region with $60^{\rm o} \leq l \leq 90^{\rm o}$ and $0^{\rm o} \leq b \leq 45^{\rm o}$. The reason is unclear yet. Further investigations must be carried out to find out the cause of this spatial segregations. Other spatial distributions are less surprising, for example, classes in group 4 are concentrated in the disk. This is to be expected, since their metallicity and [α/M] distributions match those expected for red giants that are part of the thin disk population. Classes 24 and 28, formed by metal-poor stars with high α-element abundances, corresponding to the halo population, are expected to be out of the galactic disk, as we found. Class 21 can be interpreted as the population of the bulge, with high α-element abundances and high metallicity, and is also expected to have a preferential spatial distribution like the one observed. These are the most evident examples of spatial segregation, but others can be found among the classes.

2.6.3 Additional issues

In this chapter we face the problem of determining the optimal number of clusters for the K-means classification. In our case, none of the standard criteria seem to provide a reliable answer. That is probably a consequence of the continuous nature of the dataset. In general, there are no sharp changes in the spectral properties of the stars. Indexes like CH and KL are mathematically proven to work in datasets with well separated clusters, but perform poorly in overlapping clusters or continuous distributions. In this case, K-means provides a way of artificially dividing a continuous space into meaningful slices, maximizing the similarity among objects in the same class. Thus, the number of classes can be tuned according to the degree of within-class compactness we are interested in, as shown in Sect. 2.4.

Another consequence of applying K-means to a continuous dataset is a significant observed degree of confusion between classes sharing borders in the space $T_{\rm eff} - \log g -$ [M/H]. However, these issues are not restricted to K-means. Any analysis tool, independently of whether it is supervised or not, will face the intrinsic degeneracy of these quantities in the stellar spectra. Soft clustering algorithms such as fuzzy K-means or density based algorithms such as Gaussian mixture models or DBSCAN could provide a more natural way to deal with this kind of problem, but would not solve the overlap of the classes in the space of parameters.

We have shown how the random seed used by the algorithm affects the resulting

classes. Although there is no unique solution, the variations are negligible compared to the internal dispersion of the classes. In addition, we show how the centroids of the classes are much closer to the spectra in the class than their corresponding best fit models. This suggests that K-means can be used to identify the systematic deficiencies of the modeling adopted in the determination of physical parameters and abundances with ASPCAP, and improve the agreement with the data.

Although the within-class dispersions in the parameter space are larger than the typical uncertainties derived from this kind of data, K-means provides good insight into the general characteristics of the spectra in the dataset. In this sense, K-means is not the optimal algorithm to be used for parameter determination, but can be useful in an early analysis of the data, helping to design solutions and map the general behavior of the dataset.

K-means essentially performs hyperspherical cuts in the N-dimensional space. Future works in unsupervised spectral classification should address the issues presented in this Section and search for algorithms that can more generically divide the space taking into account its density distribution. Also a soft clustering approach can arguably produce a more reliable classification. However, more complex algorithms are also more computationally expensive, therefore any further application has to address the scalability problem.

2.6.4 Conclusions

As exemplified in this chapter, K-means provides an easy way to divide complex problems into smaller pieces, which are simpler to solve. The version of ASPCAP used in DR12 was designed to work optimally on K and early-M giant stars. For dwarfs, warmer ($T_{\rm eff} > 6000$ K), cooler ($T_{\rm eff} < 3800$ K), or metal-poor stars ([M/H] < -1), the results are less accurate. Prior to a model-atmospheres spectral analysis, K-means can provide guidance on the most natural groups in the dataset. This can be very useful to design a pipeline that treats differently the distinct groups of objects, which is necessary for groups such as 2, 6, 7, and 8, for example.

Wolpert & Macready (1997) put forward what is known as the 'no free lunch' theorem for machine learning. That is to say, there is no best machine learning algorithm; it is always a matter of which one is better suited to the specific features of a given problem. Knowing the problem, we can only presume which kind of algorithm is most suitable for solving it, but finding the best solution always requires testing some algorithms and tuning their parameters. This work adds to previous applications of K-means (Sánchez Almeida et al., 2009, 2010; Morales-Luis et al., 2011; Sánchez Almeida & Allende Prieto, 2013; Sánchez Almeida et al., 2016) consolidating a guideline for the use of this algorithm in the analysis of spectroscopic data, and providing a new perspective for the APOGEE data.

In this chapter we made a serious effort to organize the spectra into classes and groups according to the similarity within their spectra. This classification is completely independent of any atmospheric and spectroscopic model. It provides a useful way to explore the data in APOGEE, since it allows a quick identification of

the main different types of objects in the survey.

Finally, the extensive online material we provide can be used to explore deeper aspects of DR12 APOGEE. We encourage the use of Tables 2.2 and A.1 for any interested researcher to explore the results of the classification. The Classification is available at http://vizier.cfa.harvard.edu/viz-bin/VizieR?-source=J/A+A/612/A98.

3

Testing the limits of chemical Tagging with star clusters

3.1 Introduction

In spectroscopic surveys such as APOGEE, there are two main approaches to the problem of finding classes of similar objects. It is possible to directly use the stellar spectra to find patterns or classes in the stellar populations or, instead, to work with the atmospheric parameters and chemical abundances. In Chapter 2 we studied how K-means performs in the direct classification of stellar spectra. We demonstrated that the spectral classification is most sensitive to $T_{\rm eff}$ and $\log g$. In this chapter, we move on to the analysis of the chemical abundances, which allows a complementary analysis of the dataset.

The analysis of chemical abundances relies on model atmospheres and line formation theory, but enables a finer description of the stellar populations. While the spectral classification mainly reflects physical parameters of the stars, i.e., effective temperature and surface gravity, pattern recognition in the chemical abundance space can potentially uncover the star formation history of our Galaxy through the identification of chemical patterns characteristic of stars with a common origin, usually referred to as *chemical tagging* (Freeman & Bland-Hawthorn, 2002).

The classification of stars based on chemical abundances has been recently explored by Blanco-Cuaresma et al. (2015) using homogeneous data for 339 stars in 35 open clusters. Applying machine learning techniques, Hogg et al. (2016) were able to identify known star clusters and the Sagittarius dwarf galaxy in the APOGEE DR12 dataset. Schiavon et al. (2017) identified a stellar population unusually rich in nitrogen in the central part of the Galaxy. Kos et al. (2018) used GALactic Archaeology with HERMES (GALAH) data to identify new members of the Pleiades cluster. The application of machine learning algorithms to the stellar abundances has also been used in works by da Silva et al. (2012); Ting et al. (2012); Jofré et al.

(2017); Anders et al. (2018) and Boesso & Rocha-Pinto (2018).

This chapter aims at exploring the limits of a large battery of unsupervised clustering algorithms in distinguishing mono-abundance stellar populations. Star clusters are likely the most chemically homogeneous populations in the Galaxy, thus any practical approach to chemical tagging should at least be able to separate known clusters from each other. Section 2 describes the details of the APOGEE data used in this work. The chapter is self contained so the information on APOGEE needed to follow the argument, partly mentioned in Chapter 2, is also repeated here. Section 3 presents the star cluster samples. Section 4 discusses the feasibility of the automated identification of stellar clusters. Section 5 presents the algorithms tested, and Section 6 describes the results. Finally, Section 7 summarizes the results and discusses the viability of applying these algorithms to blindly identifying families of stellar populations in the whole APOGEE dataset.

3.2 Data

APOGEE is observing spectroscopically hundreds of thousands of stars, primarily giants, emphasizing the regions of the Galaxy obscured by dust (the Galactic plane and the bulge) creating a chemical map (Majewski et al., 2017). In this Chapter we use data from Data Release 14 (Abolfathi et al., 2018), which includes observations for more than 260,000 stars.

With a signal-to-noise ratio per half a resolution element typically larger than 100, the $H-$band APOGEE spectra are routinely analyzed by ASPCAP (García Pérez et al., 2016). The current version of ASPCAP provides determinations of chemical abundances for about 20 elements with a precision that ranges from 0.01 dex for iron, oxygen or magnesium in solar-metallicity K-giants, to $\sim$ 0.2 dex for nitrogen in moderately metal-poor stars (Holtzman et al., 2015; Bertran de Lis et al., 2016). Among these elements, we use those that have been shown to be reliable in previous studies (Holtzman et al. 2015; Holtzman et al. 2018 in preparation; Jönsson et al. 2018, in preparation). The elements used in the clustering applications are C, N, O, Na, Mg, Al, Si, P, S, K, Ca, Fe, and Ni.

APOGEE includes observations of globular clusters and open clusters targeted for science and calibration purposes (see, e.g., Frinchaboy et al. 2013). Most clusters have a fairly limited extent on the sky and pose a challenge for the 70-arcsecond fiber collision radius of APOGEE. Hence only a limited number of members are typically observed per cluster. Only a few of the clusters targeted by APOGEE have tens of observed stars (e.g., NGC 2420, NGC 6791, the Pleiades or M67; see, e.g. Cunha et al. 2015, Linden et al. 2017; Souto et al. 2016, 2018). About two tens of clusters include 5–30 stars, and those are the ones we focus on in this chapter. APOGEE provides the abundances determined directly from the fitting of the observations with synthetic spectra based on standard model atmospheres. The optimization of the parameters is based on a χ^2 criterion, and performed using the FERRE code (Allende Prieto et al. 2006; see also Sect. 1.3 in Chapter 2).

The inferred abundances show mild trends with the stellar effective temperature

for stars members of open clusters (Holtzman et al., 2015). Even though diffusion in stellar envelopes can systematically alter photospheric abundances in a manner that depends on stellar mass (see, e.g., Souto et al. 2018), and therefore on effective temperature, such trends are interpreted mainly as a result of shortcomings in the model atmospheres and radiative transfer calculations used to derive the abundances. For example, departures from LTE or 3D effects and uncertainty with the atomic parameters. In an attempt to remove such effects, the trends with effective temperature are modeled with smooth functions of $T_{\rm eff}$, taking into account all the open clusters, producing *internally calibrated* abundances, which are the default ones included in the data releases.

Since we are interested in relative abundances among stars with similar atmospheric parameters and do not care much about absolute values, we have preferred to use the *uncalibrated* abundances. The uncalibrated abundances are provided together with the calibrated ones for each data release[1]. Uncalibrated abundances are available for far more stars than the calibrated ones, since the calibrations span a limited range of the parameters space. Dwarf stars, in particular, only have uncalibrated values, and they amount to some 20% of the total APOGEE sample. In addition, the smooth functions used to correct the trends with effective temperature may lose some of the precision of the original uncalibrated values for stars with similar parameters.

3.3 Sample

We have selected stellar clusters in APOGEE DR14 with at least five members. Membership was determined based on radial velocities and the distribution of chemical abundances. We selected the stars compatible with the mean radial velocities and velocity dispersions in Dias et al. (2002) for open clusters, and Francis & Anderson (2014) for globular clusters. We then applied one iteration of two-sigma clipping in all the chemical abundances to guarantee a single composition. Table 3.1 shows the list of selected clusters. An online table lists all the stars in the clusters[2]. Figure 3.1 shows the distribution of $T_{\rm eff}$ and $\log g$ for the whole sample.

3.4 Cluster distinguishability through their chemical abundances

In order to check whether the clusters have intrinsically different distributions of abundances, we have performed two statistical tests. The one-dimensional Kolmogorov-Smirnov two-sample test (K-S test; Smirnov 1939; Darling 1957) was applied for each pair of clusters and each chemical element, and the Cramer multivariate test (Baringhaus & Franz, 2004; Elias et al., 2006; Yeremi et al., 2014) for each pair of clusters, and for all chemical elements at the same time. In both tests, the null hypothesis is that the two samples can be generated from the same distribution. If

[1] See the data model and specifically the description of the FELEM array in http://www.sdss.org/dr14/irspec/abundances/

[2] `link_will_be_add_to_the_final_version`

TABLE 3.1— Clusters used in the chemical abundance classification. $N_\star$ is the number of cluster members, the third column presents the logarithm of the cluster ages, the fourth column gives their mean uncalibrated iron abundance, the fifth column shows the mean of the calibrated iron abundances, and the last column presents the difference between calibrated and uncalibrated iron abundances. The table's footnote gives the reference to the age of each cluster.

Cluster	$N_\star$	log(Age)	$\langle$[Fe/H]$\rangle_{\rm uncalib}$	$\langle$[Fe/H]$\rangle_{\rm calib}$	$\Delta_{calibration}$
Globular Clusters					
M2	14	11.78[i]	-1.43	-1.52	-0.09
M3	44	11.39[i]	-1.38	-1.50	-0.12
M5	74	10.62[i]	-1.17	-1.29	-0.12
M13	59	11.65[i]	-1.45	-1.56	-0.11
M15	27	12.93[i]	-2.33	-2.42	-0.09
M53	12	12.67[i]	-1.96	-2.01	-0.05
M71	14	13.70[i]	-0.70	-0.77	-0.07
M92	22	14.20[ii]	-2.35	-2.41	-0.06
M107	26	13.95[i]	-0.93	-1.03	-0.10
NGC5466	6	13.57[i]	-1.78	-1.87	-0.09
Open Clusters					
M67	29	9.45[iii]	+0.02	+0.03	+0.01
NGC188	8	9.88[iii]	+0.13	+0.14	+0.01
NGC2158	7	9.02[iii]	-0.15	-0.14	+0.01
NGC2420	7	9.30[iii]	-0.14	-0.13	+0.01
NGC6791	14	9.92[iii]	+0.41	+0.43	+0.02
NGC6819	23	9.36[iii]	+0.11	+0.12	+0.01
NGC7789	16	9.15[iii]	+0.02	+0.03	+0.01
Pleiades	51	8.13[iii]	-0.01	+0.05	+0.06
All	453	–	–	–	–

i: Forbes & Bridges (2010); Marn-Franch et al. (2009)

ii: Paust et al. (2007)

iii: Dias et al. (2002)

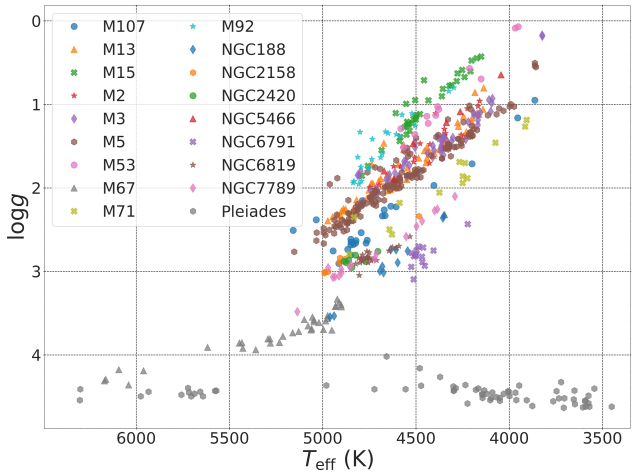

FIGURE 3.1— The figure shows the distribution of $T_{\rm eff}$ and $\log g$ for all the stars used in the unsupervised clustering exercise. For each cluster, we have used a unique combination of symbol and color, as shown in the legend.

the p-value is lower than 0.01, we can reject the null hypothesis with a confidence level of 99 percent.

In order to understand the systematics of the tests in our context, we applied them to the stars in each of the clusters. We take the median p-value over one thousand tests performed with random subsamples of each cluster. The subsamples were obtained dividing the cluster into two groups of nearly equal size. Figure 3.2 shows the results of applying the K-S test for carbon. The main diagonal presents the median p-value of the one thousand subsamples of each cluster, and the off-diagonal cells give the result for each pair of clusters. This Figure shows that it is not possible to distinguish many pairs of cluster using only their carbon abundances. For instance, M2, M3, M13, and M15 have indistinguishable carbon distributions. We see that, in some cases, even globular and open clusters have indistinguishable carbon distributions, as is the case for the pair M67-M71.

In contrast, when the K-S test is carried out for all pairs of clusters and the lowest p-value among all the 13 elements is considered, we find that, except for the pair NGC 2158-NGC 2420, all the 18 clusters are distinguishable from each other in at least one element. This is illustrated in Figure 3.3. We note that this matrix is symmetric. We explore this symmetry in Figures 3.3 and 3.4 to facilitate the direct comparison of the results of different tests. The main diagonal in the Figure is calculated as explained above for Figure 3.2, but showing only the lowest p- value among all the elements. In Figure 3.3, we combine the results for C with the ones

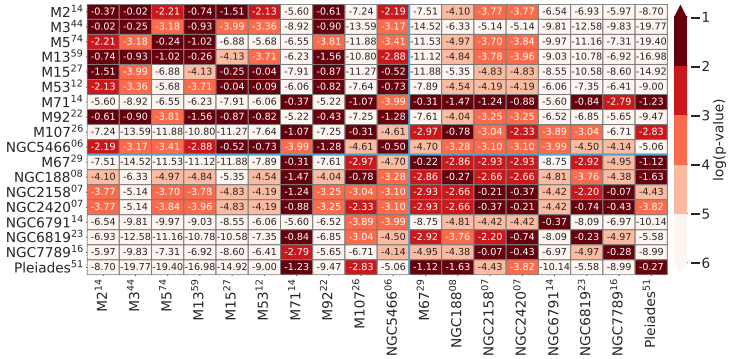

FIGURE 3.2— The figure shows the p-values of the K-S two-sample test for the carbon distributions of each pair of clusters. The cells are colored in five shades of red, from light to dark in logarithmic scale. Inside each cell, we show the p-value of the test for that particular pair of clusters. Superscripts in the cluster names indicate the number of stars in the cluster. The main diagonal shows the median p-value of one thousand tests done randomly dividing the cluster into two subsamples of nearly equal sizes. Two blue lines separate the objects in globular and open clusters, as informed in 3.1. Dark red represents values grater than 0.01, the cases where we can reject the null hypothesis with 99% confidence, i.e., when the two clusters are indistinguishable based on C only. The Figure is symmetric along the diagonal.

obtained for all elements. All the cells above the diagonal represent the minimal p-value among all elements to that pair of clusters. The cells below the main diagonal represent the results for carbon. At the top left corner of each cell, we identify the element for which the p-value is minimum and so is presented.

Three main conclusions can be drawn from Figure 3.3.

1. the probabilities stemming from the K-S test are extremely low for many of the pairs, being lower than 10^{-6} in many cases. The exceptions are a few pairs of open clusters, which have p-values typically of the order of 10^{-4}, and the indistinguishable pair NGC 2158-NGC 2420. P-values higher than average are found for the pairs M15-M92 and NGC 5466-(NGC 2158/NGC 2420), and therefore we can expect difficulties to separate the clusters in these pairs by any unsupervised clustering algorithm.

2. All clusters present high median p-values when compared with themselves. This fact underlines the cohesion of the clusters and demonstrates the consistency of the test.

3. A few elements are particularly important to distinguish the clusters from each other. Table 3.2 contains the times that each element is best for distinguishing a unique pair of clusters. Seven elements are sufficient to separate 90 percent of the pairs. Two elements, P and S, never appear as the best to distinguish

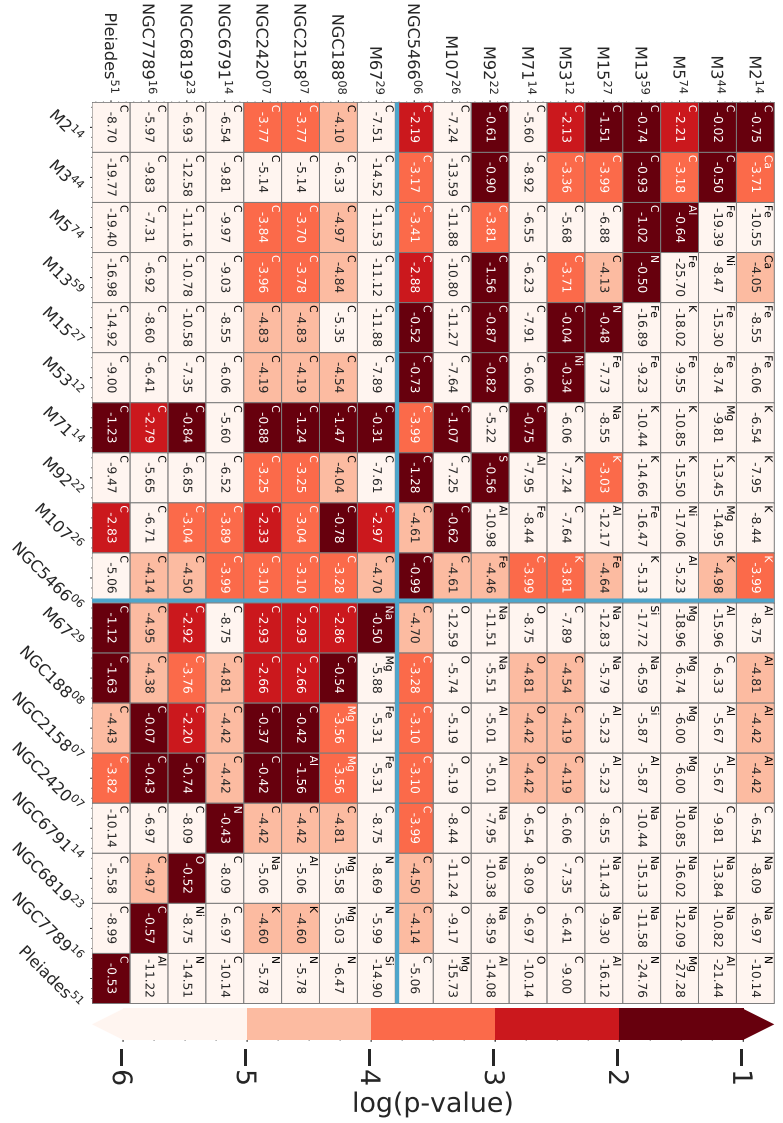

FIGURE 3.3— The Figure shows the p-values of the K-S two-sample test for each pair of clusters. The elements of the main diagonal and above the main diagonal represent the minimal p-value found among all the elements, while the elements below the main diagonal represent the p-value for carbon. The cells are colored in shades of red as shown in the color bar. The superscripts in the cluster names indicate the number of stars in the cluster. At the top left of each cell, we show the element for which the p-value is calculated. Two blue lines separate the objects in globular and open clusters, as informed in 3.1.

pairs of clusters. The elements most useful to distinguish pairs of globular clusters, are not the same as the most useful to separate open clusters, or pairs composed of a globular and an open cluster. We see that Fe and K are particularly good to separate globular clusters from each other, while C, Al, and Na are excellent to separate pair of globular and open clusters. Pairs of open clusters are best distinguished using C, N, and Mg.

The results of the Cramer test are presented in Figure 3.4. The main diagonal is also the median value of one thousand runs of the test over random subsamples of the clusters, as described for the K-S test. The Cramer test has a stochastic nature, and thus some variation is expected. The off-diagonal elements are the median value of 100 runs of the test. For all pairs with a p-value lower than 0.01, the standard deviation over the 100 runs is lower than 0.003.

Figure 3.4 shows that we cannot reject the null hypothesis (which states the two samples are compatible with a single distribution) for six pairs of clusters, M2-M5, M2-M13, M15-M92, NGC 2158-NGC 2420, NGC 2158-Pleiades, and NGC 2420-Pleiades. From these six pairs, only three have p-values higher than 10^{-5} in the K-S test. This discrepancy between the two tests is expected, since the problem of comparing multidimensional samples is not trivially reducible to its one-dimensional pieces as we intend to do in the exercise with the K-S test. However, we present the analysis with the K-S test since it is intuitive and allows us to get insight on which elements are better to distinguish the clusters from each other.

Note that even if the tests show that two clusters are chemically different, this fact does not guarantee that the samples are separable. They only evaluate whether two clusters can be generated from the very same distribution. For example, two samples generated from two Gaussian distributions with the same center and different widths would be distinguishable by these tests but would be hardly separable by any clustering algorithm.

As we stress above, the tests are valuable to establish the potentials and limitations of unsupervised clustering algorithms that we will be applying in the next sections. The Cramer test indicates that we can not expect the algorithms to completely separate M2 from M13, M2 from M15, NGC 2158 from NGC 2420, NGC 2158 from Pleiades, and NGC 2420 from Pleiades.

