UNIVERSIDADE DA CORUÑA

Tesis Doctoral

Extracción de conocimiento en bases de datos astronómicas mediante redes de neuronas artificiales. Aplicaciones en la misión Gaia

Autor:
Diego Fustes Villadóniga

Directores:
Dr. Minia Manteiga Outeiro,
Dr. Bernardino Arcay Varela

Laboratorio Interdisciplinar de Aplicaciones de la Inteligencia Artificial (LIA2)
Departamento de Tecnologías de la Información y las Comunicaciones

Febrero 2014

Declaración de Autoría

Yo, Diego Fustes Villadóniga, declaro que la tesis titulada 'Extracción de conocimiento en bases de datos astronómicas mediante redes de neuronas artificiales. Aplicaciones en la misión Gaia' y el trabajo presentado en la misma es original.

El Dr. Bernardino Arcay Varela, Catedrático de Universidad en el Área de Ciencias de la Computación e Inteligencia Artificial de la Universidade da Coruña, y la Dra. Minia Manteiga Outeiro, Profesora Titular en el Área de Física de la tierra, Astronomía y Astrofísica de la Universidade da Coruña, hacen constar que la tesis titulada 'Extracción de conocimiento en bases de datos astronómicas mediante redes de neuronas artificiales. Aplicaciones en la misión Gaia' ha sido realizada por Diego Fustes Villadóniga, bajo nuestra dirección, en el Departamento de Tecnoloxías da Información e as Comunicacións de la Universidade da Coruña y constituye la Tesis que presenta para optar al grado de Doctor en Informática de la Universidade da Coruña.

Firmado: Diego Fustes Villadóniga

Firmado: Bernardino Arcay Varela

Firmado: Minia Manteiga Outeiro

"Nunca consideres el estudio como una obligación, sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber."

Albert Einstein

Abstract

In the so-called *IT* era, the capabilities of data acquisition systems have increased to such an extent that it has become difficult to store all the information they produce, and analyse it. This explosion of data has recently appeared in the field of Astronomy, where an increasing number of objects are being observed on a regular basis. An example of this is the upcoming Gaia mission, which will pick up multiple properties of a billion stars, whose information will have a volume of approximately a *petabyte*. The analysis of a similar amount of information inevitably requires the development of new data analysis methods to extract all the knowledge it contains. This thesis is devoted to the development of data analysis methods to be integrated in the Gaia pipeline, such that knowledge can be extracted from the data collected by the mission.

In order to analyze the data from the Gaia mission, the European Space Agency organized the Data Processing and Analysis Consortium (DPAC) which is composed of hundreds of scientists and engineers. DPAC is divided into eight Coordination Units (CUs). This thesis is dedicated to algorithm development in CU8, which is responsible for source classification and astrophysical parameters (AP) estimation. Methods based on Artificial Neural Networks (ANNs) are developed to perform the tasks related to two different work packages in CU8: the GSP-Spec package (GWP-823), and the OA package (GWP-836). The GSP-Spec package is responsible for estimating stellar APs by means of the Radial Velocity Spectrograph (RVS) spectrum. This work presents the development of one of the GSP-Spec modules, which is based on the application of *feed-forward* ANNs. A methodology is described, based on the optimization of genetic algorithms and aimed at obtaining an optimal set of configuration parameters for the ANN in each case, depending on the signal to noise ratio (SNR) in the RVS spectrum and on the type of star to parameterize. Furthermore, in order to improve the AP estimates, *wavelet* signal processing techniques, applied to the RVS spectrum, are studied. Despite the effectiveness shown by ANNs in estimating APs, in principle they lack the ability to provide an uncertainty value on these estimates, making it impossible to determine their reliability. Because of this, a new architecture for the ANN is presented in which the inputs and outputs are reversed, so that the ANN estimates the RVS spectrum from the APs. Such an architecture is called Generative ANN (GANN) and is applied to the AP estimation of a set of simulated RVS spectra for the Gaia mission, where it is more effective than the conventional ANN model, in the case of faint stars with low SNR. Finally, the GANN can be applied for obtaining the posterior probability of each of the APs according to the RVS spectrum, allowing for their more complete analysis.

Given the nature of the Gaia mission, which is the first astronomical mission that will observe, in an unbiased way, the entire sky up to magnitude 20, a large number of outliers are expected. The OA package in CU8 handles the processing of this type of objects, which are defined as those that could not be reliably classified by the methods in the upstream classification packages. OA methods are based on the unsupervised learning of all outliers. Such learning has two parts: clustering and dimensionality reduction. The Self-Organizing Map (SOM) algorithm is selected as a basis for this learning. Its effectiveness is demonstrated when it is applied, with an optimal configuration, to the Gaia simulations. Furthermore, the algorithm is applied to real outliers from the SDSS catalog. Since a subsequent identification of the clusters obtained by the SOM is necessary, two different methods of identification are applied. The first method is based on the similarity between the SOM prototypes and the Gaia simulations, and the second method is based on the recovery of stored classifications in the SIMBAD catalog by cross-matching celestial coordinates. Thanks to the visualization of the SOM planes, and to both methods of identification, it is possible to distinguish between valid observations and observational artifacts. Furthermore, the method allows for the selection of objects of interest for follow-up observations, in order to determine their nature.

Resumen

En la llamada *era de las TIC*, las capacidades de los sistemas de adquisición de datos han aumentado enormemente, de forma que resulta complicado almacenar toda la información que producen, así como su análisis posterior. Esta explosión de datos ha aparecido recientemente en el campo de la Astronomía, donde cada vez se observan un número mayor de objetos, con mayor periodicidad. Un ejemplo de esto es la próxima misión Gaia, que observará múltiples propiedades de hasta mil millones de estrellas, cuya información tendrá un volumen del orden del *petabyte*. Por lo tanto, para analizar tal cantidad de datos, es necesario desarrollar nuevos métodos de análisis que permitan extraer todo el conocimiento presente en los mismos. Esta tesis está dedicada al desarrollo de métodos de análisis de datos, los cuales se integran en la cadena de procesado de Gaia, con el objetivo de extraer conocimiento de los datos recogidos por la misión.

Con el objetivo de analizar los datos de la misión Gaia, se ha organizado un consorcio a nivel europeo, llamado *Data Processing and Analysis Consortium* (DPAC), compuesto por cientos de científicos e ingenieros. DPAC se divide en ocho unidades de coordinación (CUs), estando esta tesis dedicada al desarrollo de algoritmos en la CU8, que se encarga de la estimación de parámetros astrofísicos (APs) y la clasificación de las fuentes. Se desarrollan métodos basados en redes de neuronas artificiales (ANNs) para realizar las tareas relacionadas con dos paquetes de trabajo diferentes en la CU8: El paquete GSP-Spec (GWP-823) y el paquete OA (GWP-836).

El paquete GSP-Spec se encarga de la estimación de APs de estrellas mediante el espectro del instrumento *Radial Velocity Spectrograph* (RVS). Aquí, se presentará el desarrollo de uno de los módulos de GSP-Spec, el cual se basa en la aplicación de ANNs de tipo *feed-forward*. Se presenta una metodología, basada en algoritmos genéticos de optimización, para la obtención de un conjunto óptimo de parámetros de configuración para la ANN en cada caso, en función de la relación señal a ruido (SNR) en el espectro RVS y del tipo de estrella a parametrizar. Además, con el objetivo de mejorar las estimaciones de APs, se estudian técnicas de procesado *wavelet*, aplicadas sobre el espectro RVS. A pesar de la efectividad que las ANNs muestran a la hora de estimar APs, en principio éstas carecen de la capacidad de proporcionar un valor de incertidumbre sobre dichas estimaciones, con lo cual resulta imposible conocer la fiabilidad de las mismas. Debido a ello, se presenta una arquitectura novedosa para la ANN, en la cual se invierten las entradas y salidas de la misma, de forma que la ANN estima el espectro RVS a partir de los APs. Dicha arquitectura de denomina red neuronal artificial generativa (GANN) y se aplica a la estimación de APs de un conjunto de espectros RVS

simulados para la misión Gaia, donde se muestra más efectiva que el modelo de ANN convencional, en el caso de estrellas débiles, con un bajo SNR. Finalmente, la red GANN puede aplicarse para la obtención de la probabilidad a posteriori de cada uno de los APs dado el espectro RVS, lo cual permitirá un análisis más completo de los mismos.

Dada la naturaleza de la misión Gaia, la cual es la primera misión astronómica que observará, de forma no sesgada, toda la bóveda celeste hasta magnitud 20, se espera una gran cantidad de objetos atípicos. El paquete OA en la CU8 se encarga del procesado de dicho tipo de objetos, los cuales se definen como aquellos que no han podido ser clasificados con fiabilidad por los paquetes de clasificación existentes en la cadena de procesamiento. Los métodos de OA se basan en el aprendizaje no supervisado del conjunto de observaciones atípicas. Dicho aprendizaje tiene dos partes: agrupamiento y reducción de dimensionalidad. Se seleccionan los mapas auto-organizativos (SOM) como algoritmo base para realizar dicho aprendizaje, demostrándose su efectividad cuando se aplica, con una configuración óptima, a las simulaciones de Gaia. Además, el algoritmo es aplicado a observaciones atípicas reales, provenientes del catálogo SDSS. Dado que es necesaria una identificación posterior de los grupos obtenidos por la red SOM, se aplican dos métodos de identificación diferentes. El primero está basado en la similitud entre los prototipos de la red y el conjunto de simulaciones de Gaia, mientras que el otro esa basado en la recuperación de clasificaciones almacenadas en el catálogo *Simbad*, mediante el cruce de coordenadas celestes. Gracias a la visualización de la red SOM, y a ambos métodos de identificación, es posible distinguir entre observaciones válidas y artefactos observacionales. Además, el método posibilita la selección de objetos de interés para observaciones posteriores, con el objetivo de determinar la naturaleza de los mismos.

Agradecimientos

A mis padres y el resto de mi familia, por su amor y por todo su esfuerzo que ha hecho posible esta tesis.

A mis directores, Minia y Berni, por haberme ayudado a lo largo de estos años de duro trabajo y por brindarme la oportunidad de estudiar un doctorado.

A Carlos Dafonte, por su ayuda, por haber confiado desde el principio y por seguir aguantándome después de cinco años comiendo juntos.

A los estudiantes que han realizado el proyecto de fin de carrera bajo mi tutela.