3.5 Clustering algorithms

We apply eight different clustering algorithms for the classification of stellar clusters in the chemical abundances space formed by [C/M], [N/M], [O/M], [Na/H], [Mg/M], [Al/H], [Si/M], [P/H], [S/M], [K/H], [Ca/M], [Fe/H], and [Ni/H]. We aim at testing how well they perform in separating the known clusters. The algorithms are: affinity propagation (Frey & Dueck, 2007), agglomerative clustering (Fränti et al., 2006), DBSCAN (Daszykowski & Walczak, 2010), K-means (Macqueen, 1967), mini-batch K-means (Bouveyron et al., 2007), spectral clustering (Ng et al., 2002), Gaussian mixing models (Bouveyron et al., 2007), and Bayesian Gaussian mixing

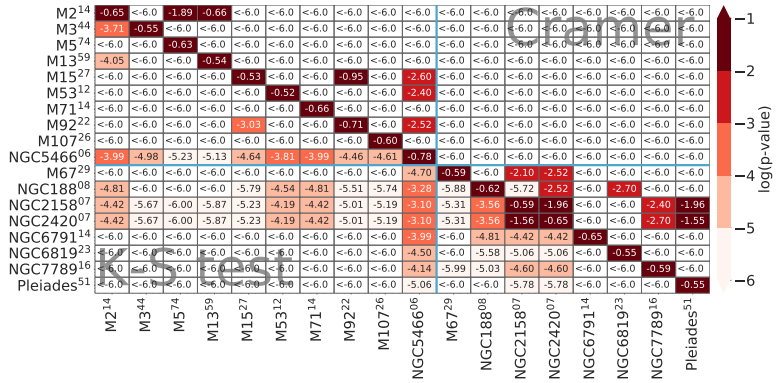

FIGURE 3.4— The image shows the median p-values of one hundred runs of the Cramer two-sample test for each pair of clusters. The elements below the main diagonal show the median value of 100 runs of the Cramer test, while the elements above the diagonal show the results for the K-S test, also included in Figure 3.3. The cells are colored in shades of red as shown in the color bar. Red colors are saturated at 0.01. Inside each cell, we show the p-value of the test for each particular pair of clusters. Superscripts in the cluster names indicate the number of stars in the cluster. Two blue lines separate the objects in globular and open clusters.

TABLE 3.2— Best chemical elements to distinguish between stellar clusters. Second to the fifth column: the number of times that each element turns out to be the best to identify clusters, and so it appears in Figure 3.3. Sixth column: median uncertainty in the measure of abundances for all the stars in the sample.

Element	N_{best}				σ_X
	glob-glob	glob-open	open-open	all comb.	
C	4	20	7	31	0.006
Al	4	15	3	22	0.015
Na	1	20	1	22	0.143
Fe	15	0	2	17	0.003
K	15	0	2	17	0.039
O	0	15	0	15	0.014
Mg	2	6	5	13	0.009
N	0	2	6	8	0.009
Ni	2	0	1	3	0.018
Si	0	2	1	3	0.009
Ca	2	0	0	2	0.025
P	0	0	0	0	0.072
S	0	0	0	0	0.074

models (Smith et al., 2013; Lartillot & Philippe, 2004). These are all the nonhierarchical clustering algorithms available in the scikit-learn library (Thirion et al., 2016). The complete description of these methods is beyond the scope of this document; we refer to Jain et al. (1999) which presents a review of these algorithms, and to the scikit-learn website which presents the documentation for the library used in this work, `http://scikit-learn.org/stable/documentation.html`.

To evaluate the performance of the algorithm for our task, we compare them through three different metrics: homogeneity score, accuracy score, and v-measure score. The *accuracy score* measures the number of coincidences between the real classification and that provided the clustering algorithm. It is a value from zero to one which represents the fraction of stars put in the right cluster. The *homogeneity score* measures in which level the predicted clusters contain only data points that are members of one real cluster: its value varies from zero to one, where one means the clusters are perfectly homogeneous. The *v-measure score* is the harmonic mean between completeness and homogeneity, where completeness is a score that evaluates the proportion of stars in the real cluster that are assigned to some of the group. Rosenberg & Hirschberg (2007) present a rigorous description of these metrics. When working with unsupervised clustering algorithms, the labels generated for the clusters can vary from one run to another. Even when the same objects are grouped together in two runs, their labels can differ. The v-measure score and homogeneity score are transparent to permutations of the labels, but the accuracy score needs all the clusters to be cross-matched. In the latter case, we have matched each group of stars found by the unsupervised tool to the star cluster with the highest number of member stars inside the group, as done in Chapter 2. However, in this chapter when the number of clusters in the real dataset does not match the number of clusters in the predicted model, or the objects in one group do not match any of the available clusters, we assign the group to the cluster with the highest number of coincident objects, even if the cluster has already been assigned to other group.

For each of these algorithms, we performed an extensive optimization of their hyperparameters searching for the highest homogeneity score. The list of hyperparameters tuned for each clustering method is shown in Tables 3.3 and 3.4, together with their optimal values. The description of each of these parameters is given in the articles cited, and also in scikit-learn documentation. The hyperparameters presented in Tables 3.3 and 3.4 are labeled exactly as in scikit-learn documentation. They are given here for the sake of comprehensiveness and reproducibility, but the details on their meaning and operation are not needed to follow the analysis we carried out.

3.5.1 Scalers

When using high-dimensional data to perform clustering, it is essential to ensure that all the variables are properly scaled. There are many standard algorithms in the literature to achieve this kind of normalization. In this chapter, we have tested all the eight clustering algorithms with two different scalers, which are known in

TABLE 3.3— List of hyperparameters explored for each algorithm. The last column shows the value of the parameter in the run with the highest homogeneity score.

Hyperparameters	Tested values	best value
Affinity Propagation		
affinity	euclidean	euclidean
convergence_iter	20	20
damping	[0.5, 0.51, ..., 0.99]	0.5
max_iter	200	200
Agglomerative Clustering		
affinity	[manhattan, cosine, euclidean]	manhattan
compute_full_tree	[False, True]	False
linkage	[complete, average]	complete
n_clusters	[9, 10, ..., 26]	26
Bayesian Gaussian Mixture		
covariance_type	[tied, diag]	diag
init_params	[kmeans, random]	kmeans
n_components	[9, 10, ..., 26]	20
n_init	5	5
random_state†	[31, 43, ..., 473]	283
warm_start	[False, True]	False
DBSCAN		
algorithm	[ball_tree, kd_tree, brute]	ball_tree
eps	[0.3, 0.31, ..., 1.49]	0.58
leaf_size	10	10
metric	euclidean	euclidean
min_samples	[2, 3]	2

†: We tried ten different random seeds, namely, 31, 43, 98, 196, 283, 294, 374, 383, 433 and 473.

TABLE 3.4— Continues from Table 3.3.

Gaussian Mixture		
covariance_type	[spherical, diag]	spherical
init_params	[kmeans, random]	kmeans
max_iter	1000	1000
n_components	[2, 3, ..., 26]	20
random_state†	[31, 43, ..., 473]	283
***K*-means**		
init	[k-means++, random]	random
n_clusters	[9, 10, ..., 26]	26
n_init	5	5
random_state†	[31, 43, ..., 473]	43
Mini-batch *K*-means		
batch_size	[10, 20, ..., 100]	90
init	[k-means++, random]	random
n_clusters	[9, 10, ..., 26]	25
random_state†	[31, 43, ..., 473]	31
reassignment_ratio	[0.1, 0.01, 0.001]	0.1
Spectral Clustering		
affinity	[rbf, sigmoid, polynomial, poly]	rbf
assign_labels	[kmeans, discretize]	discretize
degree	[3, 4, 5]	3
n_clusters	[9, 10, ..., 26]	17
n_neighbors	[2, 5, 10]	2
random_state†	[31, 43, ..., 473]	383

†: We tried ten different random seeds, namely 31, 43, 98, 196, 283, 294, 374, 383, 433 and 473.

the scikit-learn package (Thirion et al., 2016) as standard scaler and robust scaler. The standard scaler sets the mean and the variance of all dimensions to zero and one, respectively. The robust scaler sets the median to zero and scales the data according to the range between the first and the third quantile, making it more robust to outliers.

When testing the two types of scalers with all the clustering algorithms, we found the highest homogeneity scores using the standard scaler for the affinity propagation, DBSCAN, and K-means algorithms. For the other five algorithms, the highest homogeneity score is found using the robust scaler.

3.5.2 Dimensionality reduction

Dimensionality reduction is important not only to reduce the computational cost but also to eliminate redundancy among the dimensions. Having redundant dimensions can hamper the process of finding clusters by diluting the Euclidean distances in when the number of dimensions grows. This phenomenon is known as the curse of dimensionality, and it is discussed in works such as Zimek et al. (2012). We have run each clustering algorithm after reducing the dimensionality with Principal Component Analysis (PCA) (Wold et al., 1987), Linear Discriminant Analysis (LDA) (Fisher, 1936), Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000), t-distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten & Hinton, 2008), and also with no dimensionality reduction.

We have found the highest homogeneity scores using LDA for all clustering algorithms. The original data have 13 dimensions, i.e., the number of different chemical elements whose abundances are retrieved by ASPCAP. We have varied the number of dimensions from 2 to 12 for all the dimensionality reduction tools. The number of dimensions which maximizes the homogeneity score differs among the algorithms. For DBSCAN and spectral clustering, the highest homogeneity scores are found projecting the data in two dimensions only. For agglomerative clustering and K-means the optimal number of dimensions is 4. Five dimensions give the highest homogeneity score for the Gaussian mixture models and the mini batch K-means. For affinity propagation, the best number of components is 6, and for the Bayesian Gaussian mixture models the best value is 7. This analysis shows that the classical approach of using a threshold in variance to determine the optimal number of components can favor some clustering algorithms. The Figure 3.5 shows the variation of the homogeneity score with the number of dimensions for each clustering algorithm and each dimensionality reduction tool.

3.6 Results of the clustering algorithms

We evaluate the performance of the eight clustering algorithms varying their hyperparameters and combining then with four different dimensionality reductions tools and two different scaling methods. In total, we run more than three million combinations of the eight algorithms with different hyperparameters, different scalers, and different dimensionality reduction tools. Figure 3.6 shows the best

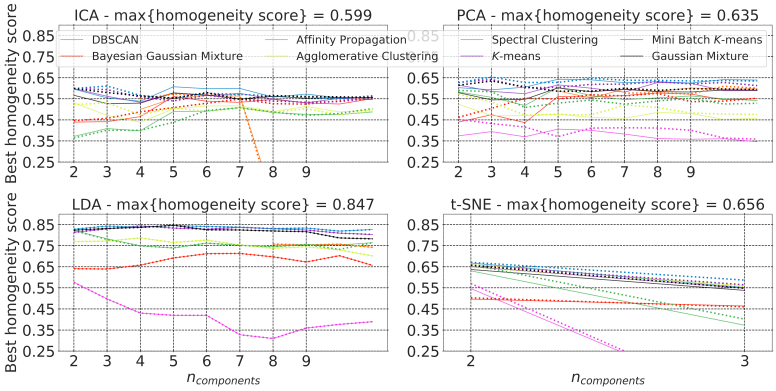

FIGURE 3.5— The panels in this figure show how the best homogeneity score varies with the variation of the number of components for each of the dimensionality reduction models. The label on the top of the different panels indicates the dimensionality reduction algorithm. Different colors are used for each clustering algorithm. The solid lines present the results applying the robust scaler, while the dotted lines show the results for the standard scaler.

scores achieved for each of the clustering algorithms when the number of clusters is 18, i.e., the number of real stellar clusters. The algorithms are ordered by their best homogeneity score. The actual values of the scores are marked on top of each bar.

These results were obtained working with the APOGEE uncalibrated elemental abundances of [C/M], [N/M], [O/M], [Na/H], [Mg/M], [Al/H], [Si/M], [P/H], [S/M], [K/H], [Ca/M], [Fe/H], and [Ni/H], from 453 stars belonging to 18 star clusters. Figure 3.7 shows the confusion matrix of the resulting clustering with the highest homogeneity score obtained when using a Gaussian mixture model. The confusion matrix compares the actual labels with those assigned by the best-performing algorithms. Each cell shows the ratio of objects belonging to the class indicated in the vertical axis classified as the class in the horizontal axis. The main diagonal presents the fraction of true positives, namely when an object is classified in the cluster it belongs to, the off-diagonal elements represent confusion between clusters, that is to say, fraction of objects from one cluster classified as belonging to another cluster. From the six pairs that appear as indistinguishable in Figure 3.4 based on the Cramer test, four of them present a degree of confusion higher than 20 percent in Figure 3.7. The pairs M3-M13, M67-NGC 6819, NGC 188-NGC 6819, on the other hand, have some mixing in the classification but are not pointed out as problematic by the Cramer test. Since the test does not guarantee separability, these cases are not surprising (see discussion in Sect.

Figure 3.5 shows the variation of homogeneity score with the number of dimensions for the four different dimensionality reduction algorithms, and for the two

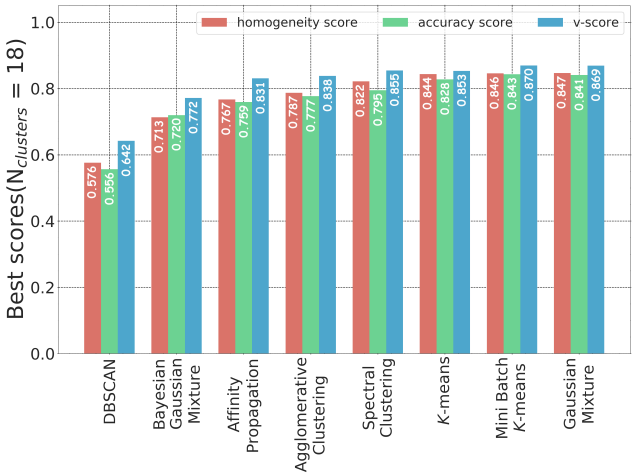

FIGURE 3.6— The bars represent the best result for each of the algorithms tested in this chapter. Different colors are used to distinguish among the metrics, as shown in the legend. In groups of three bars the leftmost represent homogeneity score, the one in the middle represents accuracy score, and the rightmost represents the v-measure score.

different scalers. The best performance regarding homogeneity is found using LDA for dimensionality reduction. We also see that the LDA performance is almost unaffected by the choice of scaler. The best result is found when we use LDA in five dimensions, with a homogeneity score of 0.847, but with two dimensions we already have a homogeneity score of 0.823. When the original 13-dimensional space are not reduced, we obtain the results shown in Table 3.5. The highest homogeneity score is found for the Gaussian mixture model, 0.708, which is much lower than what we obtain using LDA, but higher than what we observe for other dimensionality reduction tools.

LDA is a supervised dimensionality reduction algorithm based on having knowledge of the target classes to create a linear projection of the data maximizing class separation. The projection can be determined using a few clusters and then be applied to any number of stars. In the case of performing a blind search for stellar populations, we can use the known stellar populations to determine the projection and apply it to the whole sample.

For comparison with Blanco-Cuaresma et al. (2015), we present in Figure 3.8 the best scores without constraining the number of clusters to 18. In Blanco-Cuaresma et al. (2015), the best result is found with the Mitschang algorithm (Mitschang et al., 2013), with an homogeneity score of 0.86 and a v-measure score of 0.75. Our highest homogeneity score is 0.88, found using mini-batch K-means, with a v-measure score

TABLE 3.5— Highest scores found using clustering without dimensionality reduction.

Algorithm	homogeneity	accuracy	v-measure
Gaussian Mixture	0.708	0.638	0.544
Mini Batch K-means	0.639	0.578	0.595
K-means	0.615	0.490	0.577
Bayesian Gaussian Mixture	0.649	0.561	0.864
Spectral Clustering	0.558	0.532	0.819
Affinity Propagation	0.602	0.501	0.578
Agglomerative Clustering	0.568	0.477	0.592
DBSCAN	0.364	0.382	0.623

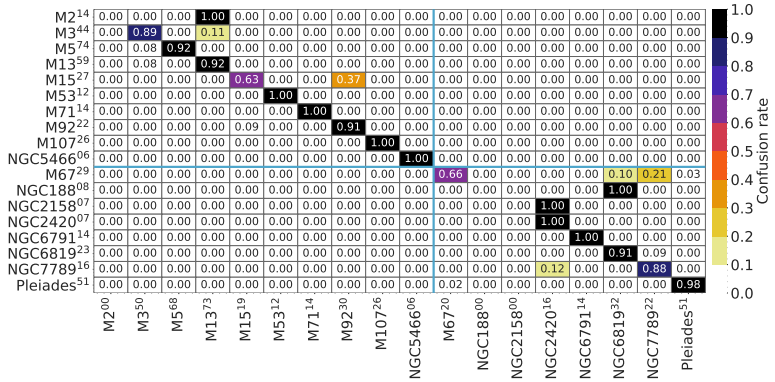

FIGURE 3.7— The confusion matrix for the best classification. The cells are color-coded according to the fraction of stars assigned as cluster members. The vertical axis represents the real clusters, and the horizontal axis represents the clusters obtained with the Gaussian Mixture model. The main diagonal shows well-classified objects, while the cells out of the diagonal represent misclassifications.

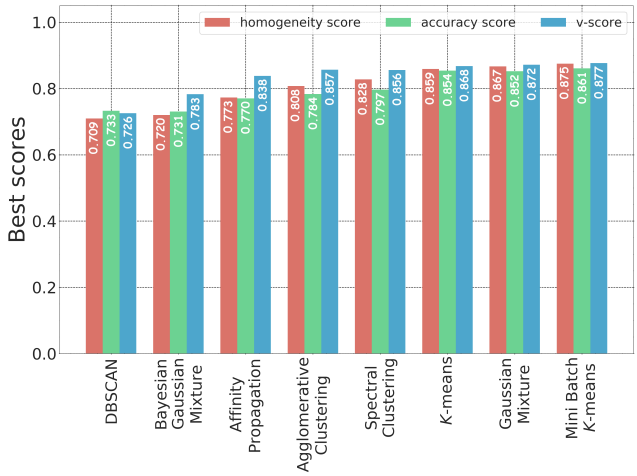

FIGURE 3.8— The bars represent the best result for each of the algorithms tested in this chapter without constraining the number of clusters. Different colors are used to distinguish among the metrics, as shown in the legend on top of the figure. In groups of three bars. The leftmost one represents the homogeneity score, the one in the middle represents the accuracy score, and the rightmost one represents the v-measure score.

of 0.88. The quality of the data used by Blanco-Cuaresma et al. (2015) is arguably better than ours, in the sense they have spectra with higher resolving power than those from APOGEE, and also they have used chemical abundances for 17 elements, while we have used only 13. However, our results show that the APOGEE data is capable of getting results on pair with those obtained by Blanco-Cuaresma et al. (2015).

3.6.1 Defining the number of clusters

In the case of a blind search for chemical populations in surveys like APOGEE, we would not have a priori knowledge of the number of clusters present. Here we examine four criteria to determine the number of clusters. We have tested the silhouette score (Rousseeuw & Kaufman, 1990), the Calnski-Harabasz index (CH index; Caliński & Harabasz 1974), the Bayesian Information Criterion (BIC; Schwarz et al. 1978), and the gap statistics (Tibshirani et al., 2001). We briefly describe these criteria in Appendix A.1.

The standard use of silhouette the score and CH index prescribes the number of clusters to be found by maximizing these indexes. For the BIC, the number of clusters is found at its minimum value, and the gap statistics chooses the optimal number of clusters as the lowest value for which the condition $\mathrm{Gap}(K+1) - s_{K+1}$ is

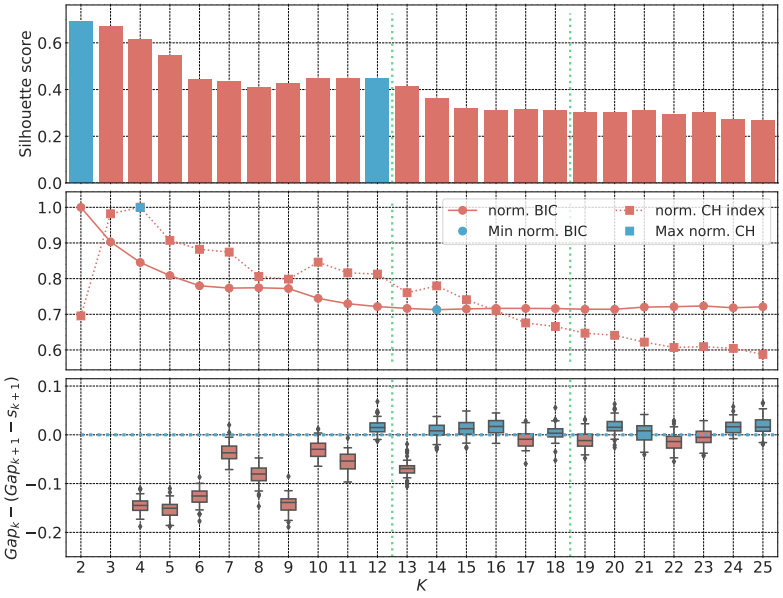

FIGURE 3.9— Criteria for finding the optimal number of clusters. We mark as blue the acceptable values according the standard prescription of the three different tests employed in this work. Vertical green lines indicate the interval in which we expect the optimal number of clusters to be. The top panel shows how the silhouette score varies with the number of clusters. The middle panel presents the normalized values for the CH index and BIC. The bottom panel shows the results corresponding to the gap statistics result as a box plot; we plot in blue those points for which the median value is greater than zero. A blue dotted line shows the zero line.

grater than zero, as defined Appendix A.1.1. Figure 3.9 shows the results for these four criteria. The Gap statistics was run 50 times with 10 random samples at each step, the median values are presented in the box plot shown in the bottom panel of Figure 3.9.

We see that using the silhouette score the optimal number of clusters is two. As we can see in Table 3.1 there is a clear split between globular and open clusters. The Table shows there is a clear difference in the iron abundance of the two populations. Probably this is the reason why silhouette score indicates there are two main clusters. A second peak in the silhouette score is found for 12 clusters. The CH index indicates four as the optimal number of clusters. The Bayesian information criterion points to 14 as the optimal number of cluster, and the gap statistics indicate the best number of clusters to be 12.

Since we have proven that six pairs of clusters are indistinguishable, we do not

expect that any of the algorithms find the 18 clusters. From Figure 3.4, we can predict that the clusters M2, M5 and M13 will be seen as a single cluster. We can also predict the combination of the clusters M15 and M92, and that the clusters NGC 2158, NGC 2420 and the Pleiades will be merged in a single cluster. Consequently, in the worst case scenario, we could expect 13 clusters to be retrieved, instead of the real 18 clusters. We mark this range with green dotted lines in Figure 3.9. Only BIC and gap statistics predict a number of clusters consistent with this interval.

3.7 Conclusions

We have explored the application of unsupervised clustering on the APOGEE survey to chemically separate star clusters from each other. We have statistically tested the feasibility of the task, and concluded that it could not be accomplished perfectly due to the intrinsic overlap of the clusters in chemical space. Eight different clustering algorithms were combined with four dimensionality reduction techniques and two scaling approaches. We have shown that the highest homogeneity score obtained from the clustering process is consistent with the outcome of the Cramer test. It reaches vaules around 0.85, meaning that 85% of the stars are properly classified.

The K-S test allows us to identify the most sensitive chemical elements for the identification of stellar populations. Seven elements are sufficient to distinguish 90 percent of the pairs of clusters. Table 3.2 shows that the best set of elements depends on the nature of the pair of objects to classify. For example, some elements are more relevant to separate globular clusters from each other (Fe and K), while other elements are more suitable to distinguish open from globular clusters (C, Al, and Na). This information is of interest when deciding whether to allocate computational and observational resources to improve the precision of abundance measurements, or to increase the spectral range of an instrument in order to expand the number of elemental species to measure.

The Cramer test suggests there are six pairs of indistinguishable clusters in our sample, namely, M2-M5, M2-M13, M15-M92, NGC 2158-NGC 2420, NGC 2158-Pleiades, and NGC 2420-Pleiades. This does not mean that the other pairs are separable since, different but overlapping distributions can be distinguishable and not separable. On the other hand, it does not mean the *indistinguishable* cluster pairs are intrinsically identical; it only implies that for the particular samples of stars and chemical elements we studied the pairs are indistinguishable. It is conceivable that the same stars measured with higher precision, the use additional chemical elements, or the availability of a larger sample of stars could lead to a different conclusion.