A mi anfitriones en mis estancias fuera: Carme, Carrasco, Coryn, Kester, Alejandra y Patrick.

A Ángela (*bicho*) por estar siempre a mi lado en estos años, en los momentos buenos y en los malos.

A todos los que se me quedan en el tintero, que han formado parte de mi vida en algún momento y por lo tanto también han contribuido a que este documento sea una realidad.

Tabla de contenidos

Abstract iii

Resumen v

Tabla de contenidos viii

Lista de Figuras x

Lista de Tablas xiii

Acŕonimos xiv

1 Introducción 1
- 1.1 Breve historia de la astronomía . . . 1
- 1.2 La misión Gaia . . . 5
- 1.3 Estadística, inteligencia artificial, aprendizaje máquina y minería de datos 9
- 1.4 El consorcio DPAC . . . 13

2 Estimación de parámetros astrofísicos mediante ANNs 19
- 2.1 Estado del arte . . . 19
- 2.2 DU-823: GSP-Spec . . . 21
- 2.3 Simulación de espectros RVS . . . 23
- 2.4 El módulo ANN en GSP-Spec . . . 27
  - 2.4.1 Objetivos de ANN . . . 27
  - 2.4.2 Configuración de la ANN . . . 27
    - 2.4.2.1 Arquitectura de la ANN . . . 28
    - 2.4.2.2 Método de aprendizaje para la ANN . . . 28
  - 2.4.3 Selección y extracción de características . . . 32
    - 2.4.3.1 Extracción de características mediante procesado wavelet 32
    - 2.4.3.2 Selección de características mediante algoritmos genéticos 34
  - 2.4.4 Obtención de APs y medidas de incertidumbre mediante ANNs generativas (GANNs) . . . 38
  - 2.4.5 Estimaciones de APs obtenidas por el módulo ANN . . . 44
  - 2.4.6 Implementación . . . 50

3 Minería de objetos atípicos mediante clasificación no supervisada 51
3.1 Estado del arte . . . 52
3.1.1 El paquete de trabajo DSC . . . 52
3.1.2 Detección de anomalías y aprendizaje activo . . . 53
3.1.3 Aprendizaje no supervisado . . . 54
3.2 DU-836: OA . . . 56
3.3 Simulación de espectrofotometría BP/RP . . . 57
3.4 Técnicas de agrupamiento para simulaciones de Gaia . . . 59
3.4.1 Preprocesado de datos BP/RP . . . 62
3.4.2 Medidas de evaluación de algoritmos de agrupamiento . . . 63
3.4.3 Algoritmo k-means . . . 64
3.4.4 Mapas auto-organizativos . . . 66
3.4.4.1 Visualización de mapas auto-organizativos . . . 71
3.4.5 Sensibilidad al ruido de las técnicas de agrupamiento . . . 75
3.4.6 Técnicas de extracción de características para el agrupamiento de espectros BP/RP . . . 79
3.4.7 Métodos de agrupamiento conjunto . . . 86
3.5 Análisis de objetos atípicos en SDSS . . . 90
3.5.1 Identificación de grupos en la red SOM mediante cruce con catálogos externos . . . 93
3.5.2 Agrupamiento conjunto de espectros de SDSS . . . 98
3.6 Métodos para el análisis de objetos atípicos en Gaia . . . 99
3.7 Implementación . . . 99
3.7.1 Implementación en APSIS . . . 100
3.7.2 Implementación en MapReduce . . . 102
3.7.3 Técnicas especiales de aceleración del entrenamiento de la red SOM 103

A Aplicaciones en otros dominios 105
A.1 Detección de vertidos de fuel mediante imágenes SAR . . . 105
A.2 Sistema de detección de intrusiones y monitorización de tráfico de red . . . 107

Conclusiones 110

Conclusions 113

Bibliografía 116

Lista de Figuras

1.1 Diagrama de Hertzsprung-Russell. Extraído de `http://lcogt.net`. 3
1.2 Representación del espectro electromagnético. Extraído de `http://www.artinaid.com/`. 3
1.3 Rotación de los campos de visión de Gaia. La precesión del eje de giro permite mapear diversas zonas del cielo. Por cortesía de ESA. 6
1.4 Distribución de tránsitos del satélite Gaia después de 5 años de misión. Coordenadas ICRS. Por cortesía de ESA. 6
1.5 Instrumentos montados en el satélite Gaia. Por cortesía de ESA. 7
1.6 Ley de dispersión de los instrumentos BP/RP. Por cortesía de ESA. . . . 8
1.7 Eficiencia de transmisión de las bandas fotométricas G (luz sin pasar por ninguno de los instrumentos), y tras pasar por los instrumentos BP, RP y RVS en función de la longitud de onda. Por cortesía de ESA. 8
1.8 Listado no exhaustivo de las diferentes ramas y métodos de la estadística. 10
1.9 Listado no exhaustivo de las diferentes ramas y métodos de la AI. 11
1.10 Listado no exhaustivo de las diferentes ramas y métodos del Aprendizaje Máquina. 12
1.11 Estructura organizativa del consorcio DPAC. 14
1.12 Flujo de datos entre paquetes de trabajo en la CU8. 17

2.1 Variación del espectro RVS (sin ruido) en función de sus APs para estrellas FGK. La variación de un parámetro se muestra fijando los valores del resto de APs. Los valores fijados son $T_{eff} = 5500K$, $log\, g = 4$ dex, $[Fe/H] = 0$ dex y $[\alpha/Fe] = 0$ dex. 24
2.2 Variación del espectro RVS en función de sus APs para estrellas A. La variación de un parámetro se muestra fijando los valores del resto de APs. Los valores fijados son $T_{eff} = 9500K$, $log\, g = 4$ dex y $[Fe/H] = 0$ dex . . 25
2.3 Relación de señal a ruido del espectro RVS en función de la magnitud de la estrella. El salto en $G_{rvs} = 10$ se debe al cambio de resolución del instrumento. 26
2.4 Variación de un espectro RVS en función de la magnitud G_{rvs}. 26
2.5 Arquitectura de una ANN feed-forward de tres capas para la estimación de APs 29
2.6 Evolución del *fitness* del algoritmo PSO en función del factor de aprendizaje (lr) y el número de neuronas en la capa oculta (nh), para ANNs entrenadas con espectros de alta y baja resolución. La barra de colores indica el fitness obtenido para cada posición en el espacio de parámetros. 31
2.7 Esquema de filtrado para obtener una representación wavelet multiresolución de una señal dada. 33

2.8 Descomposición wavelet piramidal en tres niveles de un espectro RVS. . . 33
2.9 Percentil 70 de los residuos absolutos para estrellas de tipo FGK, para distintos dominios de entrada. 35
2.10 Percentil 70 de los residuos absolutos para estrellas de tipo A, para distintos dominios de entrada. 36
2.11 Esquema del algoritmo genético de selección de características en el espectro RVS para la estimación de APs. 37
2.12 Arquitectura de una GANN de tres capas para la estimación del espectro RVS a partir de un conjunto de APs 41
2.13 Percentil 70 de los residuos absolutos para estrellas del conjunto FGK, comparativa entre ANNs y GANNs. 42
2.14 Percentil 70 de los residuos absolutos para estrellas del conjunto A, comparativa entre ANNs y GANNs. 43
2.15 Intervalos de confianza (al 70% y 90%) para la estimación de $[Fe/H]$ de estrellas A con $G_{rvs} = 7$. 44
2.16 Intervalos de confianza (al 70% y 90%) para la estimación de $[Fe/H]$ de estrellas A con $G_{rvs} = 11$. 44
2.17 Estimaciones de T_{eff} para estrellas FGK. 46
2.18 Estimaciones de $log\ g$ para estrellas FGK. 47
2.19 Estimaciones de $[Fe/H]$ para estrellas FGK. 47
2.20 Estimaciones de $[\alpha/Fe]$ para estrellas FGK. 48
2.21 Estimaciones de T_{eff} para estrellas A. 48
2.22 Estimaciones de $log\ g$ para estrellas A. 49
2.23 Estimaciones de $[Fe/H]$ para estrellas A. 49

3.1 Muestra de 50 espectros BP/RP para cada librería. Los espectros están escalados a magnitud $G = 15$ 61
3.2 Gráfico de un espectro BP/RP, correspondiente a un quásar con magnitud $G = 15$ (izquierda) y el resultado del preprocesado realizado por OA (derecha). 62
3.3 Evaluación del agrupamiento obtenido mediante el algoritmo k-means, con diversos valores de k, al presentar espectros BP/RP sin ruido. 66
3.4 Diferentes tipos de topologías aplicadas frecuentemente a los mapas auto-organizativos. 67
3.5 Evolución de $sigma(s)$ para varios valores de T, con $sigma(1) = 30$. . . . 70
3.6 Ilustración del proceso de aprendizaje de una red SOM. 70
3.7 Evaluación del agrupamiento obtenido mediante una red SOM, de diversos tamaños, al presentar espectros BP/RP sin ruido. 71
3.8 Visualizaciones estándar de un mapa auto-organizativo. 72
3.9 Visualización mejorada de la matriz U. 74
3.10 Distribución del color $G_{rp} - Gbp$ en el mapa auto-organizativo. 74
3.10 Número de impactos en la red SOM para cada una de las clases supervisadas. 77
3.11 Diagramas obtenidos mediante estadísticas sobre las clases que pueblan cada neurona en la red SOM. 77
3.12 Relación entre la magnitud G de un objeto observado por Gaia y el SNR medio del espectro BP/RP. 78
3.13 Evaluación del agrupamiento obtenido mediante una red SOM 30*30, al presentar espectros BP/RP con varios niveles de ruido. 79

3.14 Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP preprocesados con distintos métodos. 81
3.15 Ejemplo de representación de un espectro BP/RP, correspondiente a un quásar, en el dominio de Fourier, tanto a magnitud $G = 15$ como a magnitud $G = 18.5$. 82
3.16 Descomposición wavelet piramidal en tres niveles de un espectro BP/RP con $G = 15$. 83
3.17 Descomposición wavelet piramidal en tres niveles de un espectro BP/RP con $G = 18.5$. 84
3.18 Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP sin ruido, transformados con diferentes técnicas de procesado de señal. 85
3.19 Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP con $G = 18,5$, transformados con diferentes técnicas de procesado de señal. 85
3.20 Distribución de los grupos de una SOM 30*30, obtenidos mediante agrupamiento conjunto por filtrado, para espectros BP/RP con magnitud 18,5 y un porcentaje $t = 60$. 90
3.21 Transformación de un espectro de SDSS al formato BP/RP mediante el simulador GOG. 91
3.22 Visualización de una red SOM entrenada con espectros atípicos de SDSS. 92
3.23 Detecciones erróneas de objetos en SDSS, encontradas en la región superior izquierda del mapa SOM. 93
3.24 Identificaciones obtenidas mediante el cruce con espectros BP/RP simulados para Gaia. Se aplican varios percentiles para filtrar las clasificaciones no fiables. 95
3.25 Número de impactos por cada una de las clases entre las etiquetas recuperadas de Simbad. 96
3.26 Identificación de los grupos de la red SOM calculada mediante las etiquetas recuperadas de Simbad. 97
3.27 Distribución de los grupos obtenidos mediante agrupamiento conjunto por filtrado, para espectros procedentes de SDSS y $t = 20$. 98
3.28 Diagrama de flujo de datos en el procesado de datos del paquete OA. . . . 101
3.29 Esquema de la implementación del algoritmo de entrenamiento de una red SOM mediante el paradigma MapReduce. 103
3.30 Distancia media por iteración entre la neurona ganadora actual $win(s)$ y la ganadora de la iteración previa $win(s-1)$. 104

A.1 Segmentación de una imagen SAR mediante varias técnicas incluidas en Sentinazos. 107
A.2 Arquitectura general de Sentinazos. 108
A.3 Ejemplos de planos de componentes obtenidos por el IDS basado en SOMs. 109

Lista de Tablas

1.1 Descripción de las CUs presentes en DPAC. 15
1.2 Clases de objetos predefinidas en la misión Gaia. 17

2.1 Grillas simuladas para el entrenamiento de algoritmos en GSP-Spec. . . . 24
2.2 Comparación de resultados obtenidos en la estimación de APs mediante ANNs aplicadas a datos completos, con respecto a los resultados obtenidos mediante ANNs aplicadas a la selección de características obtenida mediante algoritmos genéticos. 38
2.3 Selección de la mejor configuración en función del tipo de estrella y la magnitud. 45

3.1 Número de objetos por cada librería seleccionada del conjunto de simulaciones de Gaia. 60
3.2 Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia sin ruido. 72
3.3 Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia con $G = 20$. 79
3.4 Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia sin ruido, cuando se le presentan los detalles wavelet de nivel 1. 84
3.5 Resultados obtenidos por el método de agrupamiento conjunto. Particiones obtenidas mediante una red SOM 30*30 que agrupa representaciones de espectros simulados BP/RP. 88
3.6 Resultados obtenidos por el método de agrupamiento conjunto por filtrado. Particiones obtenidas mediante una red SOM 30*30 que agrupa representaciones de espectros simulados BP/RP. 89
3.7 Matriz de confusión obtenida mediante el agrupamiento conjunto por filtrado, con magnitud $G = 15$ y porcentaje $t = 60$. 89
3.8 Número de objetos por cada tipo de etiqueta entre las recuperadas de Simbad. La etiqueta UNKNOWN se refiere a objetos para los cuales no se ha encontrado correspondencia. 96
3.9 Matriz de confusión obtenida mediante la distribución de etiquetas recuperadas de Simbad sobre la red SOM entrenada con espectros atípicos de SDSS. 97
3.10 Matriz de confusión obtenida mediante el agrupamiento conjunto de espectros atípicos de SDSS. 98

Acŕonimos

AI	**A**rtificial **I**ntelligence
AP	**A**strophysical **P**arameter
APSIS	**A**strophysical **P**arameters **I**nference **S**ystem
ANN	**A**rtificial **N**eural **N**etwork
ARD	**A**bsolute **R**esidual **D**eviation
ART	**A**daptative **R**esonance **T**heory
BP	**B**lue **P**hotometer
CCD	**C**harge **C**oupled **D**evice
CESGA	**C**entro de **S**upercomputación de **G**alicia
CI	**C**onfidence **I**nterval
CU	**C**oordination **U**nit
DPAC	**D**ata **P**rocessing and **A**nalysis **C**onsortium
DPACE	**D**ata **P**rocessing and **A**nalysis **C**onsortium **E**xecutive
DSC	**D**iscrete **S**ource **C**lassifier
DU	**D**evelopment **U**nit
ECSS	**E**uropean **C**ooperation for **S**pace **S**tandardization
ESO	**E**uropean **S**outhern **O**bservatory
ESA	**E**uropean **S**pace **A**gency
ESP	**E**xtended **S**tellar **P**arametrizer
FLAME	**F**inal **L**uminosity **A**ge and **M**ass **D**etermination
GSP	**G**eneral **S**tellar **P**arametrizer
ICA	**I**ndependent **C**omponent **A**nalysis
ICRS	**I**nternational **C**elestial **R**eference **S**ystem
KDD	**K**nowledge **D**iscovering in **D**atabases.
KNN	**K**-**N**earest in **N**eighbors.

LSST	**L**arge **S**ynoptic **S**urvey **T**elescope.
HES	**H**amburg/**E**SO **S**urvey
HR	**H**igh **R**esolution
LR	**L**ow **R**esolution
OA	**O**utlier **A**nalysis
OCA	**O**bject **C**luster **A**nalysis
MAR	**M**ean **A**bsolute **R**esidual
MDB	Gaia **M**ain **D**atabase
MLP	**M**ulti**l**ayer **P**erceptron
MPI	**M**essage **P**assing **I**nterface
MSC	**M**ultiple **S**tar **C**lassifier
MSE	**M**ean **S**quared **E**rror
PCA	**P**rincipal **C**omponent **A**nalysis
PSO	**P**article **S**warm **O**ptimization
QSOC	**Q**ua**s**ar **O**bject **C**lassifier
RP	**R**ed **P**hotometer
RVS	**R**adial **V**elocity **S**pectrometer
SDSS	**S**loan **D**igital **S**ky **S**erver
SNR	**S**ignal to **N**oise **R**atio
SOM	**S**elf **O**rganizing **M**ap
SVM	**S**upport **V**ector **M**achine
TGE	**T**otal **G**alactic **E**xtinction
UGC	**U**nresolved **G**alaxy **C**lassifier

Nunca ha cogido un avión ni se ha ido de crucero, pero ha hecho cosas más importantes en la vida como criar a cinco hijos y darles un futuro. Esta tesis va dedicada a él, mi padre, el mejor ciclista de Catabois de todos los tiempos. Va dedicada también al resto de mi familia

Capítulo 1

Introducción

1.1 Breve historia de la astronomía

La astronomía es la ciencia que trata el estudio de los cuerpos celestes del Universo. Su historia es tan antigua como la del ser humano, ya que todas las civilizaciones se han preguntado cuál es nuestro lugar en el Universo y, al mismo tiempo, han estudiado la posición y movimiento de las estrellas. Los antiguos griegos fueron de los primeros en hacer contribuciones importantes a la astronomía, mediante pensadores como Aristóteles que formalizó el primer modelo astronómico o Hiparco de Nicea, quién elaboró el primer catálogo de estrellas. Más adelante, sobrevino una época poco fructífera para el campo, en la cual el modelo geocéntrico, en el cual el Sol y demás planetas giraban alrededor de la Tierra, seguían tomándose como el verdadero. Sería Nicolás Copérnico el que propusiera por primera vez el modelo heliocéntrico en el siglo XVI, el cual fue comprobado mediante las observaciones con telescopio de Galileo Galilei. Posteriormente, el modelo heliocéntrico fue refinado por Johannes Kepler, quien determinó los movimientos elípticos de los planetas, a través de las precisas mediciones que había obtenido su maestro Tycho Brahe.

En el siglo XVII, la ley de gravitación universal de Newton revolucionó por completo la astronomía, ya que permitía explicar de forma matemática el movimiento de los diferentes cuerpos celestes. Newton también fue el inventor del telescopio refractor y el descubridor del espectro de color obtenido cuando la luz pasa por un prisma, el cual es inherente a la misma. De esta forma, nació la astrofísica como una ciencia que estudia los procesos físicos que experimentan los diferentes objetos astronómicos.

En el siglo XIX, se descubrió que, en los espectros de las estrellas, en concreto en el del Sol, se pueden observar multitud de líneas de absorción de la luz. Más tarde, se observó

que el espectro producido por gases conocidos en la Tierra, como el hidrógeno, contenían algunas de las líneas del espectro del Sol. La misma observación puede aplicarse a otros elementos químicos, cuya presencia es la causa de distintas lineas en el espectro, las cuales pueden observarse también en espectros de estrellas, cuando estas contienen dichos elementos en su atmósfera. De esta forma, mediante el estudio de los espectros estelares, se pudo determinar varias de sus propiedades intrínsecas, como su temperatura, masa y luminosidad. Además, el estudio de los gases atmosféricos estelares permitió descubrir gases antes desconocidos en la Tierra, como por ejemplo el helio, cuyo nombre fue propuesto en honor al dios griego del sol (*Helios*).

A principios del siglo XX, tanto Ejnar Hertzsprung como Henry Norris Russell descubrieron, de forma independiente, la dependencia que muestran la luminosidad de las estrellas conocidas con respecto a su color o temperatura, mediante la representación gráfica de ambas propiedades en un diagrama. Dicho diagrama se denomina hoy en día como diagrama Hertzsprung-Russell (HR), y se muestra en la figura 1.1. En dicho diagrama, la gran mayoría de las estrellas se situaban a lo largo de una región central, que se denominó la "Secuencia Principal". Además, se descubrieron dos grupos que se salen de la Secuencia, llamados gigantes y enanas blancas. La Secuencia Principal marca tanto la luminosidad como la temperatura típicas para las estrellas en función de su masa inicial, desde que se forman hasta que su brillo se estabiliza. Dependiendo de su masa inicial, las estrellas salen de la secuencia en un punto u otro para convertirse en gigantes o supergigantes rojas. Las más masivas abandonan la fase de gigante explotando como supernovas, y, en algunos casos, derivando en un remanente estelar compacto, como una estrella de neutrones o agujero negro, hecho que sería descubierto más adelante. Por otro lado, las estrellas de menor masa pierden su atmósfera lentamente para terminar convirtiéndose en una enana blanca.