We tried to separate the cluster members using unsupervised classification algorithms. They reach a homogeneity score of 0.85, where the primary sources of confusion are the pairs marked as indistinguishable by the Cramer test. The best result is found using Gaussian mixture models, but K-means and mini-batch K-means give very similar results. Gaussian mixture models offer a more elegant solution, given

not only the classification of the stars but also the probability of belonging to other groups. We have found as well that the mixture of Gaussian functions can generate decision boundaries that can adapt better to the intrinsic form of the clusters, while K-means like algorithms assume hyperspherical boundaries. However, in larger samples, where the computational cost can be a constraint, mini-batch K-means would be preferable. We notice that mini-batch K-means is a approximation to accelerate

We have tried to determine the number of clusters in the sample automatically. None of the criteria were able to recover the exact number of clusters. However, the BIC and the gap statistics give results consistent with the number of clusters we expect due to the indistinguishability of some pairs of clusters.

In summary, we have tested the limits of the distinguishability of the stellar clusters with the APOGEE data and explored the clustering algorithm to reach these limits. In this sense, we have shown that the chemical identification of stellar populations is limited by the available data and the actual chemical overlap between the clusters. We have slightly improved the results of the classification compared with previous attempts.

With the chemical information provided by APOGEE, it is not possible to completely distinguish all the stellar clusters from each other. The primary source of confusion are clusters with similar ages, which is the case for the three globular clusters M2, M3 and M13, and the pair of open clusters NGC 2158 and NGC 2420.

In a recent study, Ness et al. (2018) demonstrate the existence of stars with nearly identical chemical compositions, but with a different Galactic origin, which adds on to the limitations presented here. However, our work indicates that if chemical tagging is not possible to the level of star clusters, the existing clustering algorithms can blindly identify stellar populations with similar ages and chemical distributions in the APOGEE data.

Traditionally, chemical distribution of stellar populations is used either to distinguish large-scale Galactic components, like the thin and thick disk, or to identify populations at the star cluster level. The results found in this chapter sums to the ones found by Blanco-Cuaresma et al. (2015) and Ness et al. (2018) in demonstrating the difficulties for blind searches to identify star clusters by their chemical abundances.

There is the possibility of improving the spectroscopic data or including radial velocities or proper motion in the cluster identification, moving to a chemical/cinematical tagging. Alternatively, could also think of the chemical tagging at an intermediate level between the Galactic components and the stellar clusters. If we identify many stars that are chemically very similar to the stars in a given cluster, we expect them to have a similar chemical evolution history to the cluster's, even if the two sets are spatially disconnected.

Here we propose the concept of families of stars. Since we see that clusters like M2, M3, and M13 are indistinguishable from each other, we propose that all the stars with similar distribution of these clusters and all members of these could be identified as one family of stars.

4

Searching for stellar chemical *families* of stars in APOGEE

4.1 Introduction

In Chapter 3 we have shown the main difficulties of identifying stellar populations at the level of star clusters. We also demonstrated the similarity of the clusters M2, M3, and M13. Here we define the term *families of stars* as groups of stars with very similar chemical composition, but not necessarily coming from the very same stellar cluster.

Patter recognition in multidimensional data encounters the problem of visualizing data in more than 3 or 4 dimensions. There are many techniques that try to solve this problem. For example, the parallel plots are a way of plotting several dimensions in a two dimensional plane. However, it is very inefficient to visualize large amonts of data. Another way to deal with this problem is to reduce the dimensionality of the data by projecting it in a lower dimensional space in a meaningful way. In this chapter, we use dimensionality reduction techniques to identify families of stars in the APOGEE data. We project the 13-dimensional space of abundances in 2 dimensions in order to facilitate the identification of clusters of objects visually. We work with the same four dimensionality reduction tools we have used in Sect. 3.5.2: Principal Component Analysis (PCA) (Wold et al., 1987), Linear Discriminant Analysis (LDA) (Fisher, 1936), Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000), and t-distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten & Hinton, 2008).

4.2 Data

We have used APOGEE DR14, as in Sect. 3.2, extracting information for 263,800 stars. These are all the stars in DR14 with uncalibrated abundances for

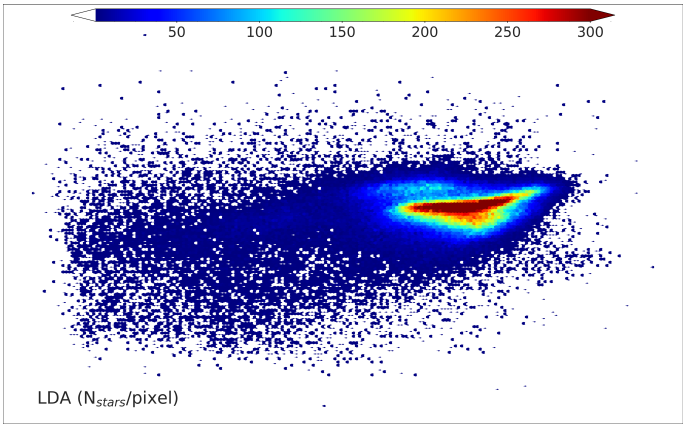

FIGURE 4.1— Density map of the LDA 2D projection of the 13-dimensional chemical abundances space. Each pixel in the image is colored according to the number of objects in that region of the map. We suppress the label of the axes since they have no direct physical meaning.

[C/M], [N/M], [O/M], [Na/H], [Mg/M], [Al/H], [Si/M], [P/H], [S/M], [K/H], [Ca/M], [Fe/H], and [Ni/H].

4.3 The crowding problem

Although LDA gives the best results when separating known stellar clusters, in Figure 4.1 we show that the 2D projection of the dataset with LDA results in a very concentrated distribution. In all images of the projections, we suppress the axis labels and tickmarks, since they have no direct physical meaning. The only observable features in the LDA map are the distributions corresponding to the thin and the thick disk components of the Galaxy. In Figure 4.2 we show the distributions of $T_{\rm eff}$, $\log g$, [C/M], and [N/M] for this projection. The four top panels of the Figure, we show the median value of these quantities for the stars in each pixel. The four bottom panels, we show the InterQuartile Ratio (IQR)[1] in each pixel. The figure shows that for many pixels the IQR is often larger than 0.1 dex for [C/M] and [N/M]. The behavior is very similar for all the other eleven elements. We are looking for a 2D projection of the 13-dimensional space which has a low IQR for the individual elements in each pixel, that is to say, a projection that guarantees that stars that are close together in 2D space have a very similar chemical pattern. In addition, the projection should not be crowded, so overdensities in the 13D space get undetectable in 2D space.

[1]The IQR is the distance between the first and the third quartiles. This is a measure similar to the standard deviation. We use median and IQR instead of mean and standard deviation to avoid the effect of outliers in the measure of the central tendency of the distribution of each pixel.

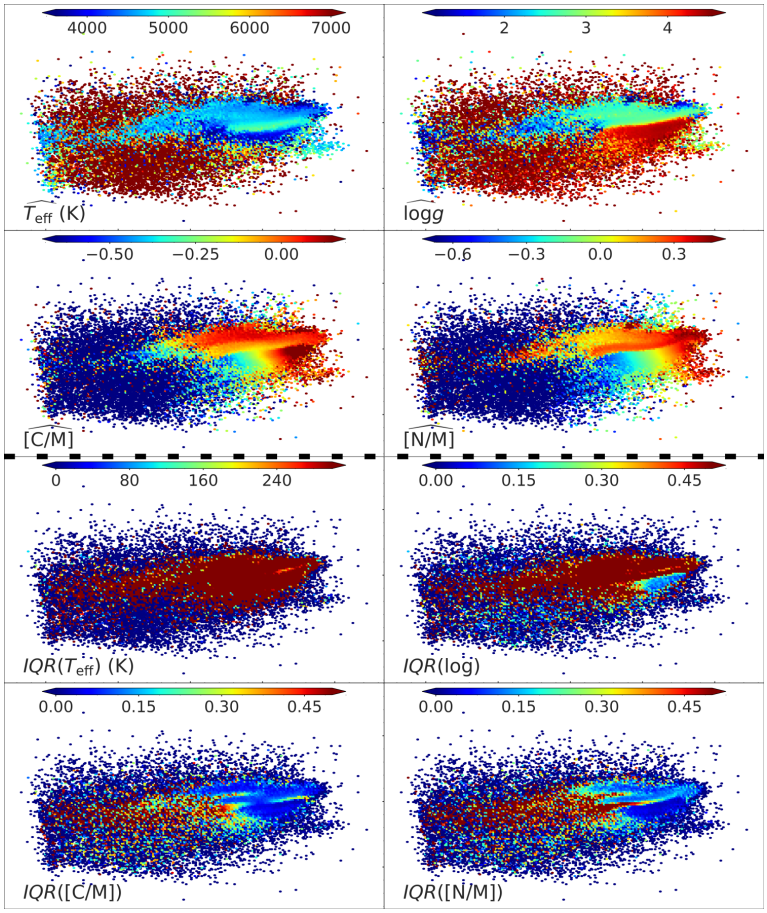

FIGURE 4.2— The color-code in the top four panels in this Figure shows the median value of the quantities indicated in the bottom left corner of each panel. The median is taken for the stars in each pixel of the image. The four bottom panels, we show the IQR for the stars in each pixel (see main text for details). We suppress the axes of the projections since they have no direct physical meaning.

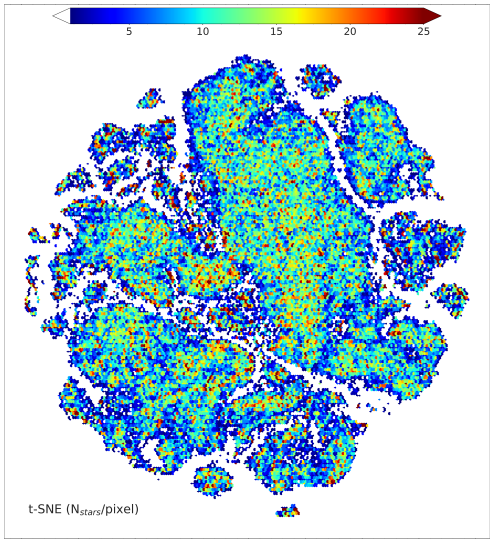

FIGURE 4.3— Density map of the t-SNE 2D projection of the 13-dimensional chemical abundances space. Each pixel in the image is colored according to the number of objects in that region of the map (see the color bar on top). We suppress the axes of the projections since they have no direct physical meaning.

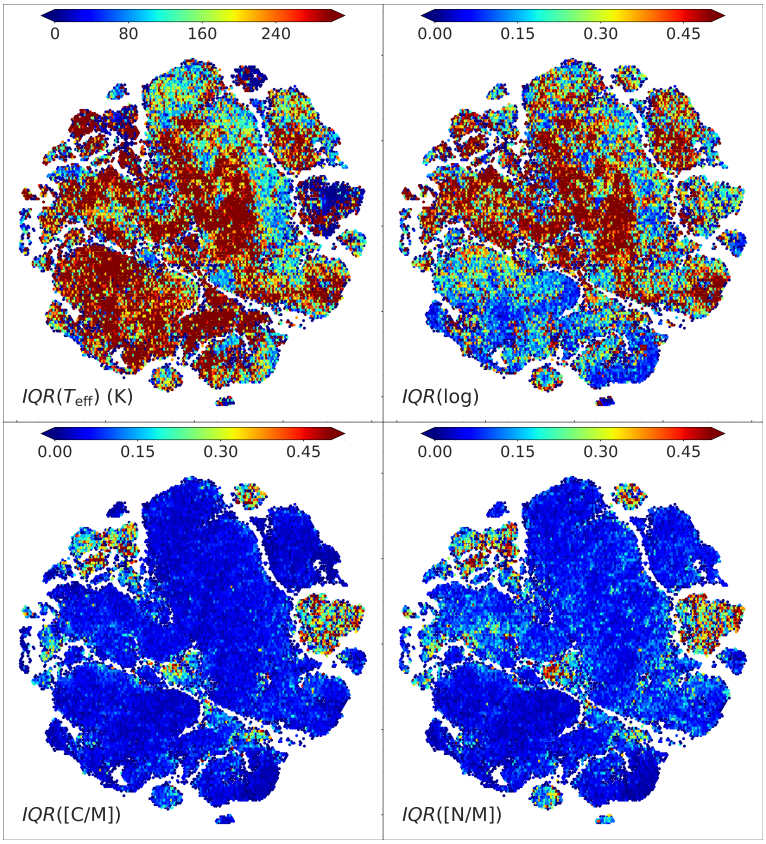

FIGURE 4.4— IQR of $T_{\rm eff}$, $\log g$, [C/M], and [N/M] in the regions of the t-SNE projection. We suppress the axes of the projections since they have no direct physical meaning.

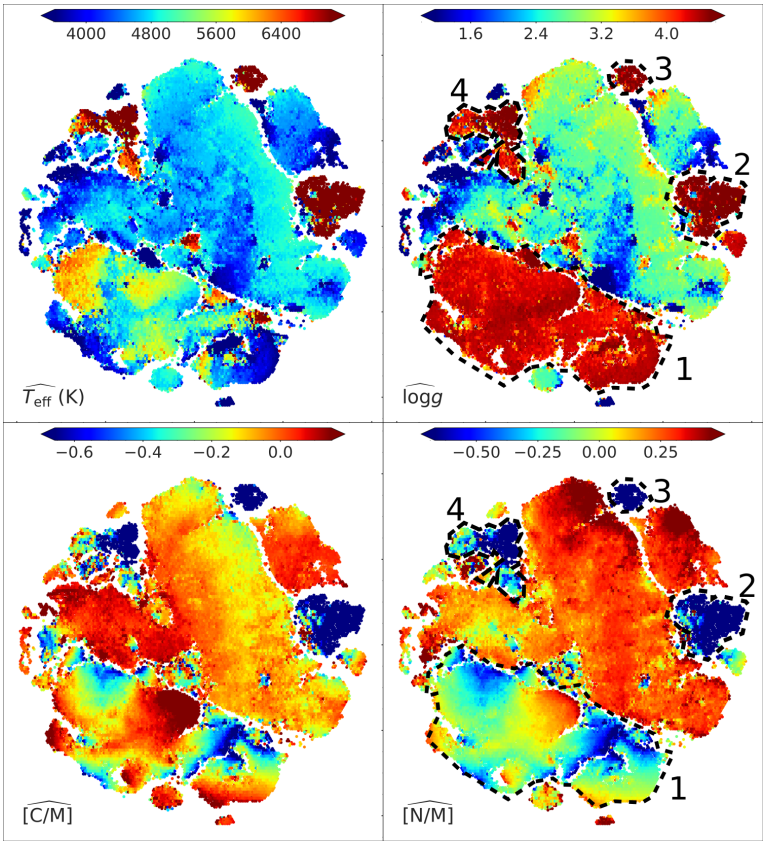

FIGURE 4.5— The Figure shows the median of $T_{\rm eff}$, $\log g$, [C/M], and [N/M] in the regions of the t-SNE projection. We suppress the axes of the projections since they have no direct physical meaning. In the right panels, dashed lines outline regions dominated by dwarf stars.

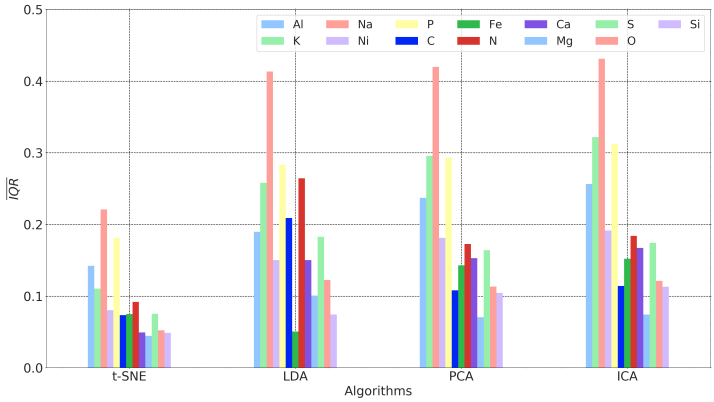

FIGURE 4.6— Mean IQR of a grid of 200 by 200 pixels for the four different dimensionality reduction projections. Different colors represent different chemical elements as given in the inset.

In Figure 4.3 we see that t-SNE produces a much less concentrated distribution. In fact, the algorithm was designed to solve the crowding problem, as stressed in Van Der Maaten & Hinton (2008). Figure 4.4 shows that the projection is very sensitive to the abundances of carbon and nitrogen. A large portion of the pixels in the figure have an IQR lower than 0.05 dex for [C/M] and [N/M]. In the right-hand panels of Figure 4.5 we highlight four regions encircled by a dashed line. Region 1 is the bulk of dwarf stars in APOGEE. We notice a clear difference in the patterns of [N/M] in the four regions. As we expect for the first dredge up, there is an increase of the nitrogen abundances when the stars evolve from dwarf to giant. Regions 2, 3 and 4 are composed of warm G-type dwarf stars. At the warm surface temperatures of these stars there are few lines in their infrared spectra, making it very difficult to determine abundances. This is the reason why the ratios [N/M] and [C/M] are very low for almost all the stars in those regions. The highlighted regions are examples of notable features of the dataset that are not easily captured by the projection with LDA. We remind the reader that these projections only consider the chemical abundances of the stars; we show the distributions of $T_{\rm eff}$ and $\log g$ to highlight the impact of the stellar evolution in the distributions of chemical elements. Figure 4.6 shows the mean IQR for a grid of 200 by 200 pixels, the same used in Figure 4.4, for each algorithm and each element. We see that t-SNE leads to the smallest IQR for all elements, except for iron which reaches a minimum with LDA.

The results presented here favor t-SNE as the best of the explored methods to find families of stars in our dataset. That is why we will focus on this algorithm in the next sections. Figure 4.7 shows the t-SNE projection colored with the median abundance of the stars in each pixel of the image. In this image we also highlight

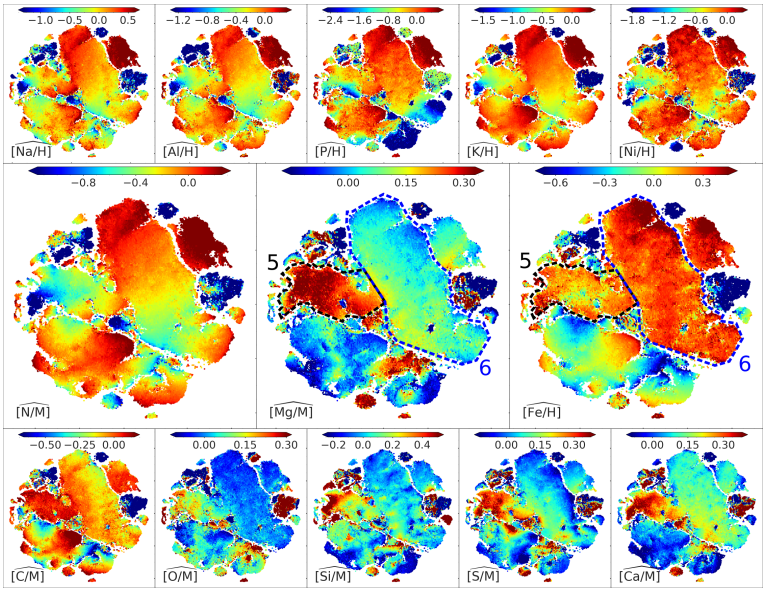

FIGURE 4.7— The Figure shows the 2D t-SNE projection colored by the median value of the stars in each pixel of the figure. In the bottom left corner of the panels we indicate the chemical element used to color the distribution. Two regions are highlighted with dashed lines. Region 5 (black line) corresponds to population from the Galactic thick disk, while Region 6 (blue line) correspond to population from the thin disk.

two other regions: Region 5 is mostly populated by the thick component of the Galactic disk, characterized by low [Fe/H] and high [Mg/M], while Region 6 is mostly composed of stars from the thin disk, characterized by high [Fe/H] and moderate [Mg/M], as shown in Figure 4.7. In fact, using the distances from Bailer-Jones et al. (2018), Figure 4.8 shows how Region 6 is mostly concentrated at small distances from the Galactic plane (Z_g) while Region 5 extends to larger distances from the plane.

4.4 t-SNE

Since we will focus on t-SNE in this Chapter, we will give a brief description of the algorithm. The complete derivation of the algorithm is given in Van Der Maaten & Hinton (2008), Van Der Maaten (2014), and Linderman & Steinerberger (2017).

t-SNE is a modified version of the Stochastic Neighbor Embedding (SNE) algorithm. Let S be a set of n objects, $S = \{\vec{x}_0, \vec{x}_1, ..., \vec{x}_n\}$, each defined for a series of N parameters, $\vec{x_i} = (a_1, a_2, ..., a_N)$. We want to transform S to $S' = \{\vec{y}_0, \vec{y}_1, ..., \vec{y}_n\}$,

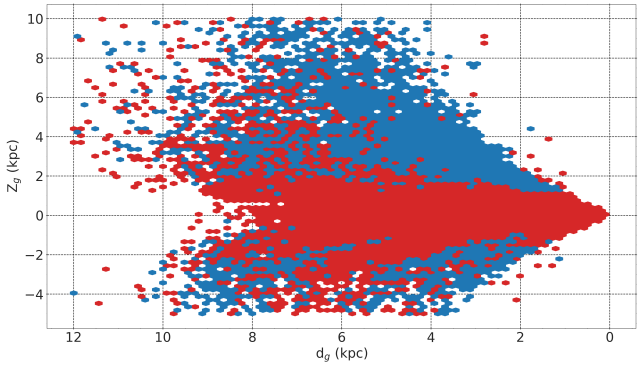

FIGURE 4.8— Distribution of distances from the galactic plane (Z_g) in the vertical axis and from the galactic center (d_g) in the horizontal axis. Each hexagonal pixel is colored according the mode in of the region in the pixel. Blue represents the Region 5 in 4.7, while Region 6 is represented in red.

where objects y are defined in less than N dimensions, typically two $\vec{y_i} = (b_1, b_2)$. To make this transformation, SNE centers a Gaussian at each point $\vec{x_i}$ and calculates a similarity metric based in the conditional probability of $\vec{x_j}$ given the Gaussian distribution centered at $\vec{x_i}$,

$$p_{j|i} = \frac{e^{\frac{||\vec{x_i} - \vec{x_j}||^2}{2\sigma_i}}}{\sum_{k \neq j} e^{\frac{||\vec{x_i} - \vec{x_k}||^2}{2\sigma_i}}}, \tag{4.1}$$

where σ_i is tunned for each point according to the local density. The tunning of σ_i is done through the perplexity value

$$\text{Perp} = 2^{-\sum_j p_{j|i} \log_2(p_{j|i})}. \tag{4.2}$$

The perplexity value is an input of the algorithm; and for each point in the dataset, σ_i is tunned to meet the value of Perp.

Once we have the pairwise similarity among the points $p_{j|i}$ we try to find a projection that reproduces the similarities in the high-dimensional space. The t-SNE approach is to calculate the similarities in the projected space with a t-distribution,

$$q_{j|i} = \frac{(1 + ||\vec{y_i} - \vec{y_j}||^2)^{-1}}{\sum_{k \neq l}(1 + ||\vec{y_k} - \vec{y_l}||^2)^{-1}} \tag{4.3}$$

There is a problem with outlier objects where $p_{j|i}$ is very different from $p_{i|j}$. It difficults the convergence of the algorithm for this points. To circumvent the problem,

a symmetric version of pairwise similarities are used, namely,

$$p_{ij} = p_{ji} = \frac{p_{j|i} + p_{i|j}}{2N},\ q_{ij} = q_{ji} = \frac{q_{j|i} + q_{i|j}}{2N}. \tag{4.4}$$

With this one can write the Kullback-Leibler cost function as

$$\text{Cost} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{4.5}$$

and gradient descent is used to minimize this function and get the best projection for each point iteratively. The gradient descent takes the form

$$\frac{\delta \text{Cost}(\vec{y_i})}{\delta \vec{y_i}} = 4 \sum_j (p_{ij} - q_{ij}) \frac{(\vec{y_i} - \vec{y_j})}{(1 + ||\vec{y_i} - \vec{y_j}||^2)}. \tag{4.6}$$

Summarizing, $\vec{y_i}$ is the projection of $\vec{x_i}$ in the lower dimensional space. The solution of Eq. 4.6 is achieved by starting with a random guess of the $\vec{y_i}$ and the posterior gradient descent optimization in a iterative fashion.