En los años 20, telescopios de mayor diámetro colector y resolución angular, permitieron a los astrónomos determinar que la Vía Láctea no era si no la luz acumulada de millones de estrellas en el disco de nuestra Galaxia. Además, el astrónomo Edwin Hubble comenzó a medir las distancias a las galaxias cercanas, por aquel entonces denominadas nebulosas, mediante la periodicidad de las estrellas variables cefeidas. Dichas distancias situaban a esas galaxias mucho más allá de la Vía Láctea. Simultáneamente, se observó que las galaxias contenían lineas espectrales que estaban desplazadas al extremo rojo del espectro, de acuerdo con el efecto Doppler, más desplazadas cuanto más lejanas se situaban las mismas. Como consecuencia, se postuló la ley de Hubble, que relaciona distancias y corrimiento al rojo, y que, al cotejarla con la Teoría de la Relatividad de Einstein, condujo a la conocida Teoría del Big Bang, que postula que el Universo comenzó como un punto infinitamente pequeño y de infinita densidad, momento a partir del cual

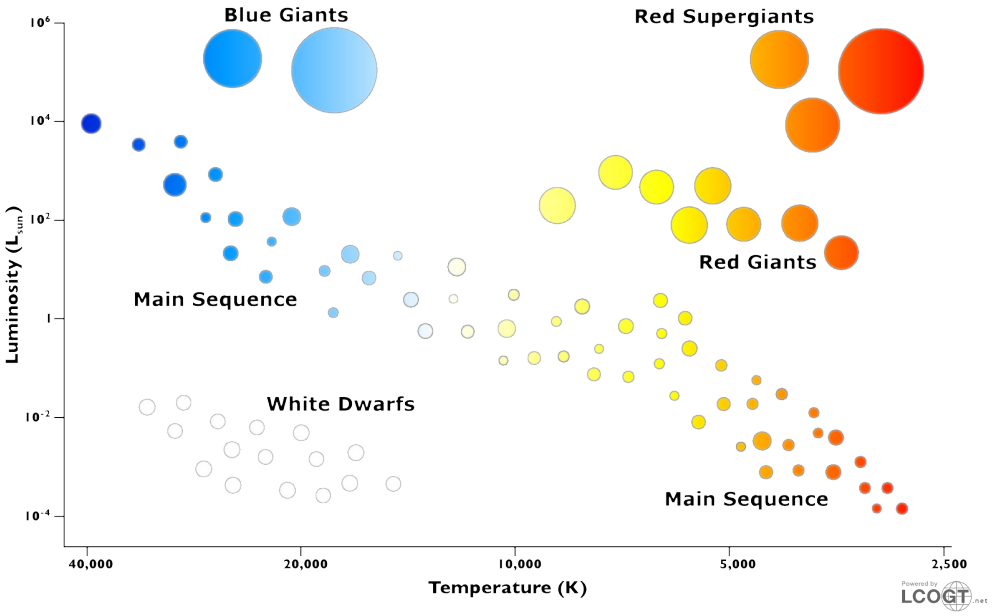


FIGURA 1.1: Diagrama de Hertzsprung-Russell. Extraído de `http://lcogt.net`.

comenzó a expandirse, proceso que continúa en el presente. Nacía así la cosmología moderna.

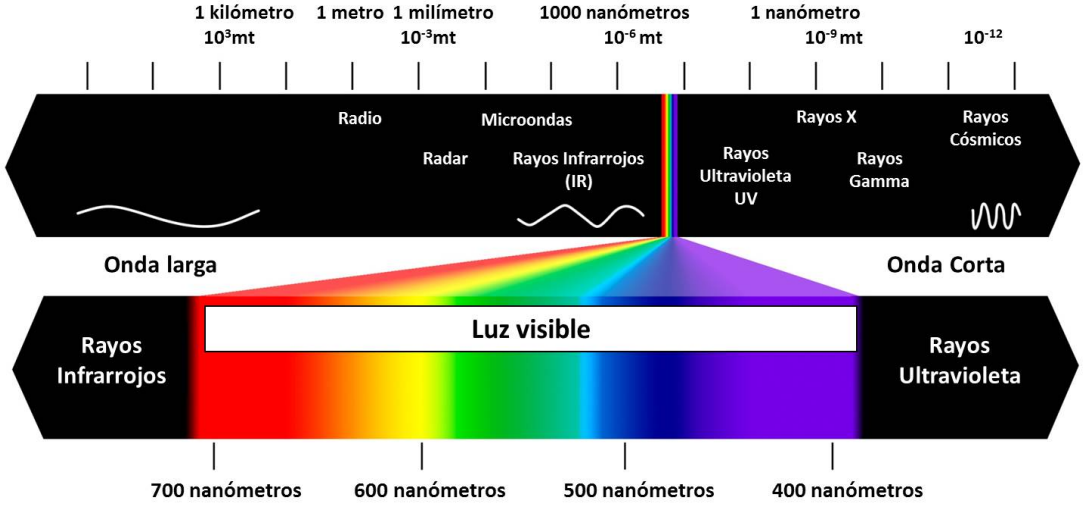


FIGURA 1.2: Representación del espectro electromagnético. Extraído de `http://www.artinaid.com/`.

A lo largo del siglo XX, las mejoras en los telescopios, en las técnicas de observación y los avances en física cuántica, trajeron consigo el descubrimiento de objetos exóticos, como las estrellas de neutrones o los agujeros negros. Evidentemente, los agujeros negros no son observables directamente, pero sabemos que están ahí por el efecto que producen en las estrellas cercanas. Además, el estudio del espectro electromagnético (ver figura 1.2) de unos objetos especiales, llamados quásares, determinó que el único proceso físico que

podía generar objetos tan intrínsecamente brillantes era el provocado por un agujero negro al engullir masa de estrellas cercanas. Dichos objetos emiten radiación en líneas espectrales con un corrimiento al rojo mayor que ningún otro tipo de objeto (exceptuando algunas galaxias), lo que les caracteriza como los objetos más distantes, brillantes y antiguos del Universo conocido.

Ya en el siglo XXI, gracias al rápido avance de las tecnologías digitales, la astronomía, así como la física en general y otros campos científicos como la biología y la química, se ha convertido en una ciencia en la que se utiliza y extrae conocimiento de información proveniente de enormes bases de datos, i.e. una "ciencia intensiva de datos". Los telescopios construidos en la última década son capaces de realizar un escaneo automático y continuado del cielo, capturando varias imágenes por segundo de una amplia región del cielo. Además, las nuevas cámaras CCD superan con creces la eficiencia de los dispositivos anteriores, permitiendo así mejorar la calidad de las imágenes. Esto permite a los astrónomos visualizar objetos que antes eran inaccesibles, debido a que la luz que recibimos de los mismos es muy débil. Todo esto ha provocado una revolución sin precedentes en el campo, como demuestra el éxito científico obtenido por proyectos como el Sloan Digital Sky Survey (SDSS), que ha cambiado nuestra forma de ver el Universo [1]. Sin embargo, esta explosión de datos lleva emparejada un coste, que viene dado por el incremento en la dificultad de su análisis.

Las bases de datos astronómicas están compuestas por un conjunto heterogéneo de mediciones sobre los astros, provenientes de modo directo o indirecto del análisis de su luz. Entre las diferentes técnicas astronómicas, la astrometría tiene como objetivo medir con precisión las posiciones de los astros en el cielo, así como sus movimientos relativos. Por otro lado, la fotometría mide el flujo de luz emitido por los astros en una o más bandas del espectro electromagnético (ver figura 1.2), con el objetivo de obtener información sobre las propiedades intrínsecas de los mismos. Para ello, se aplican una serie de filtros que bloquean la luz recibida en bandas distintas a la deseada. Por último, la espectroscopia se encarga de dispersar la luz recibida, de forma que se obtiene una medición del flujo emitido en una secuencia de longitudes de onda. La dispersión de la luz se realiza mediante diferentes dispositivos, que van desde un simple prisma a una red de difracción holográfica, específicamente diseñada para la tarea. Estas tres técnicas de medición no son independientes, sino que están relacionadas y se complementan entre sí. Por ejemplo, la espectroscopia se utiliza para medir las velocidades radiales de los astros.

Actualmente se pueden encontrar bases de datos astronómicas con millones de objetos, conteniendo para cada uno de ellos imágenes, espectros y mediciones diversas, motivo por el cual estos archivos suelen contener cantidades ingentes de datos, del orden de

varios terabytes. Para la segunda década de siglo, están en proyecto varios catálogos astronómicos que superarán el petabyte de datos, como es el caso de la misión espacial de cartografiado galáctico Gaia o el Large Synoptic Survey Telescope (LSST).

Hay varias preguntas abiertas en astronomía y astrofísica. Aquí daremos una visión general de las mismas sin entrar en detalles, ya que una discusión profunda queda fuera de la temática de la tesis. Se ha descubierto un tipo de materia que interactúa gravitacionalmente con los objetos astronómicos pero que no se puede ver, de ahí que se denomine materia oscura, véase [2]. Adicionalmente se ha observado una expansión acelerada del Universo debido a algún tipo de energía oscura desconocida. De hecho, se calcula que la energía oscura representa un 75% del contenido energético del Universo, la materia oscura un 20% y la materia bariónica (la visible por nosotros) tan sólo un 5% del mismo. Por otro lado, la formación, evolución y estructura de nuestra galaxia no se conoce por completo, véase [3]. Estas incógnitas y muchas otras podrían esclarecerse si, y sólo si, la comunidad científica tiene éxito en el análisis de los datos producidos por las próximas misiones astronómicas.

1.2 La misión Gaia

El satélite Gaia es una de las misiones clave de la agencia espacial europea (ESA, por sus siglas en inglés), la cual realizará un censo de la Vía Láctea con una precisión sin precedentes. Su lanzamiento ha sido realizado con éxito el pasado 19 de diciembre. Está previsto que Gaia realice observaciones de todos los objetos visibles con magnitudes $20 > G > 6$, con un número de objetos estimado de mil millones, que representan sobre un 1% de los objetos de la Galaxia. De esta forma, se podrá elaborar un mapa 3D detallado que nos permitirá responder diversas cuestiones abiertas acerca de la composición, formación y evolución de la Vía Láctea. Gaia no observará exclusivamente estrellas, si no todo tipo de objeto astronómico con brillo aparente dentro de los límites del satélite, bien sea galáctico (asteroides, cometas), o extragaláctico (otras galaxias, supernovas, quásares).