A series of approximations and optimizations are done to accelerate each phase of the t-SNE algorithm. We refer to the works of Van Der Maaten & Hinton (2008), Van Der Maaten (2014), and Linderman & Steinerberger (2017) for more details about the computational solutions of the problem.

4.5 DBSCAN on t-SNE

We have applied the Density-based spatial clustering of applications with noise (DBSCAN; Daszykowski & Walczak 2010) to group star in clusters on the 2D t-SNE projection. DBSCAN is ideal for this kind of problem because it is sensitive to subtle variations in density. Figure 4.9 we illustrate the workings of this clustering algorithm. The fundamental parameters in DBSCAN are the neighborhood radius (r) and the minimum number of objects to define a core object ($MinObj$). A core object is defined as a object with at least $MinObj$ objects, counting itself, within a radius r. A cluster needs to have at least one core object, and include all objects that are *density reachable* from the core object. A object $\vec{x}_n$ is density-reachable from $\vec{x}_i$ if there is a path $\vec{x}_i$, $\vec{x}_{i+1}$, ..., $\vec{x}_n$ in which all pairs ($\vec{x}_i$, $\vec{x}_{i+1}$) are core objects and *directly reachable* from each other. A object is directly reachable from $\vec{x}_i$ if it is inside of the neighborhood (a hypersphere of radius r) of this object. In Figure 4.9 we show core objects with $MinObj = 4$ in red. If a object is not density-reachable from any core object, it is defined as an outlier, as the blue object shown in Figure 4.9.

Figure 4.10 shows the result of DBSCAN applied to the 2D t-SNE projection. To avoid confusion between stellar clusters and clusters found with DBSCAN we use the word *class* to refer to the latter (not to be confused with the classes used in Sect. 4.3). The DBSCAN classification results in 1102 classes. We have used $r = 0.18$ and $MinObj = 5$. These values were chosen to optimize the number of

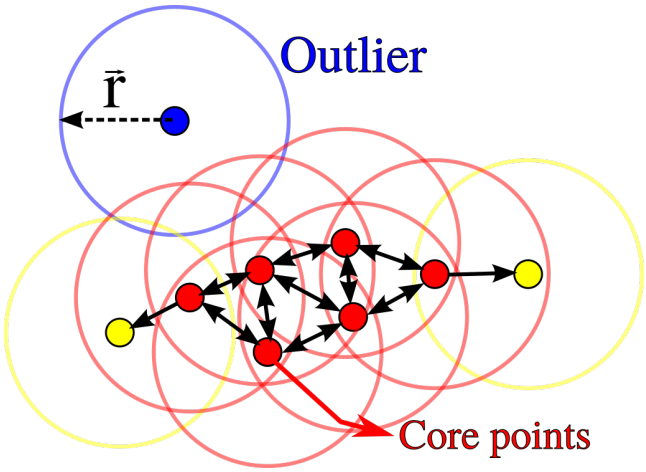

FIGURE 4.9— Illustration of the DBSCAN algorithm. In red we show the core objects, the yellow objects are density-reachable objects from any core object, and the blue object is an example of an outlier in the DBSCAN classification.

small classes, with 10 to 200 stars per class. In the left-hand panel of Figure 4.10 we show the entire sample colored by class, highlighting four classes. Zooming-in the area around each of a few highlighted classes results in the four right panels in the figure. These four classes were chosen because a significant part of their stars belong to the clusters studied in the Chapter 3.

Figure 4.11 shows the nine most populated classes, excluding the ones most representative of Regions 1 to 6, which were mentioned in Section 4.3. We see the clear separation of the classes in dwarf and giant stars. Classes 5, 16, 20, 21, and 40 were part of what we called Region 1. In Figure 4.11 we see how they differ in their distributions of atmospheric parameters and in the abundances of Mg and Fe. Class 28 is very metal-poor, with high Mg abundances, compatible with the distribution of stars in the Milky Way halo. Class 13 corresponds to the most metal-rich disk giants, while Class 1 contains the most metal poor giants in the thin disk. Class 9 is part of what we called Region 2, and these stars are near the upper limit in temperature for APOGEE.

4.5.1 Cluster Families

In order to illustrate the power of using t-SNE plus DBSCAN to identify stars with the same chemical composition, we examine in more detail the four classes highlighted in the right panels of Figure 4.10. Table 4.1 shows the number of stars in the vicinity of each real globular cluster for each class. Class 288 is composed of

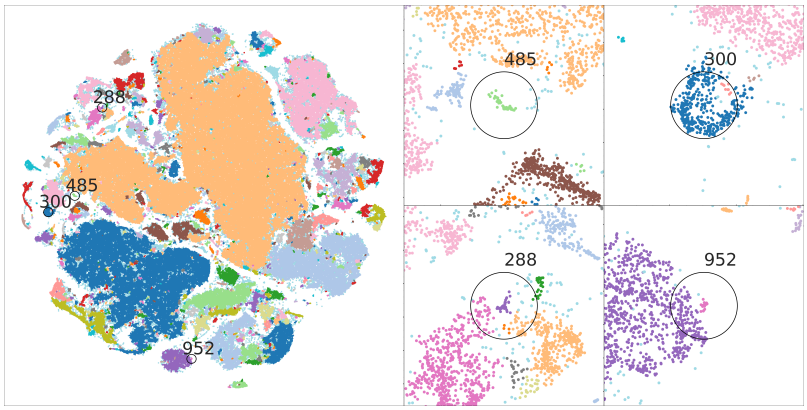

FIGURE 4.10— The left panel in the Figures shows the t-SNE projection colored by the classes in the DBSCAN classification. There are 1102 classes in the image, therefore many classes share the same color. In the right panels, we show four zooms in the regions with classes that correspond to families of stars associated with globular clusters.

17 stars from M15, 12 stars from M92 and 11 field stars. We call this class the *M15 family* because M15 is the cluster with more stars in this class.

Class 300 is composed of 411 stars, 297 of which are in the direction of ten globular clusters. We call Class 300 the *M5 family* because M5 is the cluster with the largest number of stars in the Class. All the other 114 stars are out of the plane of the Galaxy, as expected for globular clusters, but without clear spatial correlation with the clusters. In Figure 4.12, we show images for the six clusters with the largest number of stars in Class 300, identifying the stars in the class.

Class 485 contains stars from M71, M107, NGC 6316, NGC 6760, as shown in Table 4.1. Class 952 contains almost only stars from M107, having only two field stars. However, M107 appears in many classes, suggesting that some stars in M107 have a particular distribution of abundances, while others are similar to the stars in other globular clusters.

These examples of classes in the DBSCAN classification demonstrate the feasibility of using this combination of algorithms to identify families of stars with a meaningful relation. For instance, in Chapter 3 we show the clusters M2, M3, and M15 are indistinguishable, the Class 300 recover these clusters and some other very similar objects.

4.5.2 Sagittarius stream

In our sample there are 467 confirmed members of the Sagittarius stream (Majewski et al., 2013; Majewski et al., 2017), 232 of which are members of Classes 103, 3025, 410, and 509. In Table 4.2 we show the number of stars in each of these classes

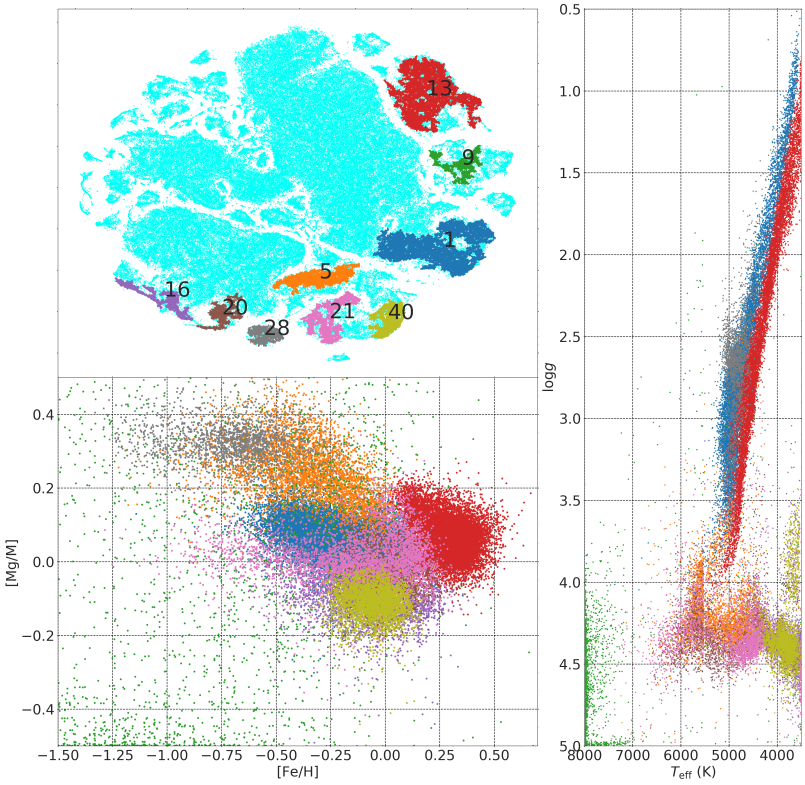

FIGURE 4.11— The top left panel highlights nine large classes in the 2D t-SNE projection. The bottom left panel we shows this classes in the plane [Mg/M] vs. [Fe/H]. The right panel shows the same classes in the plane $T_{\rm eff}$ vs. $\log g$. The classes are colored equally in all panels. All classes that are not enumerated are colored as cyan in the top left panel.

TABLE 4.1— Number of stars in each class that belong to globular clusters. *Others* refers to field stars in the class that do not belong to any globular cluster.

Cluster	$N_\star$
M15 Family - Class 288	
M15	17
M92	12
Others	11
M5 Family - Class 300	
M2	11
M3	53
M5	116
M12	12
M13	83
M53	7
M54	4
M107	4
NGC 6522	4
NGC 6544	3
Others	114
M71 Family - Class 485	
M71	13
M107	5
NGC 6316	2
NGC 6760	4
Others	32
M107 - Class 952	
M107	17
Others	3

TABLE 4.2— Number of stars in each class, the number of confirmed Sagittarius stream members and number of stars in the vicinity of the nucleus of the Sagittarius stream.

$N_{\rm class}$	$N_{\rm members}$	$N_{\rm dir}$
Class 103		
655	120	112
Class 302		
109	59	54
Class 410		
14	11	11
Class 509		
65	57	55

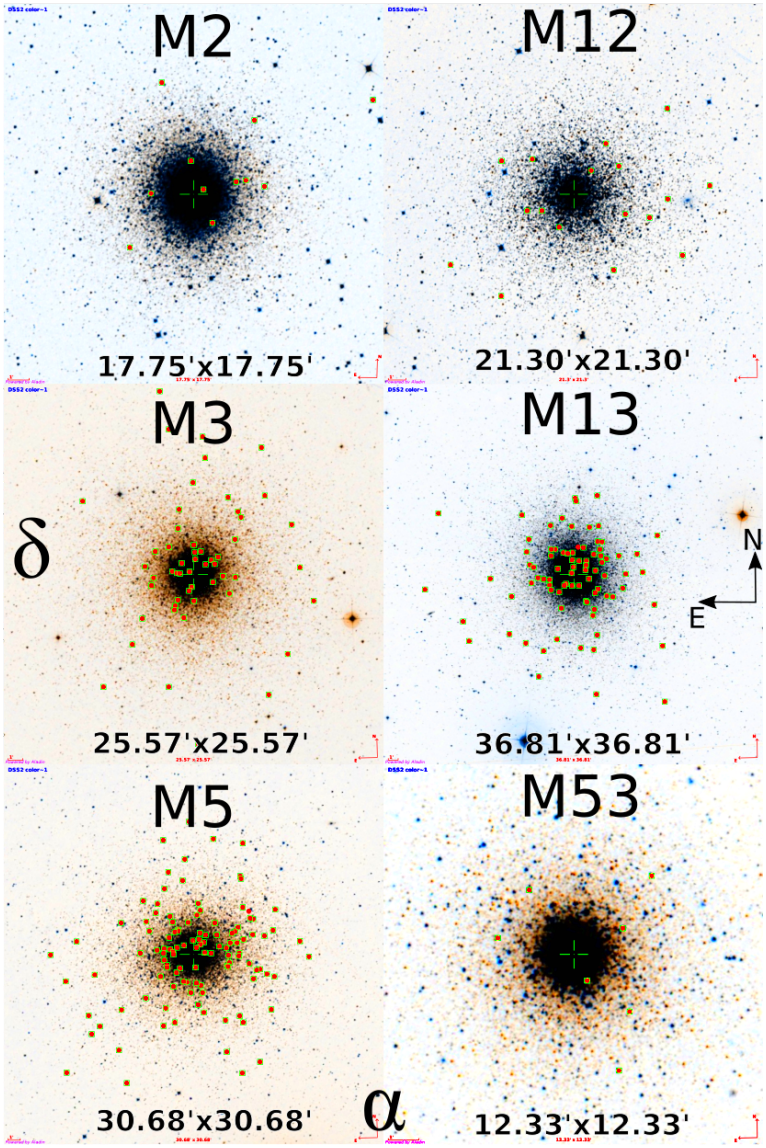

FIGURE 4.12— Images from the Digitized Sky Surveys (https://archive.stsci.edu/dss) of the six clusters with the largest number of stars in group 300; what we call the M15 family. We mark with red dots the stars in the cluster that are in Class 300. All the figures are oriented with north pointing to the top and east pointing to the left.

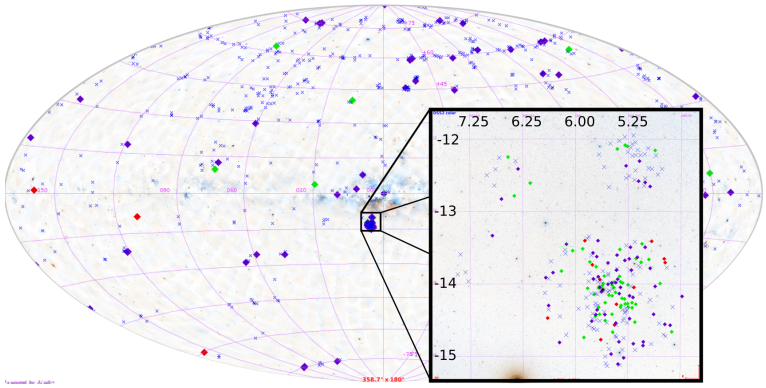

FIGURE 4.13— The distribution of the Classes 103, 302, 410, and 509 in the plane of the Galaxy. We present a zoom into the direction of the core of Sagittarius stream. Class 103 is presented with blue crosses, Class 302 with blue diamonds, Class 410 with red diamonds, and Class 509 with green diamonds. The axes correspond to the galactic longitude (abscissa) and latitude(ordinates).

($N_{\rm class}$), the number of confirmed members of Sagittarius in the class ($N_{\rm members}$), and the number of stars in the direction of the nucleus of Sagittarius stream ($N_{\rm dir}$). We see that from the total of 843 stars in this classes, 479 are either confirmed members or stars in the direction the stream. Figure 4.13 shows the distribution of these classes in galactic coordinates.

These classes demonstrate the potential of t-SNE and DBSCAN to identify stellar families blindly. These classes contain a great number of confirmed members of the Sagittarius stream and also contain stars with a very similar distribution of chemical abundances in galactic coordinates that are compatible with the known members of the stream. It sums to the findings in Sec. 4.5.1 in establish this method to identify families of stars in the APOGEE dataset.

4.6 Conclusions

In this Chapter we use t-SNE to reduce the 13-dimensional space of abundances to a 2-dimensional projection. We have shown that the projection captures the most fundamental structures in the dataset, for example, separating giants from dwarf stars given their differences in nitrogen abundances. The separation of the thin and thick disk populations is also evident. We use DBSCAN to find over-densities in the t-SNE projection. Some of the classes found by DBSCAN are associated with *families* of globular clusters. We have shown that there are also classes related to the Sagittarius stream.

We conjecture that the stars that are in classes related to the Sagittarius stream may belong to the stream. The same can be suggested for the stars belonging to

the families of globular clusters.

Further analysis is needed to determine the nature of these objects and if they are connected to the globular clusters or to the Sagittarius stream. With the proper motions provided by the Gaia mission, we could integrate the orbits of these stars and try to reconstruct their history.

There are 1102 classes in our classification. The results presented here demonstrate the feasibility of the blind identification of cluster families through their chemical abundances using simple examples of known groups. In the future, we expect to explore the other classes in the classification in order to find new members of known stellar clusters or uncover new stellar populations.

5

Conclusions

The primary objective of this thesis is to explore the application of unsupervised machine learning algorithms to find patterns in the APOGEE dataset. In spectroscopic surveys like APOGEE, there are two main approaches to pattern recognition. We can either work directly with the stellar spectra or we can use the chemical abundances measured from the spectra.

One can argue about the virtues and issues of either approaches. The direct use of the spectra has the advantage of being independent of model atmospheres. Since model atmospheres and line formation calculations involve many approximations, relying upon these models can be a disadvantage. However, one may argue that we should not throw away knowledge and so the use of the stellar abundances should be favored.

In this thesis, we have explored both approaches. In Chapter 2 we have used K-means to classify 153,847 spectra in 50 classes. We have discussed the difficulties of determining the number of clusters in this kind of dataset, and we have proposed a practical methodology to this end. K-means has a stochastic nature, related to its random initialization. We have discussed the effects of this feature in the final classification and measured its impact. We have also discussed the virtues and limitations of the classification with K-means, and proposed alternatives to this algorithm in the concluding Section of Chapter 2.

Apart from evaluating the performance of K-means, we have carried out a detailed analysis of the classes generated with K-means. The results show that the spectral classification is mainly sensitive to the atmospheric parameters of the stars, $T_{\rm eff}$ and $\log g$, and less strongly to their metallicity. We have provided extensive supporting material to help the exploration of the survey through these classes, and suggest a series of possible applications for the final catalog. The results have been published in the journal *Astronomy & astrophysics* (Garcia-Dias et al., 2018), and the catalog of the classes is available at http://vizier.cfa.harvard.edu/viz-bin/VizieR?-source=J/A+A/612/A98.

In Chapter 3 we move-on to applying pattern recognition to the chemical abun-

dances of the stars in APOGEE DR14. The primary objective of this Chapter is to evaluate the feasibility of a blind search for star clusters in chemical abundances space. To this end, we have investigated the simplest scenario, and attempted to recover stellar populations in a sample where all stars belong to known stellar clusters. We have proven that there are stellar clusters that have indistinguishable chemical distributions. Although there are clusters that are indistinguishable from each other through the chemical abundances measured by APOGEE, we show that the clustering algorithms are efficient in separating different *families of clusters*.

Based in the results obtained in Chapter 3 we define the notion of *families of stars*, and in Chapter 4 we start a blind search for them in APOGEE. Using t-SNE, we project the 13-dimensional abundance space in 2 dimensions. Then, we extract classes from this projection using DBSCAN, a density-based clustering algorithm. We demonstrate that some of the generated classes are related to families of star clusters or with the Sagittarius stream. Further investigation is necessary to determine the meaning of the other classes in the DBSCAN classification, but the Chapter provides a proof of concept for chemical tagging at the level of families of stellar clusters.

The t-SNE projection makes evident some features of the APOGEE dataset. For example, the differences between dwarfs and giants are manifest in the t-SNE projection. The projection also facilitates the separation of the thin and thick disk populations.

Further analysis is needed to determine the meaning of the classes determined with DBSCAN. We plan to crossmatch the APOGEE data with the proper motion information provided by the Gaia mission to make an analysis similar to the one presented in Chapter 4. With this, we expect to be able to identify novel stellar families in the Milky Way, or uncover relations among field stars presently thought to be unrelated.

We also plan to use the proper motion information of the stars to reconstruct their orbits and test the possibility of recovering the history of dissolved stellar clusters.

We live in an effervescent moment in the production of astronomical data. The moment offers an excellent opportunity for the application of machine learning algorithms. In this thesis, we have made a small step in testing these algorithms for the production of scientific insights. There is room for improvement in the techniques and their applications. In this thesis, we provide not only relevant scientific results in the studied field, but also a guide for the application of these algorithms to similar datasets.

References

Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42, doi: 10.3847/1538-4365/aa9e8a

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12, doi: 10.1088/0067-0049/219/1/12

Allende Prieto, C., Beers, T. C., Li, Y., et al. 2004, Origin and Evolution of the Elements

Allende Prieto, C., Beers, T. C., Wilhelm, R., et al. 2006, ApJ, 636, 804, doi: 10.1086/498131

Anders, F., Chiappini, C., Santiago, B. X., et al. 2018

Asplund, M., Grevesse, N., & Sauval, A. J. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 336, Cosmic Abundances as Records of Stellar Evolution and Nucleosynthesis, ed. T. G. Barnes, III & F. N. Bash, 25

Bailer-Jones, C. A. L. 2001, arXiv:astro-ph. https://arxiv.org/abs/0102223v1

Bailer-Jones, C. A. L., Irwin, M., & Von Hippel, T. 1998, Monthly Notices of the Royal Astronomical Society, 298, 361, doi: 10.1046/j.1365-8711.1998.01596.x

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, ArXiv e-prints. https://arxiv.org/abs/1804.10121

Baringhaus, L., & Franz, C. 2004, Journal of Multivariate Analysis, 88, 190, doi: 10.1016/S0047-259X(03)00079-4

Bensby, T., Feltzing, S., & Lundström, I. 2003, A&A, 410, 527, doi: 10.1051/0004-6361:20031213

Bensby, T., Zenn, A. R., Oey, M. S., & Feltzing, S. 2007, ApJ, 663, L13, doi: 10.1086/519792

Bertran de Lis, S., Allende Prieto, C., Majewski, S. R., et al. 2016, A&A, 590, A74, doi: 10.1051/0004-6361/201527827

Binney, J., & Merrifield, M. 1998, Galactic astronomy (Princeton University Press)

Blanco-Cuaresma, S., Soubiran, C., Heiter, U., et al. 2015, Astronomy & Astrophysics, 577, A47, doi: 10.1051/0004-6361/201425232

Boesso, R., & Rocha-Pinto, H. J. 2018, Monthly Notices of the Royal Astronomical Society, 474, 4010, doi: 10.1093/mnras/stx2742

Bouveyron, C., Girard, S., & Schmid, C. 2007, Computational Statistics & Data Analysis, 52, 502, doi: 10.1016/j.csda.2007.02.009

Bovy, J., Nidever, D. L., Rix, H.-W., et al. 2014, ApJ, 790, 127, doi: 10.1088/0004-637X/790/2/127

Caliński, T., & Harabasz, J. 1974, Communications in Statistics-theory and Methods, 3, 1

Castelli, F., & Kurucz, R. L. 2004, ArXiv Astrophysics e-prints

Chen, S. S., & Gopalakrishnan, P. S. 1998, in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 2 (IEEE), 645–648. http://ieeexplore.ieee.org/document/675347/

Cunha, K., Smith, V. V., Johnson, J. A., et al. 2015, ApJ, 798, L41, doi: 10.1088/2041-8205/798/2/L41

da Silva, R., Porto de Mello, G. F., Milone, A. C., et al. 2012, Astronomy & Astrophysics, 542, A84, doi: 10.1051/0004-6361/201118751

Daniel, S. F., Connolly, A., Schneider, J., Vanderplas, J., & Xiong, L. 2011, AJ, 142, 203, doi: 10.1088/0004-6256/142/6/203

Darling, D. A. 1957, The Annals of Mathematical Statistics, 28, 823

Daszykowski, M., & Walczak, B. 2010, in Comprehensive Chemometrics, Vol. 2 (https://www.aaai.org/), 635–654. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871, doi: 10.1051/0004-6361:20020668

Elias, F., Alfaro, E. J., & Cabrera-Cao, J. 2006, The Astronomical Journal, 132, 1052

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2011, An Introduction to Classification and Clustering (John Wiley & Sons, Ltd), 1–13. http://dx.doi.org/10.1002/9780470977811.ch1

Fisher, R. A. 1936, Annals of Eugenics, 7, 179, doi: 10.1111/j.1469-1809.1936.tb02137.x

Forbes, D. A., & Bridges, T. 2010, Monthly Notices of the Royal Astronomical Society, 404, 1203

Francis, C., & Anderson, E. 2014, Monthly Notices of the Royal Astronomical Society, 441, 1105, doi: 10.1093/mnras/stu631