Gaia se compone de 2 telescopios espaciales, separados por $106.5\,^{\circ}$, que actúan como si fuera uno sólo, ya que comparten el mismo plano focal. La luz colectada por los telescopios recorre un total de 35 metros a través de varios espejos antes de alcanzar el plano focal. Los espejos de Gaia son relativamente pequeños en comparación con los que se pueden encontrar en los telescopios terrestres, pero tienen la ventaja de operar en el espacio, donde la atmósfera no distorsiona las imágenes. Para cubrir el cielo en su totalidad, Gaia gira lentamente completando 4 rotaciones al día. Además, sigue una órbita alrededor del Sol, que coincide con el punto dos de Lagrange (L2), el cual es un lugar geométrico de gran estabilidad en el que se sitúan muchos de los satélites

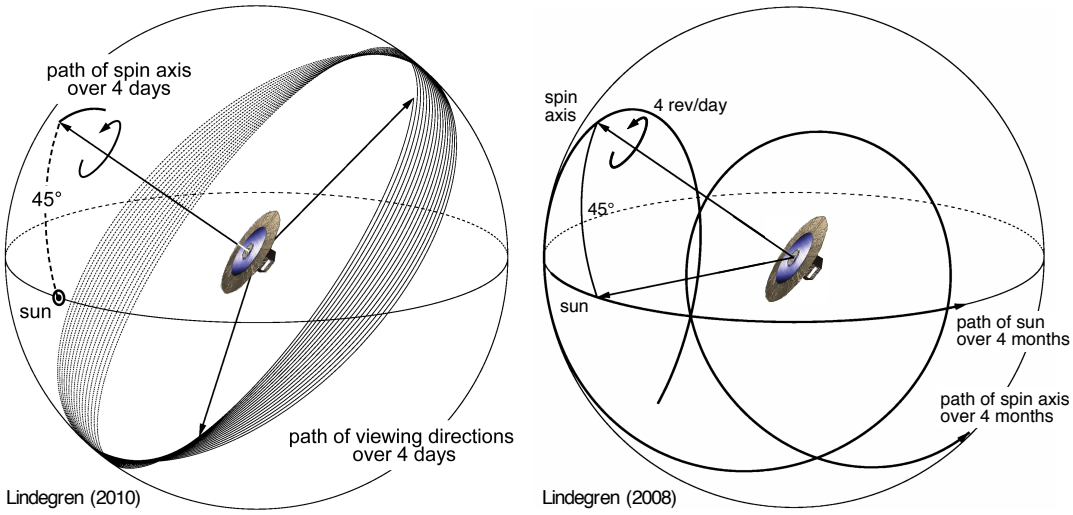


FIGURA 1.3: Rotación de los campos de visión de Gaia. La precesión del eje de giro permite mapear diversas zonas del cielo. Por cortesía de ESA.

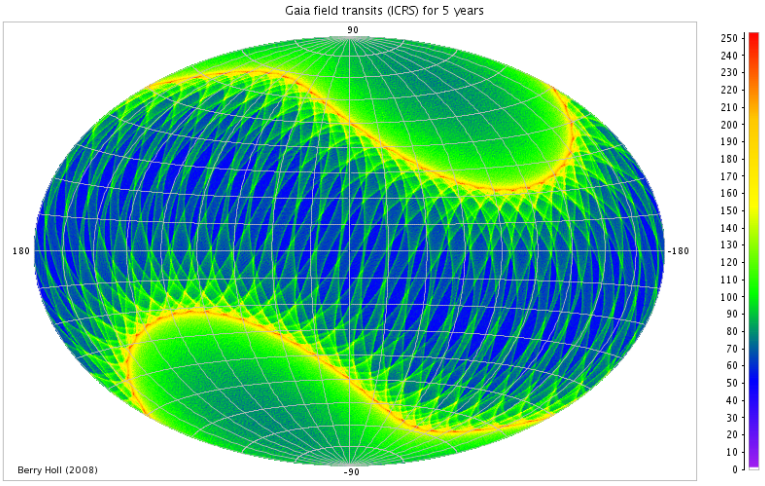


FIGURA 1.4: Distribución de tránsitos del satélite Gaia después de 5 años de misión. Coordenadas ICRS. Por cortesía de ESA.

artificiales, ya que en él se equilibran las fuerzas gravitatorias del sistema Sol-Tierra. En las cercanías de esta posición, el satélite estará además relativamente resguardado y protegido de la radiación solar, recibiendo sin embargo suficiente energía para alimentar sus paneles solares. De esta forma cubrirá sistemáticamente distintos puntos del cielo, como se muestra en la figura 1.3. Después de 5 años de misión, todas las estrellas serán observadas en 72 ocasiones como media, mejorando así la calidad de las observaciones y obteniendo información sobre variabilidad estelar. La figura 1.4 muestra la distribución de tránsitos tras 5 años de misión.

En el plano focal de Gaia se sitúan varios instrumentos que recibirán la luz de los objetos detectados en el cielo, como se muestra el la figura 1.5. La distribución de los CCDs dedicados a diversas funciones e instrumentos se muestra también en dicha figura. Las

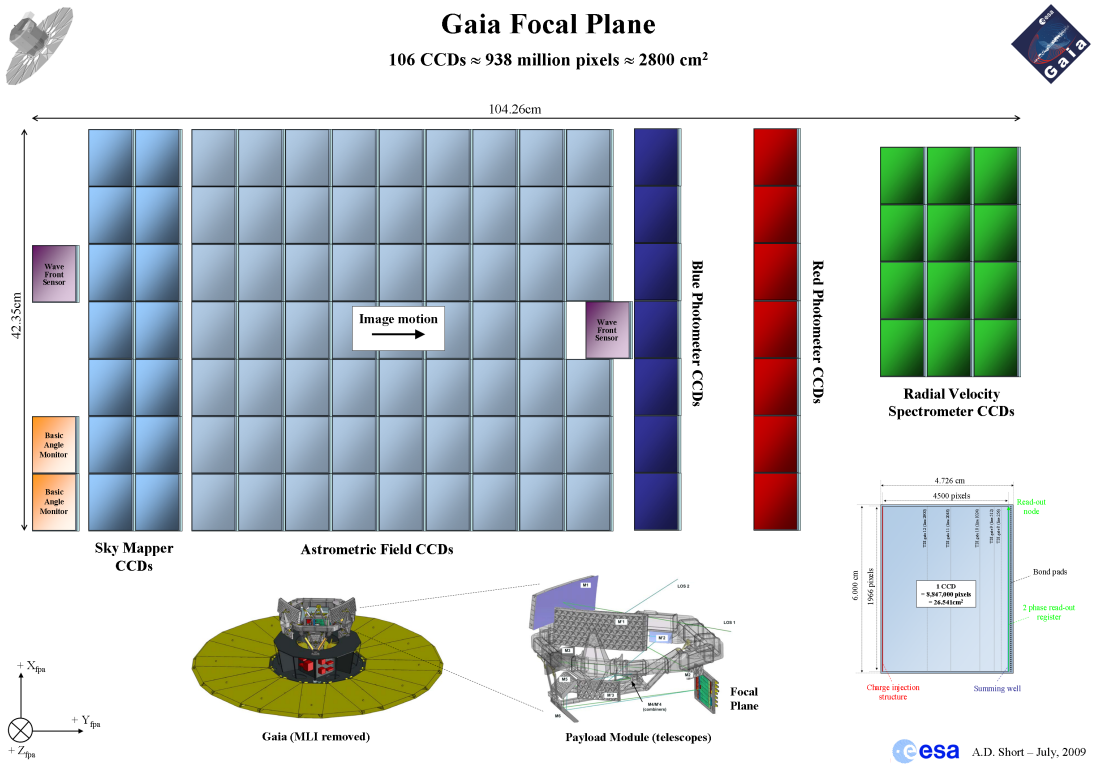


FIGURA 1.5: Instrumentos montados en el satélite Gaia. Por cortesía de ESA.

estrellas observadas en el campo de visión combinado de los dos telescopios entran en el plano focal de izquierda a derecha. Los primeros CCDs, cuyo conjunto se denomina *SkyMapper*, están dedicados a detectar los objetos en tiempo real, determinando sus posiciones y brillo. A continuación, los objetos detectados llegan a los 62 CCDs del campo astrométrico, el cual recoge toda la luz entre 330 y 1050 nm, lo cual se define como la banda G de Gaia. Esta zona del detector permite calcular con precisión, tras varias observaciones, las posiciones angulares celestes, el movimiento propio en sus dos direcciones y la paralaje, de la cual se extrae la distancia a la estrella. Antes de abandonar el plano focal, el espectro de cada objeto de extrae mediante tres conjuntos de CCDs. Los CCDs llamados BP (Blue Photometer) y RP (Red Photometer) recogen la luz tras hacerla pasar por dos prismas sensibles a las bandas azul (330-680 nm) y roja (640-1050 nm) del espectro electromagnético visible, respectivamente. Los prismas dispersan la luz produciendo espectros de baja resolución, de 3 a 27 nm en el caso de BP y de 7 a 15 nm en el caso de RP, siguiendo la ley de dispersión de la figura 1.6. De la información presente en los espectros BP/RP se podrán extraer parámetros astrofísicos fundamentales, como son la temperatura, gravedad, metalicidad y enrojecimiento de las estrellas. Por último, el instrumento RVS obtendrá espectroscopia entre 847 y 871 nm, con una resolución de R=11500 en modo de alta resolución y R=5000 en modo de baja resolución, para las estrellas relativamente brillantes entre todas las observadas por Gaia, con magnitudes V menor que 17. Esto permitirá obtener el componente de

velocidad radial, además de parámetros astrofísicos más precisos para las estrellas más brillantes. En resumen, el plano focal contiene los detectores de cuatro instrumentos, cada uno recogiendo luz en una banda determinada, como se muestra en la figura 1.7. La complejidad de esta instrumentación permite que Gaia realice observaciones astrométricas, fotométricas y espectroscópicas simultáneamente.

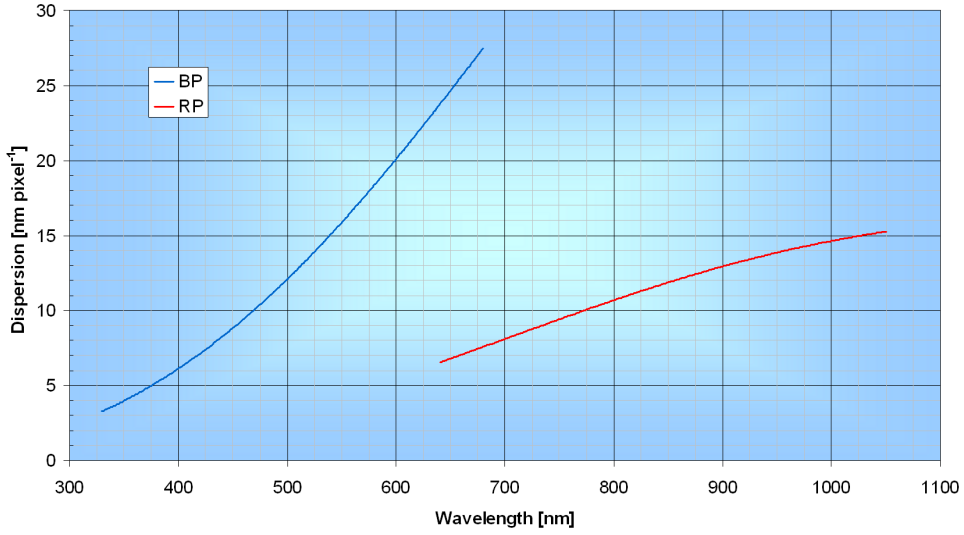


FIGURA 1.6: Ley de dispersión de los instrumentos BP/RP. Por cortesía de ESA.

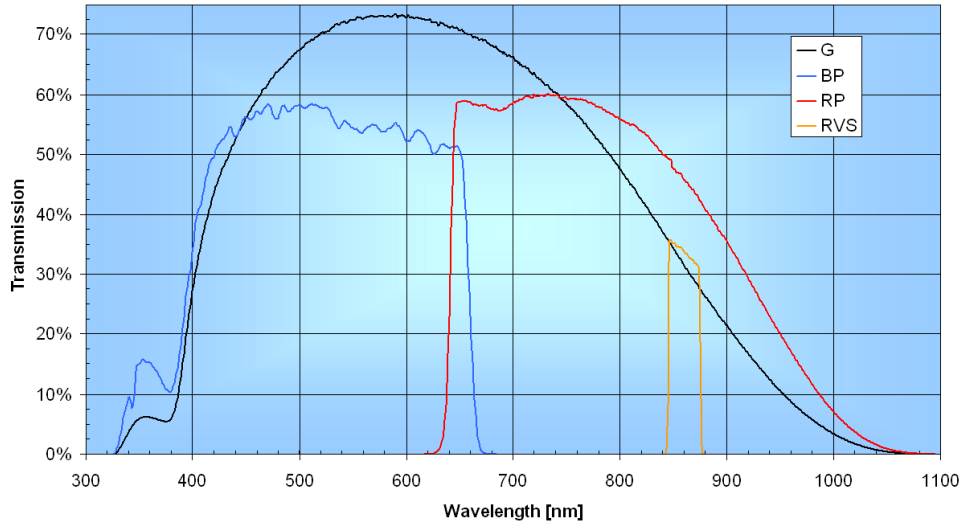


FIGURA 1.7: Eficiencia de transmisión de las bandas fotométricas G (luz sin pasar por ninguno de los instrumentos), y tras pasar por los instrumentos BP, RP y RVS en función de la longitud de onda. Por cortesía de ESA.

El principal objetivo de Gaia es realizar astrometría de alta precisión, alcanzando el microsegundo de arco, lo cual representa una mejora de dos ordenes de magnitud con respecto a su predecesor, el satélite Hipparcos. Adicionalmente, al contar con instrumentos fotométricos y un espectrómetro de resolución intermedia, la funcionalidad del satélite va mucho más allá de este objetivo primario, obteniendo información astrofísica de una gran cantidad de objetos, muchos de los cuales no habían sido observados previamente.

Con los datos de Gaia se podrán realizar diversos experimentos científicos de enorme interés y relevancia para la física y astrofísica actual, desde precisar los modelos y expectativas de la teoría de Evolución Estelar, hasta el estudio de la materia oscura, pasando por el estudio de la distribución dinámica y química de nuestra galaxia. Todo ello será posible si, y sólo si, la comunidad científica es capaz de afrontar el gran desafío que supone el procesado automático de la enorme cantidad de datos obtenidos por la misión.

1.3 Estadística, inteligencia artificial, aprendizaje máquina y minería de datos

En los últimos años, las tecnologías desarrolladas en la reconocida como "era de las TIC (Tecnologías de la Información y las Comunicaciones)", han incrementado nuestra capacidad para obtener y almacenar grandes cantidades de datos. A raíz de ello, diversos campos relacionados con el análisis de datos, como son la inteligencia artificial, la estadística, la minería de datos y el aprendizaje máquina han recibido un gran impulso. Este impulso proviene de la necesidad de analizar la gran cantidad de información que se ha comenzado a recoger en ámbitos diversos, como las redes sociales, los negocios empresariales, la biología o la astronomía. Sin embargo, existen grandes solapamientos entre los campos de análisis de datos, lo cual provoca confusiones frecuentemente. A continuación, se discuten las similitudes y diferencias entre estos 4 campos:

- **Estadística** se define como la ciencia de la recolección, análisis e interpretación de conjuntos de datos. Por lo tanto es un concepto transversal que atañe a una gran cantidad de campos. Nació en el siglo XVIII de manos de Gauss y Legendre. Cabe diferenciar entre la estadística descriptiva y la estadística inferencial. La estadística descriptiva se encarga de la visualización y resumen de conjuntos de datos. Por otro lado, la estadística inferencial se encarga de la generación de modelos en función de un conjunto de observaciones. La figura 1.8 muestra un esquema simplificado de las distintas técnicas pertenecientes a la estadística.

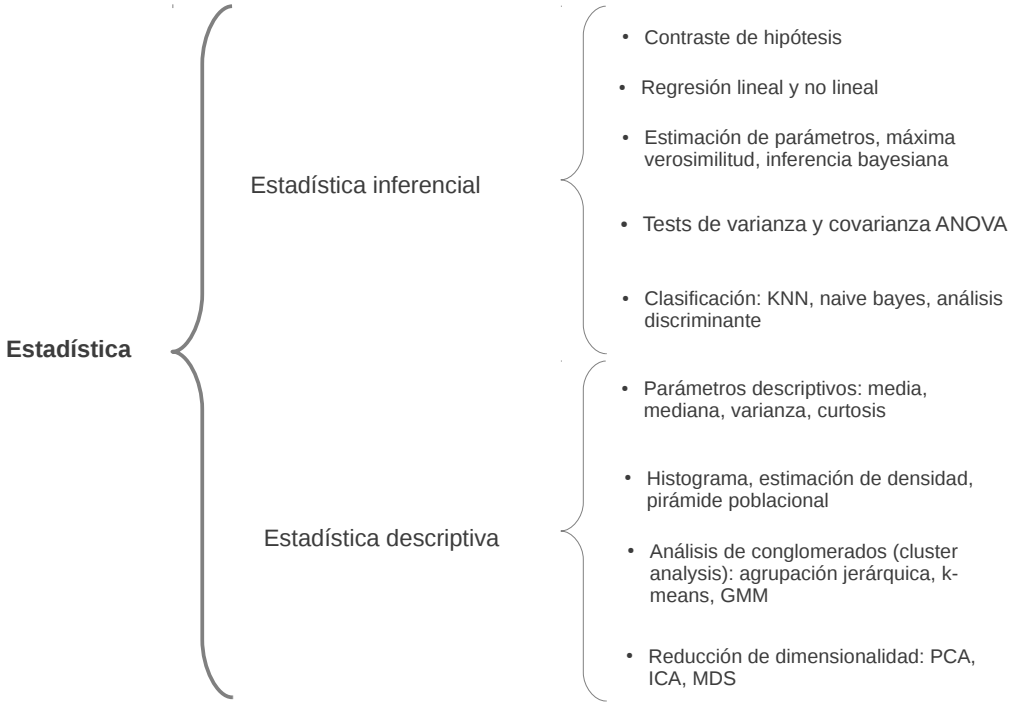


FIGURA 1.8: Listado no exhaustivo de las diferentes ramas y métodos de la estadística.

- **Inteligencia artificial** se refiere a la capacidad que se le otorga a una máquina para solucionar problemas que requieren inteligencia, de forma similar a como lo haría un humano entrenado para ello. Dichos problemas no tienen una solución analítica directa, por lo que se requiere la obtención de una solución aproximada que sea aceptable. El concepto Artificial Intelligence (AI) se acuñó en los años 60, siendo sus principales precursores Turing, McCulloch, Pitts, McCarthy, Minsky y Shannon. Dentro del campo de la inteligencia artificial se enmarcan varias técnicas, como los algoritmos de búsqueda heurística, las redes de neuronas artificiales, los algoritmos genéticos y los sistemas de razonamiento basados en reglas, así como el reconocimiento de patrones. Aparte de los métodos de aprendizaje máquina, la AI también incluye métodos de selección y extracción de características. En general, la AI incluye todo método necesario para conseguir que una máquina solucione un problema complejo de forma inteligente. La figura 1.9 muestra un esquema simplificado de las distintas técnicas pertenecientes a la AI.