Fränti, P., Virmajoki, O., & Hautamäki, V. 2006, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 1875, doi: 10.1109/TPAMI.2006.227

Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487, doi: 10.1146/annurev.astro.40.060401.093840

Frey, B. J., & Dueck, D. 2007, Science, 315, 972, doi: 10.1126/science.1136800

Frinchaboy, P. M., Thompson, B., Jackson, K. M., et al. 2013, ApJ, 777, L1, doi: 10.1088/2041-8205/777/1/L1

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, A&A, 595, A2, doi: 10.1051/0004-6361/201629512

Garcia-Dias, R., Allende Prieto, C., Sánchez Almeida, J., & Ordovás-Pascual, I. 2018, A&A, 612, A98, doi: 10.1051/0004-6361/201732134

García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144, doi: 10.3847/0004-6256/151/6/144

Giridhar, S., Muneer, S., & Goswami, A. 2006, Memorie della Società Astronomica Italiana, 77, 1130

Gray, D. F. 2008, The Observation and Analysis of Stellar Photospheres

Gressler, W., DeVries, J., Hileman, E., et al. 2014, in Proc. SPIE, Vol. 9145, Ground-based and Airborne Telescopes V, 91451A. http://adsabs.harvard.edu/abs/2014SPIE.9145E..1AG

Gustafsson, B., Edvardsson, B., Eriksson, K., et al. 2008, A&A, 486, 951, doi: 10.1051/0004-6361:200809724

Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, ApJ, 808, 132, doi: 10.1088/0004-637X/808/2/132

Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, ApJ, 833, 262, doi: 10.3847/1538-4357/833/2/262

Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148, doi: 10.1088/0004-6256/150/5/148

Hubeny, I., & Mihalas, D. 2014, Theory of Stellar Atmospheres

Hyvärinen, A., & Oja, E. 2000, Neural Networks, 13, 411

Jain, A. K. 2010, Pattern recognition letters, 31, 651

Jain, A. K., Murty, M. N., & Flynn, P. J. 1999, ACM Comput. Surv., 31, 264, doi: 10.1145/331499.331504

Jofré, P., Das, P., Bertranpetit, J., & Foley, R. 2017, Monthly Notices of the Royal Astronomical Society, 467, 1140, doi: 10.1093/mnras/stx075

Koesterke, L. 2009, in American Institute of Physics Conference Series, Vol. 1171, American Institute of Physics Conference Series, ed. I. Hubeny, J. M. Stone, K. MacGregor, & K. Werner, 73–84

Koesterke, L., Allende Prieto, C., & Lambert, D. L. 2008, ApJ, 680, 764, doi: 10.1086/587471

Kos, J., Bland-Hawthorn, J., Freeman, K., et al. 2018, MNRAS, 473, 4612, doi: 10.1093/mnras/stx2637

Krzanowski, W. J., & Lai, Y. T. 1988, Biometrics, 44, 23

Lartillot, N., & Philippe, H. 2004, Molecular Biology and Evolution, 21, 1095, doi: 10.1093/molbev/msh112

Linden, S. T., Pryal, M., Hayes, C. R., et al. 2017, ApJ, 842, 49, doi: 10.3847/1538-4357/aa6f17

Linderman, G. C., & Steinerberger, S. 2017, eprint arXiv:1706.02582. https://arxiv.org/abs/1706.02582

Macqueen, J. 1967, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281, doi: citeulike-article-id:6083430

MacQueen, J., et al. 1967, in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA. (Oakland, CA, USA.), 281–297

Majewski, S. R., Hasselquist, S., Łokas, E. L., et al. 2013, The Astrophysical Journal, 777, L13, doi: 10.1088/2041-8205/777/1/L13

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94, doi: 10.3847/1538-3881/aa784d

Manteiga, M., Carricajo, I., Rodríguez, A., Dafonte, C., & Arcay, B. 2009, The Astronomical Journal, 137, 3245, doi: 10.1088/0004-6256/137/2/3245

Marn-Franch, A., Aparicio, A., Piotto, G., et al. 2009, The Astrophysical Journal, 694, 1498

Matteucci, F. 2012, 19–52. https://www.springer.com/us/book/9783642224904#

Mészáros, S., Allende Prieto, C., Edvardsson, B., et al. 2012, AJ, 144, 120, doi: 10.1088/0004-6256/144/4/120

Mitschang, A. W., De Silva, G., Sharma, S., & Zucker, D. B. 2013, MNRAS, 428, 2321, doi: 10.1093/mnras/sts194

Morales-Luis, A. B., Sánchez Almeida, J., Aguerri, J. A. L., & Muñoz-Tuñón, C. 2011, ApJ, 743, 77, doi: 10.1088/0004-637X/743/1/77

Morgan, W. W., Keenan, P. C., & Kellman, E. 1943, An atlas of stellar spectra, with an outline of spectral classification (Chicago, Ill., The University of Chicago press [1943])

Navarro, S. G., Corradi, R. L. M., & Mampaso, A. 2012, Astronomy & Astrophysics, 538, A76, doi: 10.1051/0004-6361/201016422

Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16, doi: 10.1088/0004-637X/808/1/16

Ness, M., Rix, H.-W., Hogg, D. W., et al. 2018, The Astrophysical Journal, 853, 198, doi: 10.3847/1538-4357/aa9d8e

Ng, A. Y., Jordan, M. I., & Weiss, Y. 2002, Advances in Neural Information Processing Systems 14, 849, doi: 10.1.1.19.8100

Nidever, D. L., Holtzman, J. A., Allende Prieto, C., et al. 2015, AJ, 150, 173, doi: 10.1088/0004-6256/150/6/173

Paust, N. E. Q., Chaboyer, B., & Sarajedini, A. 2007, The Astronomical Journal, 133, 2787

Payne, C. H. 1925, PhD thesis, RADCLIFFE COLLEGE.

Reeves, H., Fowler, W., & Hoyle, F. 1970, Nature, 226, 727

Rodríguez, A., Arcay, B., Dafonte, C., Manteiga, M., & Carricajo, I. 2004, Expert Systems with Applications, doi: 10.1016/j.eswa.2004.01.007

Rosenberg, A., & Hirschberg, J. 2007, in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)

Rousseeuw, P. J., & Kaufman, L. 1990, Finding Groups in Data (Wiley Online Library)

Sakari, C. M., Shetrone, M. D., Schiavon, R. P., et al. 2016, ApJ, 829, 116, doi: 10.3847/0004-637X/829/2/116

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. 2010, ApJ, 714, 487, doi: 10.1088/0004-637X/714/1/487

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & Vazdekis, A. 2009, ApJ, 698, 1497, doi: 10.1088/0004-637X/698/2/1497

Sánchez Almeida, J., & Allende Prieto, C. 2013, ApJ, 763, 50, doi: 10.1088/0004-637X/763/1/50

Sánchez Almeida, J., Pérez-Montero, E., Morales-Luis, A. B., et al. 2016, ApJ, 819, 110, doi: 10.3847/0004-637X/819/2/110

Schiavon, R. P., Zamora, O., Carrera, R., et al. 2017, MNRAS, 465, 501, doi: 10.1093/mnras/stw2162

Schwarz, G., et al. 1978, The annals of statistics, 6, 461

SDSS Collaboration, Albareti, F. D., Allende Prieto, C., et al. 2016, ArXiv e-prints. https://arxiv.org/abs/1608.02013

Shetrone, M., Bizyaev, D., Lawler, J. E., et al. 2015, ApJS, 221, 24, doi: 10.1088/0067-0049/221/2/24

Singh, H. P., Gulati, R. K., & Gupta, R. 1998, Monthly Notices of the Royal Astronomical Society, 295, 312

Smirnov, N. V. 1939, Bull. Math. Univ. Moscou, 2, 3

Smith, C. R., Erickson, G., & Neudorfer, P. O. 2013, Maximum Entropy and Bayesian Methods: Seattle, 1991, Vol. 50 (Springer Science & Business Media)

Souto, D., Cunha, K., Smith, V., et al. 2016, ApJ, 830, 35, doi: 10.3847/0004-637X/830/1/35

Souto, D., Cunha, K., Smith, V. V., et al. 2018, ApJ, 857, 14, doi: 10.3847/1538-4357/aab612

Thirion, B., Duschenay, E., Michel, V., et al. 2016, scikitlearn

Tibshirani, R., Walther, G., & Hastie, T. 2001, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 411

Ting, Y. S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, Monthly Notices of the Royal Astronomical Society, 421, 1231, doi: 10.1111/j.1365-2966.2011.20387.x

Van Der Maaten, L. 2014, Journal of machine learning research, 15, 3221

Van Der Maaten, L., & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579

van Saders, J. L., & Pinsonneault, M. H. 2013, ApJ, 776, 67, doi: 10.1088/0004-637X/776/2/67

Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9, doi: 10.1051/aas:2000332

Wilson, J. C., Hearty, F., Skrutskie, M. F., et al. 2012, in Proc. SPIE, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, 84460H

Wold, S., Esbensen, K., & Geladi, P. 1987, Tutorial n Chemometrics and Intelligent Laboratory Systems Elsevier Science Publishers B.V, 2, 37

Wolpert, D. H., & Macready, W. G. 1997, IEEE Transactions on Evolutionary Computation, 1, 67, doi: 10.1109/4235.585893

Yeremi, M., Flynn, M., Offner, S., Loeppky, J., & Rosolowsky, E. 2014, The Astrophysical Journal, 783, 93

Zamora, O., García-Hernández, D. A., Allende Prieto, C., et al. 2015, AJ, 149, 181, doi: 10.1088/0004-6256/149/6/181

Zasowski, G., Johnson, J. A., Frinchaboy, P. M., et al. 2013, AJ, 146, 81, doi: 10.1088/0004-6256/146/4/81

Zimek, A., Schubert, E., & Kriegel, H.-P. 2012, Statistical Analysis and Data Mining, 5, 363, doi: 10.1002/sam.11161

A

Appendix

A.1 Statistics to determine the number of clusters

A.1.1 Gap statistics

Let $\vec{x}$ be an star defined by N abundance measurements, $\vec{x} = \{a_1, a_2, ..., a_N\}$, and C be a cluster of n stars. If $\vec{\mu}_k$ is the mean of all stars in the cluster C_k, $\vec{\mu}_k = 1/n_k \sum_{\vec{x}_i \in C_k} \vec{x}_i$, then the mean intra cluster abundance distance among the stars in the cluster is

$$D_k = \sum_{\vec{x}_i \in C_k} \sum_{\vec{x}_j \in C_k} ||\vec{x}_i - \vec{x}_j||^2 = 2n_k \sum_{\vec{x}_i \in C_k} ||\vec{x}_i - \vec{\mu}_k||^2. \tag{A.1}$$

For a clustering run with K clusters, the compactness of the classification is defined as

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} D_k. \tag{A.2}$$

We define the W_K^* as the compactness of a classification using the same clustering algorithm applied to the classification of the actual sample, but over a but over randomly distributed objects in the same volume as the actual sample. If we compute η different random samples we can write the Gap value as

$$\mathrm{Gap}(K) = \frac{1}{\eta} \sum_{\alpha=1}^{\eta} \log W_{K\alpha}^* - \log W_K. \tag{A.3}$$

Writing the mean and the standard deviation of the compactness of the randomly distributed dataset as

$$\overline{W_K^*} = 1/\eta \sum_{\alpha=1}^{\eta} W_{K\alpha}^*, \sigma_{W_K^*} = \sqrt{\left[\frac{1}{\eta} \sum_{\alpha} (\log W_{K\alpha}^* - \overline{W_K^*})^2\right]}, \tag{A.4}$$

we define

$$s_K = \sigma_{W_K^*}\sqrt{1+1/\eta}. \tag{A.5}$$

According to the Gap statistics prescription (Tibshirani et al., 2001), the optimal K is found to be the lowest value which satisfies the relation

$$\mathrm{Gap}(K) \geq \mathrm{Gap}(K+1) - s_{K+1}. \tag{A.6}$$

A.1.2 Calinski and Harabasz index

The between-cluster sum of squares is defined as

$$B_K = \sum_{k=1}^{K} n_K ||\vec{\mu}_{overall} - \vec{\mu}_k||^2, \tag{A.7}$$

where the $\vec{\mu}_{overall}$ is the global mean of the objects in the dataset,

$$\vec{\mu}_{overall} = \frac{1}{n}\sum_{i}^{n}\vec{x_i}. \tag{A.8}$$

The compactness, also known as the within cluster sum of squares, defined in Equation A.2. With this we can write the Calinski and Harabasz (CH) index (Caliński & Harabasz, 1974),

$$CH(K) = \frac{B_K}{W_K}\left(\frac{n-K}{K-1}\right). \tag{A.9}$$

The number of cluster is found maximizing the CH index.

A.1.3 Krzanowski and Lai index

The Krzanowski and Lai (KL) index is defined as

$$KL(K) = \left|\frac{(K-1)^{2/N}W_{K-1} - K^{2/N}W_K}{(K)^{2/N}W_K - (K+1)^{2/N}W_{K+1}}\right|, \tag{A.10}$$

where N is the number of features that defines the points in the dataset. The number of cluster is found maximizing the KL index (Krzanowski & Lai, 1988).

A.1.4 Silhouette score

Let $\bar{d}_i^w$ be the mean within-cluster distance from a object $\vec{x}_i$ to all the objects in a cluster C_k,

$$\bar{d}_i^w = \frac{1}{n_k}\sum_{\vec{x}_j \in C_k} ||\vec{x}_i - \vec{x}_j||^2, \tag{A.11}$$

and $\bar{d}_i^b$ the mean between-cluster distance from the object $\vec{x}_i$ to all the objects in the closest cluster $C_{k'}$,

$$\bar{d}_i^b = \frac{1}{n_{k'}}\sum_{\vec{x}_j \in C_{k'}} ||\vec{x}_i - \vec{x}_j||^2. \tag{A.12}$$

The silhouette score for the object $\vec{x_i}$ is defined as

$$s_i = \begin{cases} 1 - \bar{d}_i^w / \bar{d}_i^b, & \text{if } \bar{d}_i^w < \bar{d}_i^b \\ 0, & \text{if } \bar{d}_i^w = \bar{d}_i^b \\ \bar{d}_i^b / \bar{d}_i^w - 1, & \text{if } \bar{d}_i^w > \bar{d}_i^b \end{cases}. \tag{A.13}$$

For a run of a clustering algorithm, the silhouette score of the classification is given by the mean silhouette score of all objects in the sample,

$$S(K) = \frac{1}{n} \sum_i^n s_i. \tag{A.14}$$

The standard use of silhouette score prescribes the best choice to the number of cluster to maximize the silhouette score of the sample (Rousseeuw & Kaufman, 1990).

A.1.5 Bayesian Information Criteria (BIC)

When we fit a Gaussian mixture model to a dataset we determine a set of K multidimensional Gaussian distributions. Each Gaussian is defined by a mean $\vec{\mu}_k$ and a covariance matrix $\vec{\Sigma}_k$. The Bayesian Information Criteria for this model is given by,

$$BIC(K) = -\sum_k^K \frac{n_k}{2} \log |\vec{\Sigma}_k| - nK \left(N + \frac{N(N+1)}{2} \right). \tag{A.15}$$

As we defined before, K is the number of clusters, n is the number of objects in the dataset, n_k is the number of objects in the cluster k and N is the number of dimensions which defines the objects in the dataset. The optimal number of clusters is found minimizing the BIC (Schwarz et al., 1978; Chen & Gopalakrishnan, 1998).

A.2 Hint to repeatability index interpretation

We define the centroid of class i as

$$\vec{\mu}_i = \frac{1}{n_i} \sum_{\iota \epsilon \omega_i} \vec{x}_\iota \,, \tag{A.16}$$

where ω_i is the set of spectra x_ι assigned to class i, and n_i is the number of spectra in the class. So the mean difference between the classes in a particular classification c compared with the chosen classification is given by

$$\vec{\sigma}_c = \sqrt{\frac{\sum_{i=0}^{49} ||\vec{\mu}_{i,c} - \vec{\mu}_{i,chosen}||^2}{50}}. \tag{A.17}$$

Therefore, when we refer to mean difference between the matching classes over the 100 classifications we mean

$$\langle \vec{\sigma}_{compare} \rangle = \frac{1}{100} \sum_{c=0}^{99} \vec{\sigma}_c \,. \tag{A.18}$$

This is the mean pixel by pixel difference between the 99 classifications as compared with the chosen one. This vector can be compared with the mean within the cluster dispersion of the chosen classification,

$$\langle \vec{\sigma}_{within} \rangle = \frac{1}{50} \sum_{i=0}^{49} \sqrt{\frac{\sum_{\iota \epsilon \omega_i} ||\vec{x_\iota} - \vec{\mu_i}||^2}{n_i}}\,, \tag{A.19}$$

giving the main difference ratio between these quantities over the 4838 pixels of the spectra:

$$\langle \sigma_{ratio} \rangle = \frac{1}{4838} \sum_{j=0}^{4837} \frac{\sigma_{j,compare}}{\sigma_{j,within}} \approx 0.064. \tag{A.20}$$

The standard deviation of 3.3% is given by the standard deviation of $\sigma_{j,compare}/\sigma_{j,within}$ over the 4838 pixels.

A.3 Classes details and online material

In Table 2.2 we presented a summary of the 32 classes containing ≈ 99 per cent of the spectra in the data set. In this table the first column is the group and the second column is a hyper-link for the appendix supplementary figures for each class. The third column gives the main stellar type found in each class. This information was inferred based only on the range of atmospheric parameters covered by each class (Payne, 1925) and should be taken just as an idea of what kind of object is dominant in each class. The fourth column gives information about the main spatial distribution of each class. It is also a simple approximation based on their distribution of galactic coordinates and [α/M]-[M/H] (see Bensby et al. 2003, 2007). Finally, the fifth column presents some extra comments about the main features of the class.

The complete information about the classification is also available as online material at http://vizier.cfa.harvard.edu/viz-bin/VizieR?-source=J/A+A/612/A98 in the form of three tables; Table A.1 presents the classification for each spectra, APOGEE ID, and class; Table A.2 gives the mean spectra for each class, in the form of normalized fluxes and wavelengths; and A.3 contains the spectral within-class standard deviation for each class, normalized fluxes, and wavelengths. In both Tables A.2 and A.3 the last column gives the mask applied to the spectra: a binary index, where zero means the wavelength was not considered during classification and one means it was included in the classification procedure.

Tables A.4, A.5, A.6 and A.7 present the median values for the atmospheric parameters and all the individual chemical elements in each class. The upper and lower limits presented in the tables, as well as those shown in the next sections, were calculated by taking the interval around the median, which encloses 68.3 percent of the points in each class.

TABLE A.1— Spectral classification. Complete table can be found in online material.

APOGEE ID	Class
2M03183846+7216305	11
2M03470204+4125397	11
2M04425018+6644089	5
2M04575928+3416050	5
2M05373344+7441194	5
⋮	⋮

TABLE A.2— Mean spectra of the 50 classes. Complete table can be found in online material.

Class 0	Class 1	⋯	Class 49	Wavelength Å	Mask
⋮	⋮	⋮	⋮	⋮	⋮
0.99505127	0.99435151	⋯	0.16018607	16178.34	1
0.98787344	0.98545332	⋯	1.02000000	16178.57	1
0.97224899	0.96870929	⋯	1.02000000	16178.79	1
0.93429393	0.92675939	⋯	0.97670869	16179.01	1
0.89091408	0.87721260	⋯	0.09924513	16179.24	1
⋮	⋮	⋮	⋮	⋮	⋮

TABLE A.3— Within-class spectral standard deviation for the 50 classes. Complete table can be found in online material.

Class 0	Class 1	⋯	Class 49	Wavelength Å	Mask
⋮	⋮	⋮	⋮	⋮	⋮
0.01448872	0.01866457	⋯	0.00000000	16178.34	1
0.00886448	0.02040061	⋯	0.00000000	16178.57	1
0.01245967	0.02015303	⋯	0.00000000	16178.79	1
0.01557365	0.01858692	⋯	0.00000000	16179.01	1
0.01527482	0.01647230	⋯	0.00000000	16179.24	1
⋮	⋮	⋮	⋮	⋮	⋮

TABLE A.4— Median atmospheric parameters and chemical abundances for the 32 most populated classes. The median values are shown with the upper and lower limits for each class. The limits are calculated by setting the boundaries which enclose 68.3 percent of the objects in the class.

Class	$T_{\rm eff}$(K)	$\log g$	[M/H]	[C/M]	[N/M]
0	4853^{144}_{159}	$+2.93^{0.34}_{0.19}$	$-0.31^{0.11}_{0.12}$	$+0.03^{0.09}_{0.08}$	$+0.04^{0.09}_{0.11}$
1	4731^{144}_{231}	$+2.81^{0.21}_{0.35}$	$-0.22^{0.11}_{0.15}$	$-0.01^{0.10}_{0.08}$	$+0.10^{0.09}_{0.08}$
2	4712^{130}_{175}	$+2.83^{0.25}_{0.23}$	$-0.07^{0.10}_{0.11}$	$-0.05^{0.10}_{0.07}$	$+0.15^{0.08}_{0.08}$
3	7814^{171}_{617}	$+4.76^{0.22}_{0.35}$	$-2.03^{0.73}_{0.35}$	$-0.18^{0.44}_{0.46}$	$+0.18^{0.45}_{0.59}$
4	4679^{137}_{145}	$+2.84^{0.34}_{0.21}$	$+0.07^{0.10}_{0.11}$	$-0.04^{0.09}_{0.08}$	$+0.19^{0.09}_{0.09}$
5	4941^{590}_{203}	$+3.16^{1.04}_{0.38}$	$-0.45^{0.28}_{0.19}$	$+0.08^{0.11}_{0.12}$	$+0.00^{0.19}_{0.19}$
6	4589^{136}_{158}	$+2.76^{0.28}_{0.27}$	$+0.17^{0.10}_{0.12}$	$-0.03^{0.07}_{0.08}$	$+0.24^{0.09}_{0.08}$
7	4236^{97}_{101}	$+2.03^{0.23}_{0.22}$	$-0.29^{0.14}_{0.15}$	$-0.02^{0.10}_{0.08}$	$+0.17^{0.07}_{0.09}$
8	4919^{270}_{111}	$+3.45^{0.42}_{0.22}$	$-0.02^{0.15}_{0.11}$	$+0.02^{0.08}_{0.09}$	$+0.03^{0.11}_{0.14}$
9	4495^{108}_{145}	$+2.70^{0.19}_{0.28}$	$+0.30^{0.09}_{0.12}$	$-0.00^{0.05}_{0.08}$	$+0.30^{0.10}_{0.08}$
10	4791^{359}_{242}	$+4.27^{0.14}_{0.18}$	$-0.15^{0.14}_{0.17}$	$-0.08^{0.11}_{0.15}$	$+0.04^{0.22}_{0.16}$
11	6125^{478}_{698}	$+4.47^{0.50}_{0.52}$	$-0.24^{0.27}_{0.37}$	$+0.06^{0.17}_{0.18}$	$+0.26^{0.54}_{0.55}$
12	4761^{241}_{166}	$+4.35^{0.08}_{0.12}$	$+0.13^{0.13}_{0.11}$	$-0.05^{0.07}_{0.14}$	$+0.06^{0.13}_{0.11}$
13	6396^{482}_{1681}	$+4.35^{0.61}_{2.10}$	$-0.88^{0.57}_{1.06}$	$+0.07^{0.33}_{0.33}$	$+0.33^{0.48}_{0.62}$
14	4095^{98}_{107}	$+1.85^{0.24}_{0.24}$	$-0.09^{0.13}_{0.14}$	$-0.03^{0.09}_{0.08}$	$+0.19^{0.06}_{0.07}$
15	4065^{241}_{159}	$+4.25^{0.08}_{0.13}$	$-0.18^{0.13}_{0.15}$	$+0.02^{0.03}_{0.18}$	$-0.30^{0.37}_{0.31}$
16	3620^{104}_{101}	$+1.30^{0.24}_{0.26}$	$+0.17^{0.13}_{0.15}$	$+0.04^{0.02}_{0.03}$	$+0.14^{0.08}_{0.07}$
17	4361^{104}_{221}	$+4.26^{0.08}_{0.11}$	$+0.01^{0.13}_{0.09}$	$-0.01^{0.04}_{0.17}$	$+0.01^{0.13}_{0.34}$
18	3868^{95}_{88}	$+1.65^{0.22}_{0.25}$	$+0.18^{0.15}_{0.18}$	$+0.02^{0.03}_{0.05}$	$+0.21^{0.08}_{0.08}$
19	3753^{95}_{95}	$+1.12^{0.26}_{0.33}$	$-0.34^{0.17}_{0.24}$	$-0.02^{0.13}_{0.08}$	$+0.12^{0.09}_{0.11}$
20	3724^{131}_{128}	$+4.23^{0.16}_{0.37}$	$-0.22^{0.19}_{0.16}$	$+0.02^{0.03}_{0.24}$	$-0.24^{0.45}_{0.54}$
21	3500^{1}_{0}	$+0.58^{0.32}_{0.35}$	$-0.21^{0.20}_{0.27}$	$+0.19^{0.04}_{0.16}$	$+0.25^{0.07}_{0.08}$
22	4137^{114}_{102}	$+2.09^{0.27}_{0.24}$	$+0.23^{0.13}_{0.14}$	$+0.01^{0.04}_{0.07}$	$+0.27^{0.10}_{0.08}$
23	3503^{108}_{3}	$+4.15^{0.56}_{0.40}$	$-0.54^{0.24}_{0.55}$	$+0.07^{0.26}_{0.23}$	$+0.28^{0.60}_{0.65}$
24	4582^{330}_{322}	$+2.22^{0.54}_{0.60}$	$-1.20^{0.25}_{0.22}$	$-0.08^{0.32}_{0.25}$	$+0.21^{0.45}_{0.24}$
25	3561^{83}_{59}	$+0.79^{0.25}_{0.38}$	$-0.31^{0.17}_{0.29}$	$+0.01^{0.16}_{0.06}$	$+0.11^{0.08}_{0.09}$
26	3899^{77}_{103}	$+1.37^{0.21}_{0.33}$	$-0.47^{0.14}_{0.20}$	$-0.03^{0.12}_{0.10}$	$+0.11^{0.11}_{0.10}$
27	4243^{641}_{282}	$+4.83^{0.17}_{0.84}$	$-0.21^{0.21}_{0.20}$	$+0.03^{0.04}_{0.24}$	$-0.44^{0.74}_{0.50}$
28	4027^{119}_{165}	$+1.47^{0.35}_{0.66}$	$-0.81^{0.19}_{0.33}$	$-0.09^{0.15}_{0.30}$	$+0.09^{0.18}_{0.10}$
29	3973^{244}_{361}	$+4.97^{0.03}_{0.79}$	$-0.40^{0.20}_{0.48}$	$+0.05^{0.25}_{0.25}$	$-0.35^{0.79}_{0.59}$
30	4202^{1744}_{659}	$+3.14^{1.16}_{1.05}$	$-0.54^{0.83}_{0.84}$	$+0.24^{0.57}_{0.43}$	$+0.09^{0.78}_{0.56}$
31	4458^{386}_{528}	$+2.17^{0.83}_{0.73}$	$-1.34^{0.47}_{0.58}$	$+0.04^{0.45}_{0.31}$	$+0.18^{0.53}_{0.32}$

TABLE A.5— Median chemical abundances for the 32 most populated classes. The median values are shown with the upper and lower limits for each class. The limits are calculated by setting the boundaries which enclose 68.3 percent of the objects in the class.