- **Aprendizaje máquina** es el proceso mediante el que una máquina aprende a desarrollar una determinada tarea a través de ejemplos. Por lo tanto se trata de una rama de la AI. Podemos distinguir entre aprendizaje supervisado, en el cual se le ofrece a la máquina el resultado deseado (llamado etiqueta) con cada ejemplo de entrenamiento y aprendizaje no supervisado, donde las etiquetas no

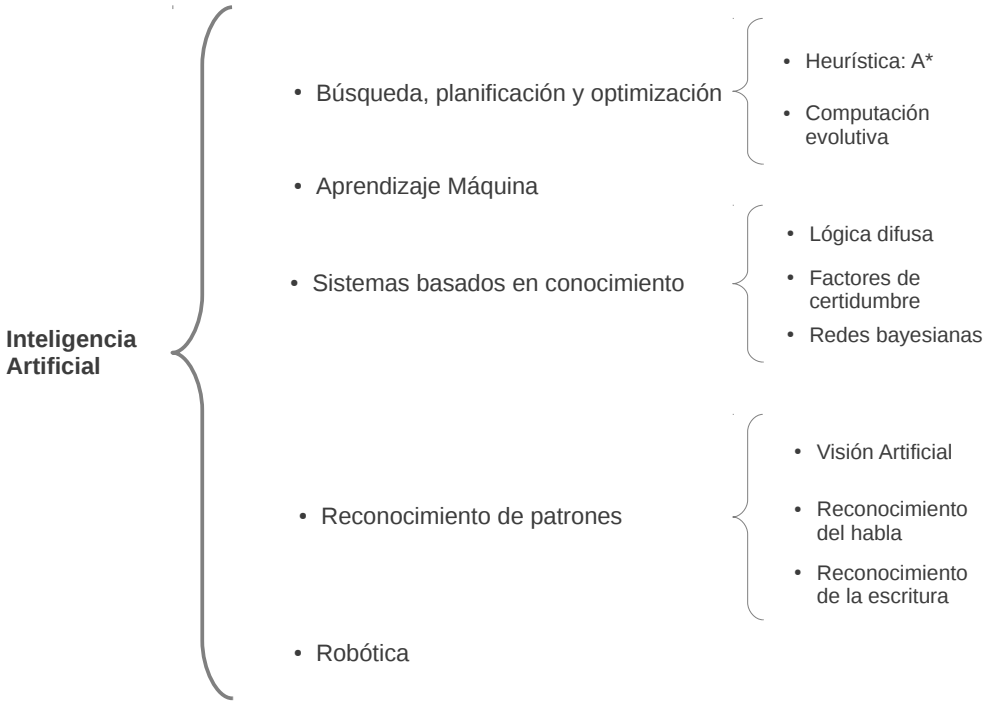


FIGURA 1.9: Listado no exhaustivo de las diferentes ramas y métodos de la AI.

le son mostradas a la máquina. Una vez la máquina ha aprendido la función deseada a través de los ejemplos, ésta se aplica para predecir las etiquetas de nuevos ejemplos presentados a la misma. En los últimos años, se han desarrollado un conjunto de modelos de aprendizaje que tienen una base probabilística, dirigida por los datos, con los cual este campo se ha acercado más a la estadística. La figura 1.10 muestra un esquema simplificado de las distintas técnicas pertenecientes al aprendizaje máquina.

- **Minería de datos** es una disciplina que se encarga de los procesos necesarios para la obtención de conocimiento en grandes bases de datos. También recibe el nombre de "extracción de conocimiento en bases de datos" o KDD, por sus siglas en inglés. El término fue acuñado en los años 90, cuando las grandes corporaciones comenzaron a almacenar sus datos en formato digital. La minería de datos incluye métodos provenientes de disciplinas como la estadística, la AI, las bases de datos y la teoría de la complejidad computacional. La minería de datos contempla métodos de aprendizaje máquina tanto supervisados como no supervisados, aunque se enfoca en métodos que generen modelos comprensibles, de forma que se pueda extraer fácilmente conocimiento, como son los árboles de decisión o los sistema de inducción de reglas.

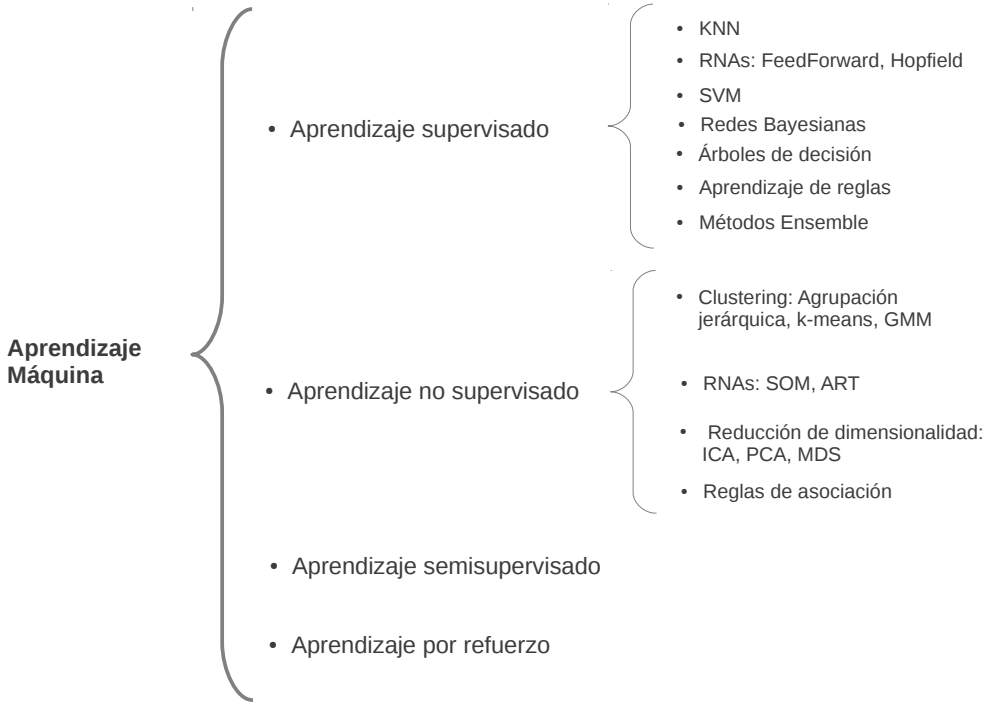


FIGURA 1.10: Listado no exhaustivo de las diferentes ramas y métodos del Aprendizaje Máquina.

Como puede observarse en las definiciones, estos campos se superponen ampliamente. Por ejemplo, las técnicas de agrupamiento como k-means son al mismo tiempo técnicas de estadística descriptiva, minería de datos y aprendizaje no supervisado, dependiendo de su uso. Además, muchas técnicas de aprendizaje supervisado se basan en modelos estadísticos. Sin embargo, otras técnicas son exclusivas de un campo, por ejemplo los algoritmos genéticos son algoritmos genuinos de la AI, mientras que las reglas de asociación son una técnica principalmente relacionada con la minería de datos. Lo importante a la hora de resolver un problema que requiera de un procesado automático inteligente es escoger adecuadamente qué técnicas usar y cómo aplicarlas en el dominio. No existe una técnica que supere a las demás en todos los casos, sino que en cada caso de aplicación debe evaluarse cuáles son más efectivas. De hecho, existe un teorema denominado "No Free Lunch" que demuestra que todos los algoritmos obtendrían una precisión parecida si se aplicasen a todos los posibles problemas.

En el capítulo 2 nos centraremos en la obtención de parámetros astrofísicos mediante redes de neuronas artificiales. Se trata de un problema conocido como *Regresión* en los campos de aprendizaje máquina y estadística inferencial, que trata de aprender la función que relaciona un conjunto de variables predictoras X con un conjunto de variables a predecir Y. Por otro lado, en el capítulo 3, se abordará el problema de la

minería de datos en bases de datos astronómicas. En dicho capítulo se contemplarán técnicas tanto de AI como estadísticas para obtener un modelo que simplifique un gran conjunto de datos de una forma óptima. Estas técnicas se basan en el agrupamiento de datos (*clustering* en inglés) y en la reducción de dimensionalidad. Así, se desarrollarán varias técnicas de visualización y navegación de los datos diseñadas con el objetivo de ayudar a los astrofísicos en el proceso de extracción de conocimiento mediante la base de datos obtenida con Gaia.

1.4 El consorcio DPAC

Con el objetivo de llevar a cabo el procesado de los datos que se obtendrán con la misión Gaia, se ha organizado un gran consorcio a nivel europeo llamado Gaia Data Processing and Analysis Consortium (DPAC). Dicho consorcio está formado por unos 400 científicos e ingenieros repartidos por toda la región europea. DPAC se ha organizado en diversas unidades de coordinación, llamadas Coordination Units (CUs), que a su vez pueden dividirse en paquetes de trabajo llamados Development Units (DUs). Como se puede observar en la figura 1.11, cada CU tiene una serie de funciones bien definidas, las cuales se describen brevemente en la tabla 1.1. Las CUs cuentan con el soporte de una serie de centros de cálculo, llamados Data Processing Centers (DPCs), que se encargan de ofrecer los recursos hardware que requiere la enorme cantidad de cómputo implicada en el procesado, además de dar soporte software a los desarrolladores de las CUs. Finalmente, el DPAC Executive (DPACE) se encarga de coordinar todo el sistema y de evaluar su estado general.

El DPAC ha adoptado una estrategia de desarrollo basada en ciclos de entrega de software y documentación. Cada DU debe especificar sus requisitos, implementar su software, probarlo y documentarlo en cada uno de los ciclos temporales, acordados previamente para el desarrollo. De esta forma la complejidad del producto generado se incrementa con el tiempo. También se genera una versión de la MDB con cada uno de los ciclos. Este esquema será el que se implemente durante la operación real del satélite. Sin embargo, el DPAC ha comenzado los ciclos de desarrollo mucho antes, sobre el año 2008, con el objetivo de garantizar la calidad de los productos finales. En lugar de los datos operacionales, CU2 se encarga de generar una simulación para cada uno de los ciclos de desarrollo previos al lanzamiento.

Aquí nos centraremos en la CU8, en la cual se enmarca esta tesis. CU8 se encargará de clasificar los astros en una serie de clases predefinidas, en función de su naturaleza. Además, CU8 obtendrá los principales parámetros astrofísicos de los objetos clasificados, permitiendo una descripción astrofísica completa de los mismos. Para ello, CU8 se ha