Class	[α/M]	[Al/H]	[Ca/H]	[C/H]	[Fe/H]	[K/H]
0	$+0.08^{0.13}_{0.04}$	$+0.02^{0.10}_{0.09}$	$+0.04^{0.09}_{0.12}$	$+0.14^{0.11}_{0.07}$	$-0.29^{0.22}_{0.23}$	$+0.04^{0.14}_{0.06}$
1	$+0.07^{0.12}_{0.04}$	$-0.01^{0.10}_{0.08}$	$+0.10^{0.09}_{0.09}$	$+0.12^{0.10}_{0.06}$	$-0.23^{0.18}_{0.21}$	$+0.01^{0.13}_{0.06}$
2	$+0.05^{0.05}_{0.03}$	$-0.05^{0.10}_{0.08}$	$+0.15^{0.08}_{0.08}$	$+0.07^{0.07}_{0.04}$	$-0.09^{0.16}_{0.16}$	$-0.02^{0.08}_{0.04}$
3	$-0.26^{0.34}_{0.41}$	$-1.00^{1.73}_{0.00}$	$+0.42^{0.58}_{1.17}$	$+0.31^{0.65}_{0.84}$	$+0.12^{0.38}_{1.88}$	$+0.77^{0.23}_{1.26}$
4	$+0.04^{0.04}_{0.03}$	$-0.04^{0.09}_{0.09}$	$+0.19^{0.09}_{0.10}$	$+0.05^{0.06}_{0.04}$	$+0.07^{0.17}_{0.15}$	$-0.02^{0.07}_{0.04}$
5	$+0.11^{0.14}_{0.09}$	$+0.07^{0.13}_{0.12}$	$-0.01^{0.20}_{0.18}$	$+0.18^{0.14}_{0.14}$	$-0.35^{0.37}_{0.44}$	$+0.12^{0.13}_{0.11}$
6	$+0.04^{0.03}_{0.03}$	$-0.03^{0.08}_{0.08}$	$+0.24^{0.09}_{0.09}$	$+0.04^{0.05}_{0.04}$	$+0.19^{0.19}_{0.16}$	$-0.01^{0.06}_{0.04}$
7	$+0.09^{0.12}_{0.04}$	$-0.02^{0.11}_{0.09}$	$+0.17^{0.07}_{0.10}$	$+0.12^{0.12}_{0.05}$	$-0.34^{0.17}_{0.21}$	$+0.05^{0.13}_{0.05}$
8	$+0.05^{0.07}_{0.03}$	$+0.02^{0.09}_{0.09}$	$+0.01^{0.12}_{0.14}$	$+0.08^{0.08}_{0.05}$	$+0.03^{0.20}_{0.15}$	$+0.03^{0.10}_{0.05}$
9	$+0.04^{0.03}_{0.03}$	$-0.00^{0.05}_{0.08}$	$+0.29^{0.10}_{0.08}$	$+0.03^{0.04}_{0.04}$	$+0.39^{0.11}_{0.18}$	$+0.01^{0.06}_{0.04}$
10	$+0.04^{0.08}_{0.04}$	$-0.07^{0.11}_{0.14}$	$+0.01^{0.20}_{0.16}$	$+0.07^{0.10}_{0.08}$	$-0.08^{0.16}_{0.19}$	$+0.10^{0.11}_{0.07}$
11	$+0.01^{0.12}_{0.05}$	$-0.04^{0.31}_{0.96}$	$+0.74^{0.26}_{0.77}$	$+0.05^{0.12}_{0.07}$	$-0.20^{0.27}_{0.73}$	$+0.06^{0.15}_{0.07}$
12	$+0.00^{0.03}_{0.02}$	$-0.04^{0.07}_{0.12}$	$+0.04^{0.12}_{0.12}$	$+0.03^{0.06}_{0.04}$	$+0.19^{0.22}_{0.16}$	$+0.07^{0.05}_{0.04}$
13	$+0.01^{0.26}_{0.25}$	$-0.93^{1.51}_{0.07}$	$+1.00^{0.00}_{0.63}$	$+0.08^{0.32}_{0.18}$	$-1.26^{1.10}_{1.24}$	$+0.22^{0.25}_{0.21}$
14	$+0.06^{0.10}_{0.04}$	$-0.03^{0.09}_{0.09}$	$+0.19^{0.06}_{0.07}$	$+0.08^{0.10}_{0.05}$	$-0.11^{0.15}_{0.17}$	$+0.06^{0.10}_{0.05}$
15	$-0.05^{0.06}_{0.04}$	$+0.01^{0.03}_{0.14}$	$-0.24^{0.31}_{0.30}$	$-0.05^{0.08}_{0.05}$	$-0.05^{0.17}_{0.18}$	$+0.02^{0.16}_{0.35}$
16	$+0.00^{0.03}_{0.02}$	$+0.00^{0.03}_{0.03}$	$+0.16^{0.08}_{0.07}$	$-0.01^{0.04}_{0.02}$	$+0.41^{0.09}_{0.25}$	$+0.01^{0.12}_{0.04}$
17	$-0.02^{0.04}_{0.04}$	$-0.02^{0.05}_{0.14}$	$-0.01^{0.13}_{0.30}$	$-0.02^{0.06}_{0.05}$	$+0.14^{0.24}_{0.15}$	$+0.10^{0.06}_{0.05}$
18	$+0.02^{0.03}_{0.02}$	$+0.00^{0.03}_{0.05}$	$+0.21^{0.08}_{0.08}$	$+0.01^{0.04}_{0.03}$	$+0.32^{0.18}_{0.27}$	$+0.05^{0.09}_{0.06}$
19	$+0.08^{0.15}_{0.05}$	$-0.04^{0.13}_{0.09}$	$+0.14^{0.09}_{0.11}$	$+0.10^{0.15}_{0.06}$	$-0.35^{0.21}_{0.33}$	$+0.16^{0.09}_{0.11}$
20	$-0.07^{0.04}_{0.21}$	$+0.02^{0.04}_{0.21}$	$-0.14^{0.49}_{0.49}$	$-0.07^{0.05}_{0.21}$	$-0.06^{0.19}_{0.20}$	$-0.27^{0.09}_{0.19}$
21	$+0.27^{0.06}_{0.18}$	$+0.09^{0.06}_{0.11}$	$+0.29^{0.07}_{0.08}$	$+0.31^{0.04}_{0.22}$	$+0.22^{0.27}_{0.30}$	$+0.13^{0.07}_{0.13}$
22	$+0.03^{0.04}_{0.03}$	$-0.00^{0.04}_{0.08}$	$+0.26^{0.10}_{0.08}$	$+0.03^{0.04}_{0.04}$	$+0.33^{0.17}_{0.22}$	$+0.04^{0.07}_{0.05}$
23	$-0.08^{0.23}_{0.22}$	$+0.07^{0.27}_{0.06}$	$+0.26^{0.57}_{0.64}$	$-0.05^{0.25}_{0.08}$	$-0.14^{0.23}_{0.25}$	$-0.23^{0.24}_{0.11}$
24	$+0.24^{0.07}_{0.07}$	$-0.14^{0.36}_{0.25}$	$+0.26^{0.41}_{0.29}$	$+0.30^{0.09}_{0.09}$	$-1.47^{0.87}_{1.02}$	$+0.14^{0.13}_{0.12}$
25	$+0.08^{0.18}_{0.05}$	$-0.02^{0.14}_{0.06}$	$+0.16^{0.09}_{0.09}$	$+0.09^{0.22}_{0.06}$	$-0.24^{0.19}_{0.30}$	$+0.17^{0.08}_{0.12}$
26	$+0.14^{0.11}_{0.09}$	$-0.03^{0.12}_{0.11}$	$+0.12^{0.11}_{0.11}$	$+0.18^{0.12}_{0.10}$	$-0.55^{0.19}_{0.38}$	$+0.18^{0.07}_{0.14}$
27	$-0.14^{0.11}_{0.20}$	$+0.02^{0.05}_{0.22}$	$-0.48^{0.67}_{0.47}$	$-0.18^{0.14}_{0.20}$	$-0.19^{0.26}_{0.50}$	$-0.27^{0.42}_{0.26}$
28	$+0.24^{0.04}_{0.11}$	$-0.10^{0.16}_{0.31}$	$+0.10^{0.21}_{0.11}$	$+0.27^{0.06}_{0.10}$	$-1.14^{0.42}_{0.73}$	$+0.18^{0.06}_{0.15}$
29	$-0.16^{0.19}_{0.24}$	$+0.05^{0.25}_{0.25}$	$-0.36^{0.76}_{0.51}$	$-0.19^{0.23}_{0.23}$	$-0.46^{0.46}_{0.80}$	$-0.32^{0.22}_{0.28}$
30	$+0.02^{0.34}_{0.43}$	$+0.20^{0.59}_{0.34}$	$+0.11^{0.76}_{0.59}$	$+0.06^{0.44}_{0.45}$	$-0.23^{0.73}_{1.07}$	$+0.01^{0.54}_{0.51}$
31	$+0.20^{0.11}_{0.23}$	$-0.04^{0.42}_{0.32}$	$+0.19^{0.49}_{0.35}$	$+0.29^{0.19}_{0.27}$	$-1.10^{0.50}_{1.10}$	$+0.09^{0.23}_{0.34}$

TABLE A.6— Median chemical abundances for the 32 most populated classes. The median values are shown with the upper and lower limits for each class. The limits are calculated by setting the boundaries which enclose 68.3 percent of the objects in the class.

Class	[Mg/H]	[Mn/H]	[Na/H]	[Ni/H]	[N/H]	[O/H]
0	$-0.11^{0.14}_{0.13}$	$+0.20^{0.10}_{0.09}$	$+0.20^{0.11}_{0.09}$	$-0.39^{0.11}_{0.13}$	$+0.03^{0.13}_{0.06}$	$-0.07^{0.11}_{0.09}$
1	$-0.04^{0.13}_{0.15}$	$+0.21^{0.09}_{0.08}$	$+0.19^{0.10}_{0.08}$	$-0.31^{0.09}_{0.15}$	$-0.00^{0.11}_{0.05}$	$-0.06^{0.10}_{0.07}$
2	$+0.07^{0.12}_{0.12}$	$+0.18^{0.07}_{0.07}$	$+0.16^{0.07}_{0.07}$	$-0.19^{0.09}_{0.10}$	$-0.03^{0.08}_{0.04}$	$-0.06^{0.08}_{0.07}$
3	$-2.16^{0.64}_{0.34}$	$+0.26^{0.37}_{0.51}$	$+0.46^{0.42}_{0.49}$	$-2.50^{2.02}_{0.00}$	$-0.31^{1.31}_{0.69}$	$-0.35^{1.35}_{0.65}$
4	$+0.19^{0.14}_{0.13}$	$+0.16^{0.07}_{0.07}$	$+0.13^{0.06}_{0.07}$	$-0.08^{0.10}_{0.10}$	$-0.05^{0.07}_{0.03}$	$-0.02^{0.08}_{0.06}$
5	$-0.22^{0.22}_{0.18}$	$+0.20^{0.14}_{0.16}$	$+0.26^{0.14}_{0.15}$	$-0.49^{0.22}_{0.18}$	$+0.08^{0.13}_{0.10}$	$-0.02^{0.14}_{0.16}$
6	$+0.28^{0.14}_{0.14}$	$+0.16^{0.07}_{0.07}$	$+0.12^{0.06}_{0.06}$	$+0.01^{0.11}_{0.12}$	$-0.07^{0.06}_{0.03}$	$+0.02^{0.08}_{0.06}$
7	$-0.12^{0.15}_{0.17}$	$+0.23^{0.09}_{0.07}$	$+0.16^{0.12}_{0.06}$	$-0.37^{0.14}_{0.17}$	$-0.02^{0.10}_{0.04}$	$+0.02^{0.11}_{0.07}$
8	$+0.13^{0.19}_{0.14}$	$+0.13^{0.07}_{0.07}$	$+0.13^{0.08}_{0.09}$	$-0.16^{0.12}_{0.10}$	$-0.01^{0.15}_{0.05}$	$-0.07^{0.11}_{0.10}$
9	$+0.41^{0.09}_{0.15}$	$+0.16^{0.07}_{0.07}$	$+0.11^{0.05}_{0.06}$	$+0.13^{0.11}_{0.14}$	$-0.09^{0.06}_{0.04}$	$+0.08^{0.07}_{0.07}$
10	$-0.04^{0.22}_{0.19}$	$+0.06^{0.09}_{0.09}$	$+0.11^{0.13}_{0.10}$	$-0.28^{0.12}_{0.14}$	$+0.01^{0.12}_{0.06}$	$+0.01^{0.12}_{0.16}$
11	$-0.37^{0.24}_{0.22}$	$+0.00^{0.15}_{0.14}$	$+0.20^{0.19}_{0.19}$	$-0.37^{0.23}_{0.29}$	$+0.03^{0.11}_{0.07}$	$+0.13^{0.32}_{0.31}$
12	$+0.25^{0.22}_{0.19}$	$+0.04^{0.06}_{0.06}$	$+0.05^{0.08}_{0.06}$	$-0.04^{0.15}_{0.12}$	$-0.02^{0.09}_{0.04}$	$+0.07^{0.11}_{0.13}$
13	$-1.19^{0.48}_{0.81}$	$-0.10^{0.44}_{0.30}$	$+0.19^{0.42}_{0.28}$	$-1.22^{0.74}_{1.23}$	$+0.05^{0.37}_{0.42}$	$+0.55^{0.45}_{0.67}$
14	$+0.02^{0.15}_{0.18}$	$+0.18^{0.07}_{0.07}$	$+0.10^{0.10}_{0.05}$	$-0.20^{0.11}_{0.15}$	$-0.06^{0.06}_{0.04}$	$+0.08^{0.10}_{0.07}$
15	$-0.24^{0.17}_{0.17}$	$-0.02^{0.08}_{0.04}$	$-0.04^{0.11}_{0.05}$	$-0.29^{0.12}_{0.13}$	$-0.02^{0.07}_{0.07}$	$-0.18^{0.09}_{0.06}$
16	$+0.18^{0.20}_{0.19}$	$+0.00^{0.11}_{0.04}$	$-0.02^{0.05}_{0.04}$	$+0.11^{0.17}_{0.20}$	$-0.24^{0.10}_{0.03}$	$-0.12^{0.20}_{0.05}$
17	$+0.04^{0.24}_{0.16}$	$-0.04^{0.06}_{0.04}$	$-0.01^{0.08}_{0.05}$	$-0.12^{0.14}_{0.10}$	$-0.01^{0.07}_{0.04}$	$-0.09^{0.17}_{0.11}$
18	$+0.25^{0.19}_{0.21}$	$+0.11^{0.09}_{0.09}$	$+0.02^{0.05}_{0.05}$	$+0.03^{0.18}_{0.18}$	$-0.16^{0.07}_{0.06}$	$+0.00^{0.19}_{0.13}$
19	$-0.27^{0.23}_{0.29}$	$+0.18^{0.10}_{0.10}$	$+0.08^{0.15}_{0.05}$	$-0.40^{0.19}_{0.25}$	$-0.03^{0.09}_{0.06}$	$+0.14^{0.11}_{0.18}$
20	$-0.39^{0.19}_{0.23}$	$-0.05^{0.08}_{0.11}$	$-0.05^{0.09}_{0.18}$	$-0.31^{0.15}_{0.18}$	$-0.06^{0.08}_{0.22}$	$-0.19^{0.04}_{0.20}$
21	$-0.07^{0.24}_{0.30}$	$-0.02^{0.15}_{0.04}$	$+0.24^{0.10}_{0.18}$	$-0.14^{0.26}_{0.35}$	$-0.04^{0.11}_{0.06}$	$+0.23^{0.18}_{0.18}$
22	$+0.32^{0.16}_{0.18}$	$+0.16^{0.08}_{0.07}$	$+0.07^{0.06}_{0.06}$	$+0.06^{0.15}_{0.16}$	$-0.12^{0.06}_{0.04}$	$+0.10^{0.09}_{0.09}$
23	$-0.84^{0.33}_{0.44}$	$-0.04^{0.24}_{0.21}$	$+0.02^{0.37}_{0.10}$	$-0.42^{0.24}_{0.27}$	$-0.01^{0.35}_{0.15}$	$-0.18^{0.28}_{0.11}$
24	$-1.19^{0.55}_{0.39}$	$+0.35^{0.09}_{0.09}$	$+0.41^{0.14}_{0.14}$	$-1.19^{0.31}_{0.39}$	$+0.17^{0.13}_{0.11}$	$+0.00^{0.14}_{0.12}$
25	$-0.26^{0.19}_{0.28}$	$+0.09^{0.13}_{0.08}$	$+0.07^{0.20}_{0.05}$	$-0.33^{0.17}_{0.27}$	$-0.06^{0.11}_{0.09}$	$+0.07^{0.17}_{0.18}$
26	$-0.34^{0.20}_{0.28}$	$+0.27^{0.10}_{0.11}$	$+0.16^{0.13}_{0.10}$	$-0.53^{0.17}_{0.21}$	$+0.02^{0.10}_{0.08}$	$+0.13^{0.11}_{0.13}$
27	$-0.76^{0.38}_{0.42}$	$-0.11^{0.12}_{0.17}$	$-0.11^{0.28}_{0.20}$	$-0.49^{0.17}_{0.25}$	$-0.18^{0.20}_{0.16}$	$-0.30^{0.19}_{0.21}$
28	$-0.69^{0.33}_{0.59}$	$+0.38^{0.09}_{0.10}$	$+0.28^{0.08}_{0.13}$	$-0.90^{0.28}_{0.34}$	$+0.12^{0.07}_{0.08}$	$+0.02^{0.10}_{0.11}$
29	$-1.27^{0.56}_{0.35}$	$-0.15^{0.21}_{0.22}$	$-0.13^{0.27}_{0.24}$	$-0.72^{0.25}_{0.44}$	$-0.20^{0.30}_{0.19}$	$-0.29^{0.21}_{0.35}$
30	$-0.63^{1.13}_{1.21}$	$+0.04^{0.37}_{0.58}$	$+0.14^{0.61}_{0.44}$	$-0.74^{0.78}_{0.90}$	$+0.11^{0.70}_{0.45}$	$-0.00^{0.71}_{0.42}$
31	$-1.27^{0.51}_{0.82}$	$+0.35^{0.12}_{0.36}$	$+0.31^{0.20}_{0.28}$	$-1.37^{0.45}_{0.77}$	$+0.16^{0.19}_{0.18}$	$-0.00^{0.27}_{0.25}$

TABLE A.7— Median chemical abundances for the 32 most populated classes. The median values are shown with the upper and lower limits for each class. The limits are calculated by setting the boundaries which enclose 68.3 percent of the objects in the class.

Class	[Si/H]	[S/H]	[Ti/H]	[V/H]	$N_\star$
0	$-0.57^{0.42}_{1.12}$	$-0.35^{0.13}_{0.16}$	$-0.30^{0.12}_{0.13}$	$-0.37^{0.11}_{0.13}$	15066
1	$-0.43^{0.31}_{0.47}$	$-0.24^{0.13}_{0.17}$	$-0.21^{0.11}_{0.15}$	$-0.26^{0.10}_{0.15}$	14177
2	$-0.22^{0.20}_{0.30}$	$-0.05^{0.12}_{0.13}$	$-0.06^{0.10}_{0.11}$	$-0.11^{0.10}_{0.11}$	12482
3	$+0.14^{0.36}_{0.71}$	$-0.60^{0.43}_{1.03}$	$-1.21^{0.34}_{0.38}$	$-2.50^{1.30}_{0.00}$	10628
4	$-0.06^{0.19}_{0.20}$	$+0.15^{0.12}_{0.14}$	$+0.08^{0.10}_{0.11}$	$+0.04^{0.11}_{0.12}$	10253
5	$-0.87^{0.69}_{1.63}$	$-0.53^{0.32}_{0.24}$	$-0.45^{0.29}_{0.20}$	$-0.52^{0.28}_{0.19}$	9144
6	$+0.06^{0.21}_{0.18}$	$+0.30^{0.13}_{0.15}$	$+0.17^{0.11}_{0.12}$	$+0.15^{0.12}_{0.13}$	8096
7	$-0.43^{0.16}_{0.18}$	$-0.26^{0.17}_{0.18}$	$-0.29^{0.14}_{0.15}$	$-0.31^{0.14}_{0.14}$	5325
8	$-0.13^{0.24}_{0.53}$	$-0.02^{0.17}_{0.14}$	$-0.02^{0.15}_{0.11}$	$-0.08^{0.16}_{0.11}$	5054
9	$+0.26^{0.18}_{0.19}$	$+0.49^{0.01}_{0.16}$	$+0.30^{0.09}_{0.12}$	$+0.31^{0.11}_{0.14}$	4820
10	$-0.31^{0.31}_{0.50}$	$-0.21^{0.17}_{0.20}$	$-0.16^{0.16}_{0.18}$	$-0.24^{0.15}_{0.19}$	4771
11	$-0.42^{0.33}_{1.99}$	$-0.28^{0.28}_{0.40}$	$-0.28^{0.25}_{0.36}$	$-0.31^{0.31}_{0.38}$	4271
12	$+0.03^{0.26}_{0.27}$	$+0.14^{0.18}_{0.15}$	$+0.13^{0.14}_{0.12}$	$+0.05^{0.17}_{0.14}$	4051
13	$-1.65^{1.50}_{0.85}$	$-0.53^{0.38}_{1.29}$	$-0.85^{0.45}_{1.03}$	$-0.85^{0.54}_{1.11}$	3696
14	$-0.18^{0.15}_{0.15}$	$+0.03^{0.17}_{0.18}$	$-0.10^{0.13}_{0.14}$	$-0.10^{0.13}_{0.14}$	3545
15	$-0.55^{0.23}_{0.36}$	$-0.19^{0.18}_{0.19}$	$-0.20^{0.13}_{0.15}$	$-0.28^{0.15}_{0.18}$	3450
16	$+0.28^{0.22}_{0.23}$	$+0.50^{0.00}_{0.18}$	$+0.13^{0.13}_{0.14}$	$+0.29^{0.15}_{0.19}$	3157
17	$-0.24^{0.25}_{0.26}$	$+0.01^{0.18}_{0.14}$	$-0.02^{0.13}_{0.09}$	$-0.11^{0.15}_{0.12}$	3109
18	$+0.20^{0.24}_{0.24}$	$+0.46^{0.04}_{0.25}$	$+0.16^{0.15}_{0.18}$	$+0.22^{0.17}_{0.20}$	2955
19	$-0.35^{0.19}_{0.25}$	$-0.24^{0.24}_{0.30}$	$-0.35^{0.16}_{0.24}$	$-0.33^{0.18}_{0.26}$	2874
20	$-0.62^{0.26}_{0.37}$	$-0.19^{0.22}_{0.22}$	$-0.26^{0.19}_{0.17}$	$-0.30^{0.21}_{0.21}$	2771
21	$+0.29^{0.21}_{0.25}$	$+0.02^{0.31}_{0.39}$	$-0.25^{0.20}_{0.27}$	$+0.03^{0.25}_{0.25}$	2556
22	$+0.21^{0.20}_{0.20}$	$+0.47^{0.03}_{0.18}$	$+0.22^{0.13}_{0.14}$	$+0.25^{0.15}_{0.16}$	2494
23	$-0.76^{0.32}_{0.48}$	$-0.38^{0.42}_{0.65}$	$-0.81^{0.42}_{0.78}$	$-0.67^{0.34}_{0.60}$	2425
24	$-1.90^{0.55}_{0.60}$	$-1.40^{0.30}_{0.23}$	$-1.21^{0.24}_{0.22}$	$-1.30^{0.24}_{0.26}$	2388
25	$-0.20^{0.17}_{0.26}$	$-0.15^{0.26}_{0.36}$	$-0.34^{0.17}_{0.27}$	$-0.25^{0.18}_{0.29}$	2288
26	$-0.54^{0.18}_{0.24}$	$-0.44^{0.18}_{0.27}$	$-0.47^{0.14}_{0.20}$	$-0.48^{0.15}_{0.22}$	2252
27	$-0.82^{0.78}_{1.32}$	$-0.24^{0.22}_{0.22}$	$-0.35^{0.19}_{0.22}$	$-0.31^{0.22}_{0.23}$	1431
28	$-1.00^{0.28}_{0.43}$	$-0.93^{0.24}_{0.38}$	$-0.81^{0.19}_{0.33}$	$-0.87^{0.24}_{0.36}$	1117
29	$-1.12^{0.70}_{1.37}$	$-0.31^{0.25}_{0.50}$	$-0.61^{0.24}_{0.72}$	$-0.49^{0.31}_{0.45}$	1081
30	$-0.99^{1.23}_{1.51}$	$-0.55^{0.79}_{0.81}$	$-0.63^{0.72}_{0.76}$	$-0.68^{0.77}_{0.98}$	562
31	$-1.90^{1.03}_{0.60}$	$-1.37^{0.50}_{0.49}$	$-1.36^{0.46}_{0.50}$	$-1.51^{0.54}_{0.47}$	474

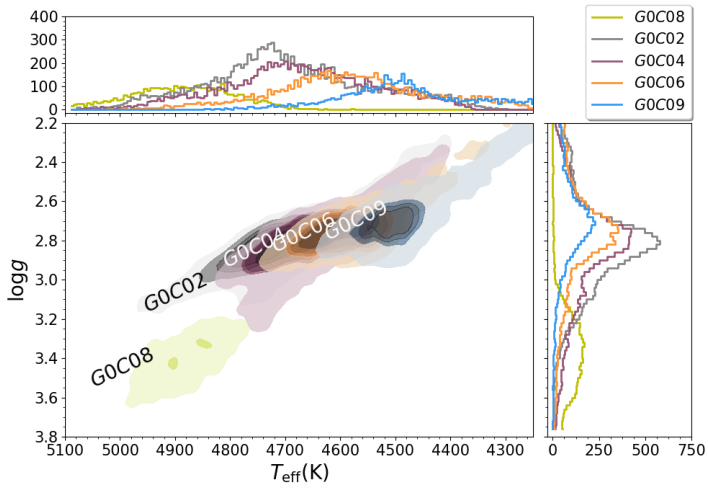

FIGURE A.1— $T_{\rm eff} - \log g$ distribution for classes in group 0. The same rules and color from Figure 2.7 were applied to the contours here. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. The histogram line color matches the color of the contours.

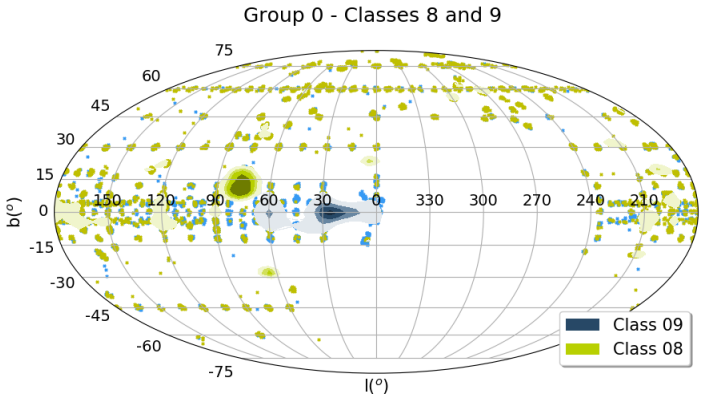

FIGURE A.2— Mollweide's projection of the Galactic coordinates distribution of the stars in classes 8 and 9, which belong to group 0. Yellow and dark blue contours enclose 68.3 percent of the stars in classes 8 and 9, respectively. Yellow symbols represent stars in class 8 and blue symbols represent stars in class 9 out of the regions containing 68.3 percent of the points. The contour shades follow the same rule as in Figure 2.7.

A.3.1 Group 0 (Classes 2, 4, 6, 8 and 9) - Metal-rich RC/warm RGB

From the distribution of $\log g$ and $T_{\rm eff}$ values in Figure A.1, one can spot this group among the Red Clump (RC) stars and at the warmest end of the red giant branch (RGB) (Binney & Merrifield, 1998). Comparing these classes with Bovy et al. 2014's catalog of red clump stars, we found that 31, 26, 26, 1, and 21 percent of the stars in classes 2, 4, 6, 8, and 9, respectively, belong to the red clump. The classes increase in metallicity in the sense $-0.07\pm^{0.10}_{0.11} = \widehat{[\rm M/H]}_{c2} < \widehat{[\rm M/H]}_{c8} < \widehat{[\rm M/H]}_{c4} < \widehat{[\rm M/H]}_{c6} < \widehat{[\rm M/H]}_{c9} = 0.30\pm^{0.09}_{0.12}$. As metallicity increases, the position of the RC moves towards cooler regions in the plane $T_{\rm eff}$- $\log g$, as shown in Figure A.1. Chemical abundances for individual elements also vary inside this group; for example, [Si/H] varies as follows: $-0.22\pm^{0.20}_{0.30} = \widehat{[\rm Si/H]}_{c2} < \widehat{[\rm Si/H]}_{c8} < \widehat{[\rm Si/H]}_{c4} < \widehat{[\rm Si/H]}_{c6} < \widehat{[\rm Si/H]}_{c9} = 0.26\pm^{0.18}_{0.19}$. This group is similar to group 5 in terms of atmospheric parameters, but classes here are more metal rich. For this group there is some confusion among classes, as shown in Figure 2.5. About 30 percent of the spectra belonging to class 4 in the chosen classification are assigned to class 2 in other classifications.

Classes 2 and 8 are similar in chemical abundances, but differ in $\log g$. Besides metallicity differences, classes 8 and 9 also differ in their spatial distribution over the Galactic plane, as shown in Figure A.2. While stars in class 8, with lower [M/H], lie preferentially at higher galactic longitudes, stars in class 9, which are cooler and more metal rich, are mainly towards the galactic center. In general the spectral fittings for class 9 are poor, with the observed spectral lines being systematically deeper than the synthetic ones. Classes 2, 4, and 6 follow approximately the same spatial distribution of the APOGEE sample.

In the top panel of Figure 2.10 we have a comparison of the mean spectra for all the classes in this group. For group 0, we see that their mean spectra are very similar in shape, but with different line strengths (s). The intensity of lines grows in the sense $s_{c8} < s_{c2} < s_{c4} < s_{c6} < s_{c9}$, following their median temperatures. Together, these classes include ≈ 27 percent of the spectra in DR12.

A.3.2 Group 1 (Classes 7, 14, 19, 25, 26 and 28) - Metal poor cool RGB

As shown in Figure A.3, the classes in group 1 are composed of cooler stars in the RGB ($3500 \lesssim \widehat{T_{\rm eff}} \lesssim 4200$ K and $0.79 \lesssim \widehat{\log g} \lesssim 2.03$) (Binney & Merrifield, 1998). All classes are mainly formed of low latitude stars, composed of a mixture of thin and thick disk population, except for class 28 which is mainly projected towards the Galactic center and with high α abundances, $\widehat{[\alpha/\rm M]} = 0.24\pm^{0.04}_{0.11}$. All of them are classes composed of stars in the RGB, but with different ranges of metallicities[1], surface gravities[2], and temperatures[3].

[1] $-0.81\pm^{0.19}_{0.33} = \widehat{[\rm M/H]}_{c28} < \widehat{[\rm M/H]}_{c26} < \widehat{[\rm M/H]}_{c19} < \widehat{[\rm M/H]}_{c25} < \widehat{[\rm M/H]}_{c7} < \widehat{[\rm M/H]}_{c14} = -0.09 \pm 0.13$.

[2] $0.79\pm^{0.25}_{0.37} = \widehat{\log g}_{c25} < \widehat{\log g}_{c19} < \widehat{\log g}_{c26} < \widehat{\log g}_{c28} < \widehat{\log g}_{c14} < \widehat{\log g}_{c7} = 2.03 \pm 0.22$.

[3] $3561\pm^{84}_{60} = \widehat{T_{\rm eff}}_{c25} < \widehat{T_{\rm eff}}_{c19} < \widehat{T_{\rm eff}}_{c26} < \widehat{T_{\rm eff}}_{c28} < \widehat{T_{\rm eff}}_{c14} < \widehat{T_{\rm eff}}_{c7} = 4236\pm^{97}_{100}$ K.

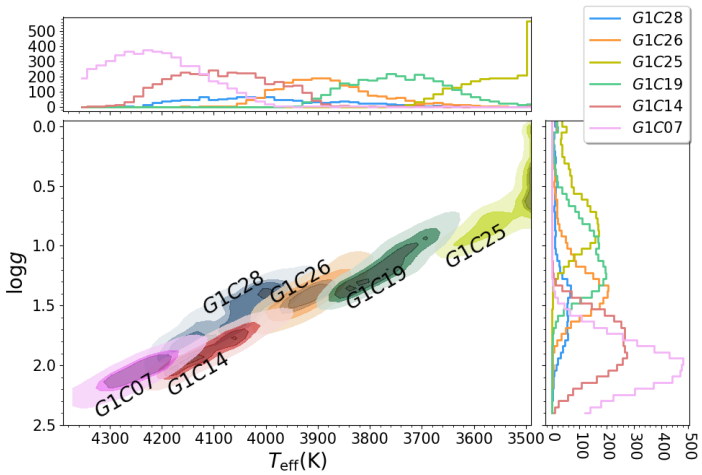

FIGURE A.3— The distribution in $T_{\rm eff} - \log g$ for classes in group 1. The same rules and colors from Figure 2.7 were applied to the contours here. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. The color of the histogram matches the color of the contours.

Concerning the stability of the classes, class 25 is very stable, having a mean coincidence ratio of 82 percent. As shown in Figures 2.7 and A.3, this class consists of giant stars at the tip of the RGB. Confusions higher than 10 percent occurs between classes inside the group. The highest confusion rates are 12 percent and 16 percent between class 7 and classes 14 and 28, respectively, 16 percent between classes 14 and 26, 16 percent between classes 19 and 26, and 30 percent between classes 26 and 28. Again, classes overlapping in the 3D space $T_{\rm eff} - \log g - [{\rm M/H}]$ present the highest degrees of confusion. Between classes in this group and other classes out of the group, the confusion rate is above 5 percent only between class 14 and class 22 (10 percent).

Tables A.4 and A.5 show that the classes in this group are selecting stars within a narrow range of the parameters, including the abundances. They typically have $\hat{\sigma}_{T_{\rm eff}} \approx 100$ K, $\hat{\sigma}_{\log g} \approx 0.30$, and, for example, in class 14, the within class dispersion of the parameter reaches $\hat{\sigma}_X \leq 0.1$ for [α/M], [N/M], [C/M], [Na/H], [Mn/H] and [K/H].

Class 28 is particularly spread in $\widehat{[{\rm C/M}]} = -0.09 \pm^{0.15}_{0.30}$, $\widehat{[{\rm Fe/H}]} = -1.14 \pm^{0.42}_{0.73}$, and $\widehat{[{\rm Al/H}]} = -0.10 \pm^{0.16}_{0.31}$. In Figure 2.10, second panel from top to bottom, we show the mean spectra of the stars in this group. As in group 0, the spectra have very similar shapes but with different line strengths.

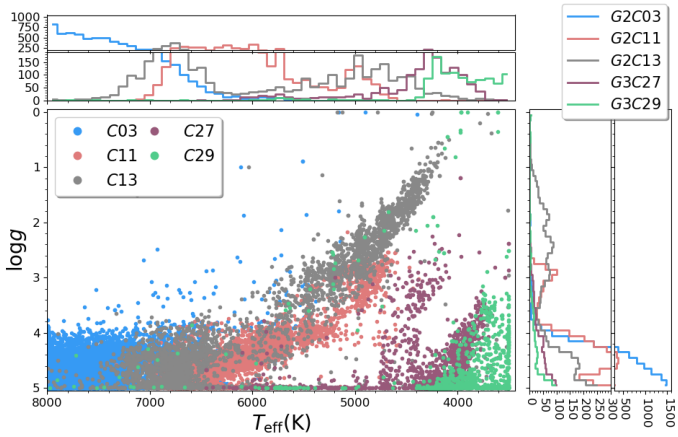

FIGURE A.4— Scatter plot for $T_{\rm eff} - \log g$ distributions of the classes in groups 2 and 3. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. To aid visualization, both panels are split into two plots with different scales. The histogram line color match the color of the scatter plot.

A.3.3 Group 2 (Classes 3, 11, and 13) - Warm stars

This group assembles the warmest stars in APOGEE DR12. The sample includes 15,233 spectra flagged as telluric standards, warm objects ideal for characterizing the telluric lines that plague the IR, of which 67 percent are in class 3, 16 percent in class 11, and 12 percent in class 13. According to target-selection flags, 96 percent of the 10,628 objects in class 3 are telluric standards, while classes 11 and 13 have up to 50 percent of stars of this kind. The differences between the classes in this group are mainly found in $T_{\rm eff}$ and [M/H], as seen in panel b of Figure 2.7; class 3 is the warmest, containing A and B type stars, according to a match with the SIMBAD catalog (Wenger et al., 2000), while classes 11 and 13 are RGB stars, cooler and richer in metals compared to class 3 (see Table A.4). The third panel in Figure 2.10 shows the differences between the mean spectra of the classes in group 2. The mean spectrum of class 3 is almost featureless, while the mean spectrum in class 13 has the strongest lines in the group. Moreover, there is a difference in their spatial distribution; while class 3 mainly occupies low latitudes, classes 11 and 13 are found primarily out of the Galactic plane and towards the Galactic center.

As we would expect, since their spectra are clearly distinguishable, classes in this group are among the most stable classes in the classification, with mean coincidence rates of 94 percent, 73 percent, and 80 percent for classes 3, 11, and 13, respectively. As class 11 is cooler than classes 3 and 13, it has the highest mean confusion rate with other classes. For example, it has about 10 percent of confusion with classes

5 and 24. Classes 5 and 24 are among the most metal-poor in the classification, emphasizing the role that the degeneracy between $T_{\rm eff}$ and [M/H] plays in the determination of the stellar parameters.

All the chemical elements have very wide distributions except for [K/H] in class 11. Nevertheless, the atmospheric parameters of the stars in this group are out of the DR12 model grid, and thus it should be seen as a failure of the model fittings, as suggested by the ASPCAP flag *star warn* found in $\approx$ 35 percent of the objects in this class.

A.3.4 Group 3 (Classes 27 and 29) - Fast rotators

This group is formed by fast rotating stars. In both classes the spectra are poorly fitted by their model spectra. As a consequence, artifacts are observed in several abundances, for example, the abundances of [C/M], [α/M], [Al/H], [K/H], [Na/H] and [Si/H] are not continuous. They appear in clumps, having gaps of at least 0.2 dex in abundance between them.

In terms of atmospheric parameters, this group is very close to group 6 (dwarfs), but their spectra are remarkably different. The spectra of group 3 have fewer, shallower, and broader lines than those found in group 6, as can be seen in the fourth and seventh panel in Figure 2.10. This shows that the algorithm is sensitive to rotation, since it is able to split the stars affected by $\log g$ line broadening from those affected by rotational line broadening. On the other hand, ASPCAP determines that the vast majority of the stars in this group have $\log g$ greater than 4.9 (see Figure A.4), but since the rate of stars flagged with a fast rotation warning are 81 percent and 93 percent for classes 27 and 29, respectively, we cannot trust the determinations of the stellar parameters of these stars. The rate of stars flagged with a rotation warning in the entire DR12 dataset is 7 percent.

Class 29 is the most unstable of the classes, excluding the outliers (see Section A.3.9). It has a confusion rate of 62.8 percent with class 27, which means that for some classifications class 29 dissolves mainly in classes 13, 23, 27 and 29. Class 27 is more stable, with 63 percent of coincidence, having some degree of confusion with class 10 (13 percent), which has the shallower lines in group 6.

About one quarter of the stars in class 27 and about half of the stars in class 29 are either young embedded cluster members or known calibration cluster members. Statistically we expect fast rotating stars to be younger than those that rotate more slowly (van Saders & Pinsonneault, 2013). In addition, the great majority of stars form in star clusters, dispersing latter on, and thus the fastest rotating stars are expected to be in young embedded clusters.

A.3.5 Group 4 (Classes 16, 18 and 22) - Metal-rich cool RGB

Group 4 classes include metal rich stars covering the RGB with effective temperatures from 3620 to 4140 K, and with metallicities from 0.17 to 0.22 in the order $[\widehat{\rm M/H}]_{c16} < [\widehat{\rm M/H}]_{c18} < [\widehat{\rm M/H}]_{c22}$. Some stars in this group are near the edge of the model grid, at [Fe/H] = 0.50 (36 percent in class 16, 26 percent in class 18, and 24

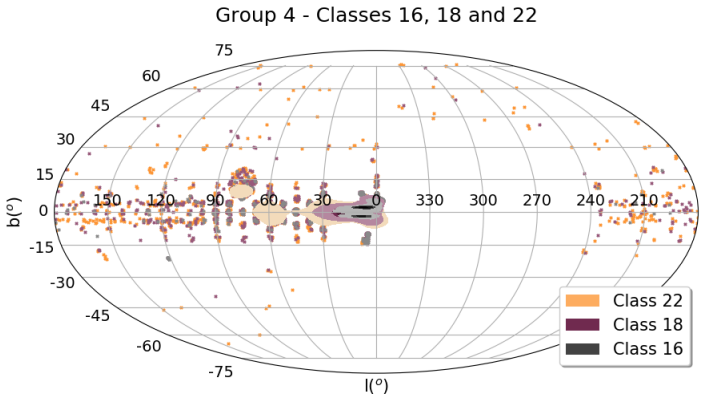

FIGURE A.5— Galactic coordinates in Mollweide's projection for objects in classes 22 (orange triangles and contours), 18 (purple triangles and contours), and 16 (gray circles and contours), all belonging to group 4. The contours and individual points follow the same rule as in Figure 2.7.

percent in class 22). This also happens for $T_{\rm eff}$ in class 16, which has 43 percent of the stars cooler than 3600 K.

The stars in these classes are very concentrated in the galactic disk, with [α/M] close to the solar value. As shown in Figure A.5, the spatial distribution of class 16 is more concentrated towards the Galactic center than classes 18 and 22.

Classes 16 and 18 are very stable, with a coincidence rate of 91 percent and 80 percent, respectively. Class 22 is much less stable having a coincidence rate of 29 percent. The highest degree of confusion for class 22 occurs with class 9 (38 percent), but classes 14 and 18 also contaminate class 22. These three classes, 9, 14, and 18, share borders with class 22 in the space $T_{\rm eff}$ − [M/H], as shown in Figure 2.7, and they are also superposed in $\log g$, as can be seen by comparing Figures A.3 and A.6. Once again, we see that the overlap in the space $T_{\rm eff}$ − [M/H] − $\log g$ is the main cause of confusion between classes. The abundance distributions for these classes are narrow, as reflected in Tables A.4 and A.5.

A.3.6 Group 5 (Classes 0, 1 and 5) - Metal-poor RC/warm RGB

Just like group 0, this group is made of classes that include stars from the RC and the warmest end of the RGB. For classes 0, 1, and 5, the ratios of red clump stars are 30, 31, and 16 percent according to a comparison with the compilation carried out by Bovy et al. (2014). This group is more metal-poor than group 0, with $-0.45 \lesssim \widehat{[\mathrm{M/H}]} \lesssim -0.22$. The group lacks stars in the direction of the Galactic center, being homogeneously distributed in all other directions. Relative to group 0, group 5 is more dense in regions with Galactic latitudes higher than 30 degrees. All

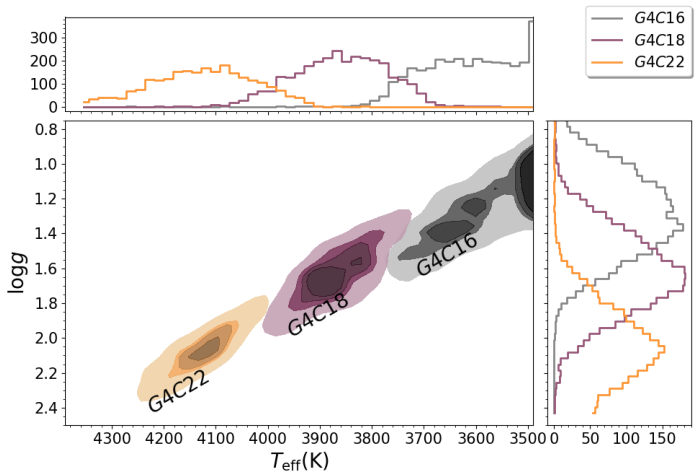

FIGURE A.6— $T_{\rm eff} - \log g$ distribution for classes in group 4. The same rules and color from Figure 2.7 were applied to contours here. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. The color of the histograms match the color of the contours.

three classes are a mixture of thin and thick disk populations, but class 5 is more populated by high [α/M] stars than other classes in the group, as shown in Figure A.7.

As shown in Figure A.8, class 5 almost completely overlaps with classes 0 and 1 in $T_{\rm eff} - \log g$ space. The median temperatures of class 0 stars are about 150 K warmer than class 1 stars. Class 5 is particularly broad in $T_{\rm eff}$ and $\log g$, covering temperatures from 4125 to 7170 K, with a median value of $\widehat{T_{\rm eff}} = 4942 \pm^{584}_{202}$ K and $\log g = 3.16 \pm^{+1.04}_{-0.38}$. Figure A.7 shows the distribution of the stellar parameters in the planes $T_{\rm eff}$ – [M/H], $T_{\rm eff}$ – [α/M] and [α/M] – [M/H]. The dispersion there is likely to be an artifact due to the degeneracy between $T_{\rm eff}$ and [M/H] in the ASPCAP parameter determination pipeline. Also the class is broadly spread in $\widehat{[{\rm Si/H}]} = -1.38 \pm^{0.96}_{1.38}$, which may also be an artifact of ASPCAP. In this range of atmospheric parameters the pipeline is probably confusing warmer temperatures with lower metallicities, as discussed in Holtzman et al. (2015).

A.3.7 Group 6 (Classes 10, 12, 15, 17 and 20) - Dwarfs stars

With $\log g$ ranging from 4.23 to 4.35, group 6 has only dwarf stars. The classes differ because of their different temperatures and abundance patterns. Figure A.9 shows the distribution of $\log g$ and $T_{\rm eff}$ for this group.

Class 12 is over-abundant in Mg ($\widehat{[{\rm Mg/H}]} = +0.38 \pm^{0.32}_{0.28}$), and classes 15 and 20 have low [α/M], especially in [Ca/H] and [O/H]. Some bimodality is found for

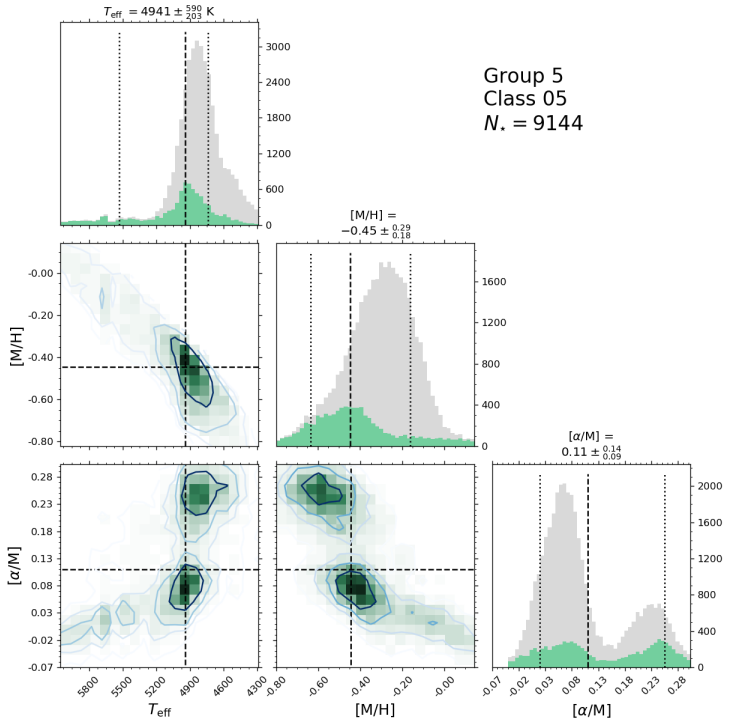

FIGURE A.7— Properties of class 5 (group 5), which contains 9,144 stars ($N_\star$). The panels in the uppermost diagonal contain histograms for $T_{\rm eff}$, [M/H] and [α/M], from left to right, respectively. In these plots vertical black dashed lines show the median value and the limits enclosing 68.3 percent of the data points around the median value. The green histograms correspond to the objects in class 5 and the gray histogram shows the distribution of the whole group 5. As indicated by labels in the axes, the other three panels show 2D histograms for $T_{\rm eff}$ – [M/H], $T_{\rm eff}$ – [α/M] and [α/M] – [M/H]. The contours enclose 68.3, 45, 30, and 15 percent of the objects in the class.

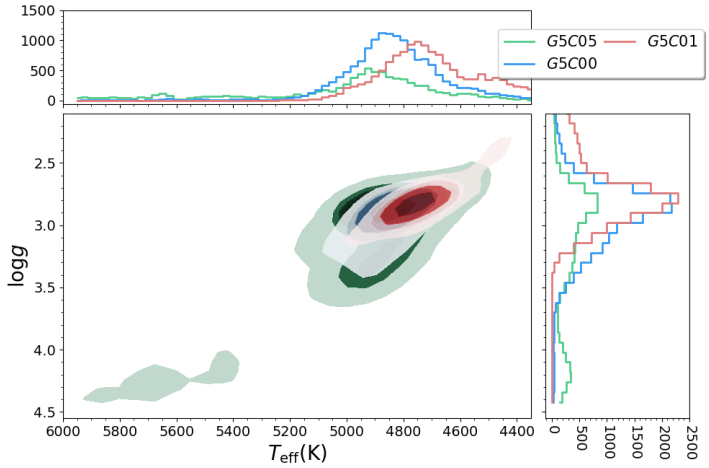

FIGURE A.8— $T_{\rm eff} - \log g$ distribution for classes in group 5. The same rules and colors from Figure 2.7 were applied to the contours here. Top and right panels show histograms with the distributions of $T_{\rm eff}$ and $\log g$, respectively. The color of the histograms matches the color of the contours.

[Al/H] and [K/H] for classes 15 and 20. However, 99 percent of the objects in the group have their chemical abundances flagged with a warning and are not reliable, so this strange behavior is likely to be an artifact of ASPCAP.

Figure 2.10 shows how the FeI line around 16210 Å is blended with the CN and CO lines for classes 15 and 20. In other regions of the spectra, blends like this are present. This is caused by the enhancement of molecular lines at low $T_{\rm eff}$ values.

Class 20 presents two separate blobs of [α/M] abundances, one around solar values and the other around $[\widehat{\alpha/\mathrm{M}}] = -0.3$, but almost 70 percent of the stars in this class are flagged with the star warning, so the abundance determination for these stars is not reliable. The abundance distributions of these classes are very narrow, as shown in Tables A.4 and A.5.

The classes here are relatively stable. Class 17 is the most unstable (50 percent of mean coincidence rate), but has a significant degree of confusion only with classes 10, 12, and 15. Class 20 is the most stable in the group with a mean coincidence rate of 81 percent. Other significant confusion rates are found only between classes inside the group, showing that the classes are stable as a group.

A.3.8 Group 7 (Classes 21, 23, 24, 30 and 31) - Sparse classes

This group is formed by the most peculiar classes, with a number of objects corresponding to at least 0.5 percent of the whole DR12 sample. The group is very

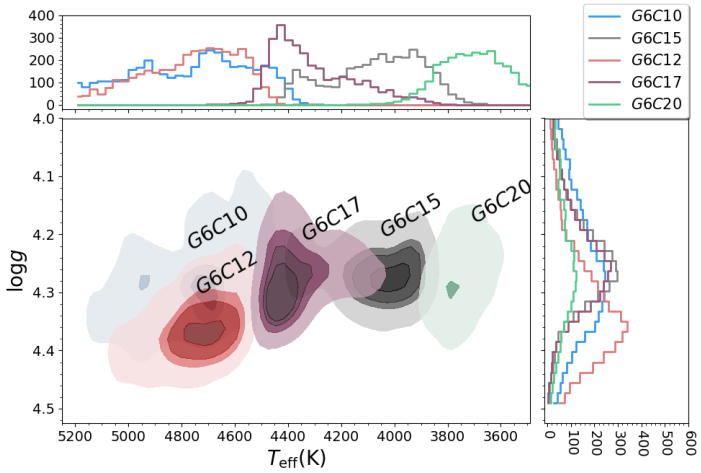

FIGURE A.9— $T_{\rm eff} - \log g$ distribution for classes in group 6. The same rules and colors from Figure 2.7 were applied to the contours here. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. The histogram line color matches the color of the contours.

diverse, so in this case we describe each class individually. All classes that represent less than 0.5 percent of the sample are treated as outliers and are discussed in Section A.3.9. Figure A.10 shows the $T_{\rm eff}$- $\log g$ distribution for the group.

Class 21 - M-giants/Bulge

Ninety-seven percent of the stars in class 21 are at the edge of the grid of models in $T_{\rm eff}$. That is to say, their temperatures are likely to be lower than the minimum $T_{\rm eff}$ of the models in the spectral library. The class presents other anomalies; except for [C/M], [N/M], [α/M], [Al/H], [K/H], [Mn/H] and [Na/H], all other abundances are also at the edge of the grid of models. Lacking sufficiently cool spectra, ASPCAP probably tries to compensate this deficiency with an unphysical combination of the abundances. The problem has been corrected in DR13 (SDSS Collaboration et al., 2016). This class is the most stable class with a coincidence rate of 95 percent. Figure 2.10, bottom panel, shows that the mean spectra of this class looks totally different from the other classes, with very strong molecular bands, so K-means easily identifies these spectra as a class. Spatially, the stars are concentrated at low latitude, specially towards the galactic center, as shown in Figure A.11. This class also gathers 23 percent of the bulge targets in DR12, according to its target flags in the APOGEE catalog.

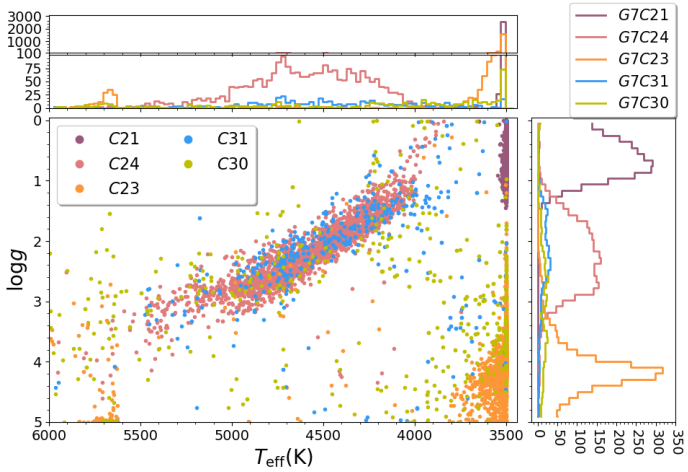

FIGURE A.10— Scatter plot for $\log g$ versus $T_{\rm eff}$ of the classes in group 7. Top and right panels show histograms of the distributions of $T_{\rm eff}$ and $\log g$, respectively. Top panel is divided in two plots with different scales, as indicated in the vertical axis of the plots. The color of the lines in the histograms matches the color of the scatter plot, as indicated in the legends.

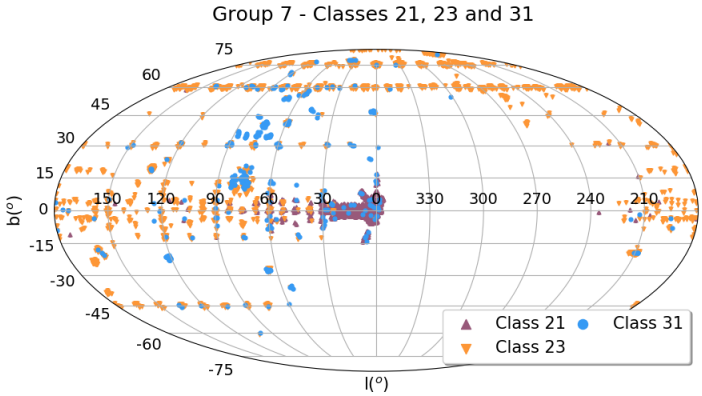

FIGURE A.11— Galactic coordinates distribution of classes 21 (purple triangles), 23 (orange triangles) and 31 (blue circles).

Class 23 - Metal-poor M dwarfs

This class is dominated by metal-poor ($\widehat{[M/H]} \approx -0.54$) M dwarfs. The distribution of $[\alpha/M]$ is divided into four clumps, showing a problem in the determination of these abundances, since very similar spectra correspond to differences of 0.25 in $[\alpha/M]$. The mean spectrum is similar to that of class 20, but with cooler stars; here more than 60 percent of the stars are at the minimum $T_{\rm eff} = 3500$ K. This similarity in their spectra causes a mean confusion rate with class 20 of 12 percent. However, class 23 is quite stable, with a mean coincidence rate of 87 percent. Similar to what happened to class 21, this class has many anomalies in its parameters, gaps in chemical abundances, and a high concentration at the borders of the grid of models. This can also be related to limitations in ASPCAP. As shown in Figure A.11, there seems to be no anisotropy in the spatial distribution of this class.

Class 24 - K-giants from the Halo

This is a very metal-poor class with stars lying over the whole RGB, $\widehat{T_{\rm eff}} = 4583\pm^{322}_{330}$ K and $\widehat{\log g} = 2.22\pm^{0.60}_{0.54}$, as shown in Figure A.10. With a median metallicity of $\widehat{[M/H]} = -1.20\pm^{0.22}_{0.25}$, it is one of the most metal-poor classes in the classification, certainly the most well-defined class among the metal-poor ones. This class is also α enhanced, with $\widehat{[\alpha/M]} = 0.24 \pm 0.07$. We find that 593 out of 2388 ($\approx$ 25 percent) of these objects are globular cluster members used for calibration of APOGEE. Its spacial distribution is more dense in Galactic latitudes above 30º. Class 24 has a very low stability, having a coincidence rate of 18 percent. Its stars are classified as class 11 members 59 percent of the time.

Class 30 - M31 GCs

In APOGEE DR12, 236 integrated spectra of unresolved Globular Clusters (GCs) from M31 were observed; each of these spectra is duplicate in the dataset. In order to remove the contamination from the unresolved M31 stellar population in these spectra, 141 background spectra near to the clusters were obtained (Zasowski et al., 2013). Altogether they add up to 613 spectra in the region of M31. This class has the largest number of objects in this region, 171, with 33 background spectra and 69 duplicated GCs spectra. In general the spectra present high absorption in the continuum, as shown for the mean behaviour by the yellow line in the bottom panel of Figure 2.10. Its spectra are poorly fitted by the ASPCAP, and their wide chemical abundances and atmospheric parameters distributions (see Tables A.4 and A.5) should not be trusted since they are all flagged with ASPCAP warnings. Sakari et al. (2016) have determined the abundance for 25 of the GCs in DR12 (eight are in this class) and we refer to their work as a better source of chemical abundances for these objects. This group also has 62 stars in embedded clusters, two member candidates of the GC Palomar 1, six bulge giants, and many metal-poor RGB stars. The class has 562 spectra, from which 93 percent are flagged with star warnings, so the ASPCAP values cannot be trusted.

Class 31 - M31 GCs/high persistence

Class 31 also has some spectra from the region of M31 (84 out of 613), 20 of which are background spectra and 64 are duplicated spectra of 32 clusters. In this class the spectra seem to be less affected by continuum absorption. As shown by the light blue circles in Figure A.11, this class has a peculiar spacial distribution, being more dense in $60^o \leq l \leq 90^o$ and $0^o \leq b \leq 45^o$. Further investigation is needed to determine why the stars in that direction have these characteristics. In this class there are 88 calibration cluster members and 38 spectra that overlap with the Kepler mission sample. Comparing the position of the stars of this class in Figure A.11 with Figure 2 in Zasowski et al. (2013) one sees that the position of these objects matches the positions where the halo population, the Kepler mission, and some of the calibration clusters were observed. Thirty-five percent (170) of the spectra in this class are flagged with a warning.

Thirty one percent of the stars in this class are flagged as *high persistence* observations. Persistence refers to the latent image of a previous exposure appearing in subsequent images, due to a slow release of an appreciable fraction of accumulated charge in the previous exposure over the subsequent ones. It affects the bluest chip particularly (Nidever et al., 2015). The intensity of the persistence effect depends on the brightness of the spectra and their history of previous observations. In DR12, a flag is used to inform the relevance of the persistence effect on each spectra (Holtzman et al., 2015). Some of the affected spectra by persistence present an obvious excess/deficit of flux in the blue chip. This behavior is flagged as a *positive/negative jump in blue chip*.

A.3.9 Group 8

Ninety-nine percent of the stars in APOGEE are in the classes presented in Sects. A.3.1 to A.3.8. Here we briefly discuss the remaining 1 percent. Figure A.12 shows the number of spectra in the classes of group 8. Figure A.13 shows the spatial distribution of these classes, which are represented by different symbols and colors. Figure A.14 shows the spectra in classes from 32 to 38 in the same wavelength window used for Figure 2.10; we plot the spectra as semi-transparent black lines to highlight the locations where the spectra are closer to each other. Figure A.14 shows the mean spectrum of each class is drawn with a white dashed line.

Class 32 - Bulge giants

This class has 269 spectra, 71 of which are of supergiant stars in the bulge, 33 are bulge giants, and 44 are spectra in the region of M31 (20 of the background and 24 of 12 duplicated GCs). Forty-one percent of the spectra in this class are flagged as having a negative jump in the blue chip, 19 percent of them as having high persistence, and 99 percent of them are flagged as *star bad*, assigned if there is warning on any of the following issues: $T_{\rm eff}$, $\log g$, model fitting χ^2, rotation, S/N (signal-to-noise ratio), or if the difference between photometric and spectroscopic

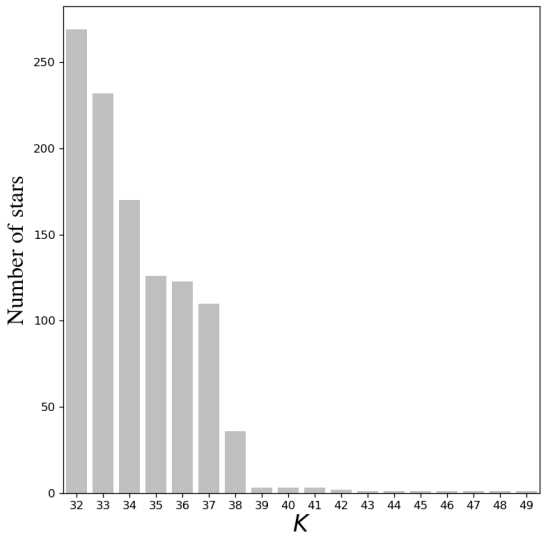

FIGURE A.12— Number of objects in minor classes.

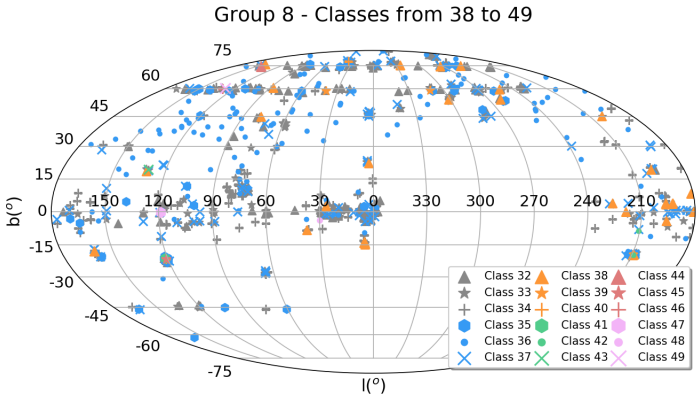

FIGURE A.13— Galactic coordinates for the targets in group 8.

temperature is greater than 500K.

Class 33 M31 - GCs/high persistence

Class 33 has 116 spectra in the region of M31, 18 background spectra, and 98 spectra from 49 GCs. There are 39 spectra flagged as emission line stars in DR12, eight of which are in this class. Figure A.14 second panel from the top shows all 232 spectra overlapped. Emission lines are not visible in this figure because all spectra were truncated at 1.02 of the normalized flux. In spite of this constraint, the algorithm is able to identify emission lines since they affect the form of the continuum around them. Ninety-five percent of the 232 spectra in this class are flagged as *star bad*, 56 percent are flagged with the rotation warning, and 33 percent of these spectra are flagged as high persistence spectra. As one can see in Figure A.14, there is no clear resemblance between the spectra in the class. The wavelengths where the spectra are more similar seem to be emission-dominated by lines, suggesting that the spectra are either actually emission line stars or have some problems with the sky subtraction.

Class 34 - Bad pixels

Seventy six percent of the 170 stars in this class are flagged as high persistence observations. Figure A.14, we see they are mainly giant stars whose spectra have sequences of bad pixels, as those seen between 16,205 and 16,220 Å.

Class 35 - M31 GCs/high persistence

Class 35 has 88 spectra in the region of M31, 38 background spectra, and 50 spectra from 25 duplicated GCs. These 88 spectra represent 70 percent of the 126 spectra in the class. There are 99.2 percent of the objects in this class are flagged as *star bad*, and 94 percent of them have S/N per pixel lower than 30. As we see in Figure A.14, central panel, all the spectra are very noisy.

Class 36 - 1m Telescope

In DR12, there are 817 spectra observed with the 1m telescope in DR12, and 93 of these are in this class. With 123 spectra, it corresponds to 76 percent of the spectra in the class. Apart from a few cases, the spectra seem to contain sequences of a few bad pixels like the ones seen in class 34, Fig. A.14, but in different regions of the spectrum.

Class 37 - Emission line stars/M31 GCs

This class has 13 emission line stars. There are also 11 spectra in the M31 region, one spectrum from the background, and ten spectra of five GCs. There are six objects identified by SIMBAD as galaxies.

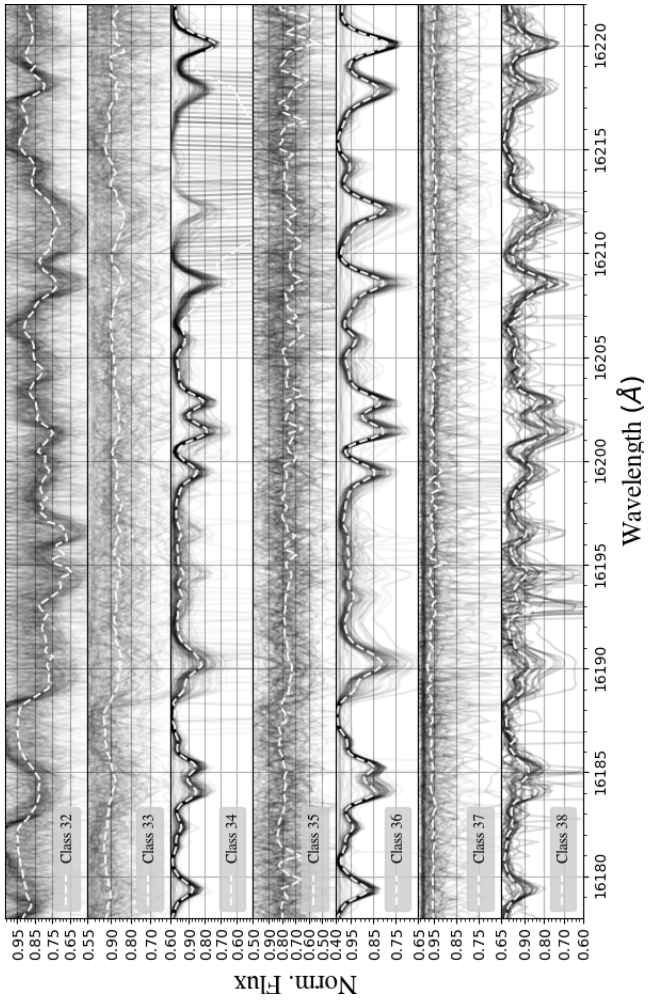

FIGURE A.14— Spectra of the objects in classes from 32 to 38. Each spectra is plotted as a semi-transparent line, in a way that the darkest regions represent the most dense regions in this flux window. The wavelength coverage here is the same as in Figure 2.10.

Class 38 - Negative flux

This class has 36 spectra, eight of which are embedded cluster members, four are Sagittarius dwarf galaxy members, and one is an integrated spectra of the Pal1 GC. Eighty three percent of the spectra in the class have pixels with negative counts.

Classes from 39 to 49

Except for class 42, all classes here have extreme negative flux values in some pixels. These negative counts imply high Euclidean distances between these spectra and those restricted to positive fluxes. Therefore they are segregated within these classes. Here we give a brief description of these objects.

- **Class 39:** Three noisy spectra, one of them flagged as an embedded cluster member.
- **Class 40:** Two duplicated spectra of a globular cluster in M31 and one spectrum of the background in the M31 region.
- **Class 42:** Two stars with a very similar pattern of sequences of pixels with flux equal to zero.
- **Class 43:** One spectrum of the Pal1 globular cluster. This spectrum has deep asymmetric lines.
- **Class 44:** One noisy spectrum with negative spikes.
- **Class 45:** One background spectrum in the region of M31.
- **Class 46:** One stellar spectrum with broad absorption lines.
- **Class 47:** One spectrum with great negative spikes.
- **Class 48:** One spectrum with high persistence and a positive jump in the blue chip.
- **Class 49:** One noisy spectrum with wide absorption lines.