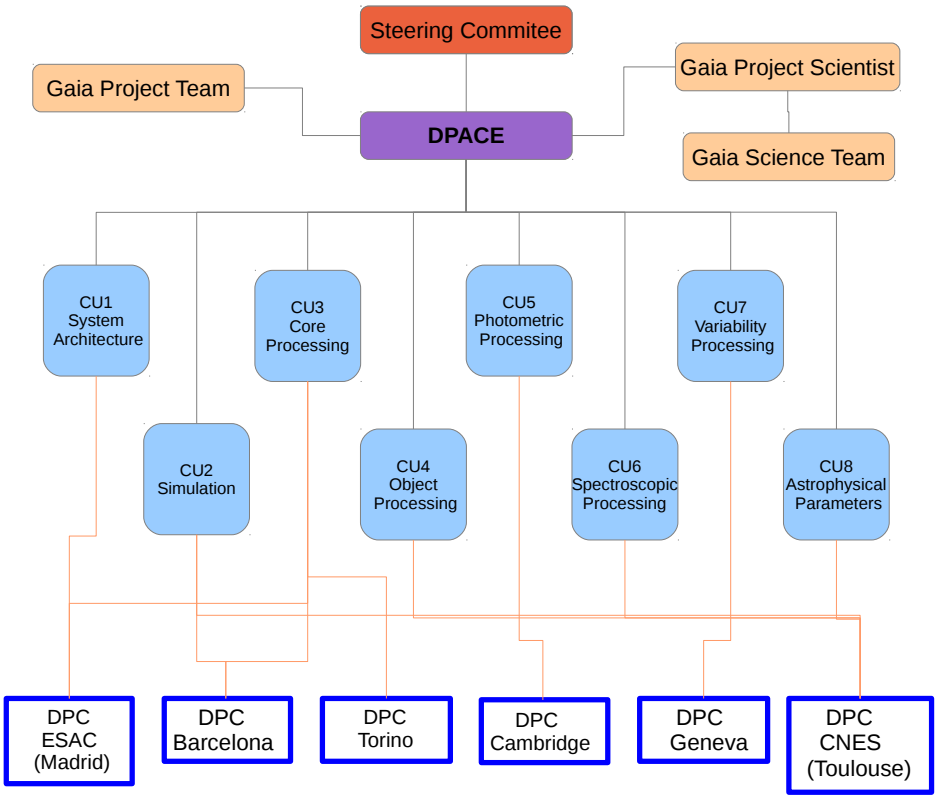


FIGURA 1.11: Estructura organizativa del consorcio DPAC.

dividido en varios DUs, cada uno con tareas bien definidas, las cuales se describen a continuación:

- Discrete Source Classifier (DSC) realiza la clasificación probabilística en un conjunto predefinido de clases, las cuales se describen en la tabla 1.2. Para ello utiliza técnicas de clasificación supervisada, en concreto Support Vector Machines (SVMs). El clasificador tiene una estructura jerárquica, en la cual cada subclasificador se encarga de estimar probabilidades en función de un tipo de datos diferente, bien sea BP/RP, astrometría o variabilidad. Finalmente, las probabilidades de los subclasificadores se combinan para obtener una clasificación combinada. Además, DSC también proporcionará una clasificación más detallada, mediante la cual se podrá distinguir entre subtipos de estrellas. El funcionamiento de DSC se tratará más a fondo en el capítulo 3.

- General Stellar Parametrizer-Photometry (GSP-Phot) determinará los principales parámetros astrofísicos para todas las estrellas mediante espectrofotometría BP/RP. Los parámetros estimados son: temperatura efectiva (T_{eff}), gravedad

Tabla 1.1: Descripción de las CUs presentes en DPAC.

CU	DESCRIPCIÓN
1	Diseño de la base de datos, de la arquitectura software y hardware. Adaptación del estándar de desarrollo ECSS.
2	Simulación detallada de todos los aspectos de la misión Gaia.
3	Monitorización del comportamiento del satélite y sus instrumentos. Obtención de parámetros astrométricos.
4	Procesado astrométrico y fotométrico de objetos complejos: objetos múltiples, extendidos, asteroides, etc.
5	Calibración fotométrica, tratamiento de imágenes, corrección de aberraciones cromáticas.
6	Calibración espectroscópica, determinación de velocidades radiales.
7	Análisis de variabilidad, incluyendo clasificación de estrellas variables y análisis estadístico.
8	Obtención de parámetros astrofísicos y clasificación de fuentes.
9	Validación y divulgación del catálogo obtenido por Gaia. Aplicaciones que faciliten casos científicos con los datos de Gaia.

superficial logarítmica ($log\ g$), metalicidad ($[Fe/H]$) y enrojecimiento (A_0). GSP-Phot se compone de varios algoritmos de regresión supervisada, principalmente de carácter bayesiano, de forma que también se proporcionarán valores de incertidumbre sobre las estimaciones.

- General Stellar Parametrizer-Spectroscopy (GSP-Spec) determinará parámetros astrofísicos para las estrellas más brillantes ($G < 15$), mediante el uso de espectroscopia RVS. Además de los parámetros fundamentales que estima GSP-Phot, también obtendrá la abundancia de elementos alfa ($[\alpha/Fe]$) y la velocidad de rotación ($vsini$). Finalmente, estimará abundancias químicas individuales para estrellas muy brillantes, con magnitud G menor que 12. La estimación de abundancias de elementos químicos ligeros es de gran importancia, ya que permite conocer si las estrellas observadas provienen de otras más antiguas que contenían dichos elementos en su núcleo y que han estallado en una supernova. GSP-Spec se compone de varios algoritmos de diferente índole. El capítulo 2 de esta tesis se centra en el desarrollo de uno de ellos, el cual está basando en ANNs.

- Extended Stellar Parametrizer (ESP) complementa las estimaciones de GSP centrándose en tipos especiales de estrellas, bien sea por tener parámetros extremos o procesos físicos poco convencionales. Los algoritmos de ESP hacen uso de datos

BP/RP, RVS o ambos. Algunos de ellos incorporan algoritmos con un tratamiento físico del espectro, en lugar de algoritmos basados en aprendizaje máquina.

- Multiple Star Classifier (MSC) determina los parámetros astrofísicos de sistemas binarios, compuestos por una estrella principal y una estrella secundaria. Además de los parámetros fundamentales de ambas estrellas, MSC determina parámetros del sistema, como el cociente entre las luminosidades. Los métodos utilizados por MSC son similares a los empleados en GSP-Phot.

- Quasar Classifier (QSOC) se encarga de estimar los principales parámetros de los quásares: el desplazamiento al rojo, el ancho equivalente de las líneas de emisión y la pendiente del continuo. También clasifica los quásares en 3 subtipos. El principal método de regresión en este caso son los árboles de decisión denominados *Extremely Randomized Forest*.

- Unresolved Galaxy Classifier (UGC) realiza una clasificación de las Galaxias en varios subtipos (elíptica, espiral, irregular, etc.) y determina parámetros astrofísicos de las galaxias como su desplazamiento al rojo, su tasa de formación estelar y la extinción interestelar total. UGC también se fundamenta en el uso de SVMs especializadas para la tarea.

- Object Clustering Algorithm (OCA) aplica técnicas de clasificación no supervisada para determinar los grupos (*clusters* por su nombre en inglés) naturales que forman todos los objetos observados por Gaia, a través de los datos obtenidos por el instrumento BP/RP, de los cuales se extraen parámetros estadísticos. De esta forma, OCA complementa a DSC, buscando grupos de objetos que no han sido bien modelados en los conjuntos de entrenamiento. OCA utiliza un algoritmo jerárquico de estimación de densidad no paramétrica denominado HMAC [4].

- Outlier Analysis (OA) tiene una misión similar a la de OCA, pero en este caso trabaja tan sólo con un subconjunto de objetos, formado por aquellos que no han podido ser clasificados con certeza por parte de DSC. Esto permite un análisis más detallado de dichos objetos. El capítulo 3 tratará en profundidad las herramientas y métodos que se han desarrollado para este paquete de trabajo.

- Final Luminosity, Age and Mass Estimator (FLAME) hace uso de las llamadas isocronas para determinar la luminosidad, edad y masa de las estrellas, a partir de los parámetros estimados por GSP.

- Total Galactic Extinction (TGE) calcula un mapa de extinción interestelar bidimensional de la Vía Láctea, lo cual puede usarse para el cálculo de la extinción interestelar total de la misma y también como asistencia para otros módulos en CU8 que necesitan corregir la extinción en los datos antes de aplicar sus métodos.

TABLA 1.2: Clases de objetos predefinidas en la misión Gaia.

NOMBRE	**DESCRIPCIÓN**
STAR	Objetos estelares simples en general, incluyendo todos los estados de evolución (menos enanas blancas).
QSO	Objetos extragalácticos cuasi-estelares, también denominados quásares.
GALAXY	Galaxias de todos los tipos y edades.
PHYSBINARY	Sistema binario de estrellas ligado gravitacionalmente.
NONPHYSBINARY	Sistema binario de objetos, por superposición en el cielo. Puede tratarse de un par estrella-estrella, estrella-galaxia, estrella-quásar, etc.
WD	Enanas blancas. Estrellas de baja masa, en su última etapa de evolución.

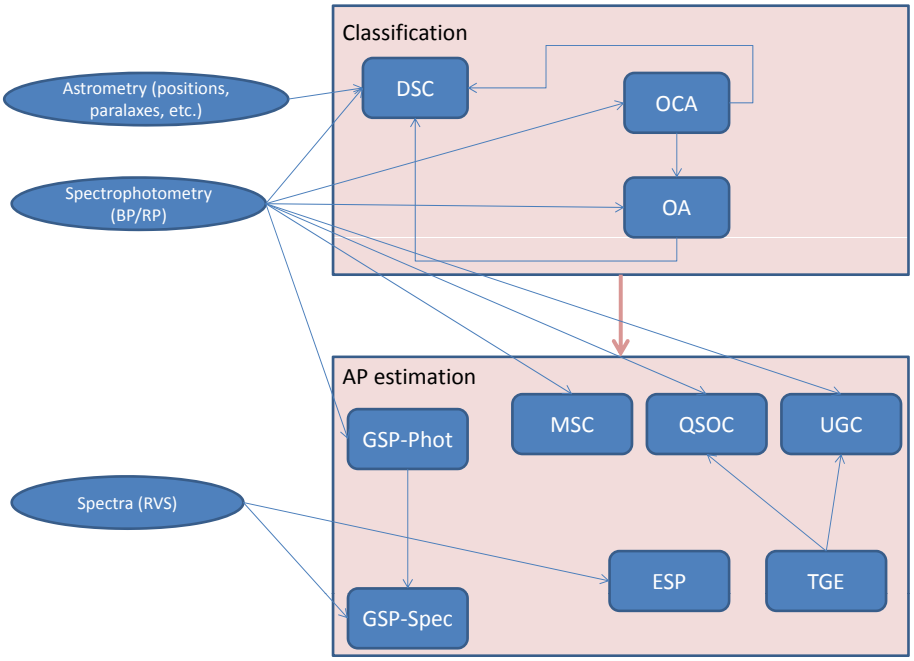


FIGURA 1.12: Flujo de datos entre paquetes de trabajo en la CU8.

El procesado en cadena de todos los DUs en CU8 se integra en un sistema denominado APSIS. Las dependencias entre los WPs de APSIS, así como el flujo de datos se muestran en el esquema de la figura 1.12. Más información sobre APSIS y los algoritmos que lo componen puede encontrarse en [5] y las referencias que allí se encuentran.

En esta tesis se describe en detalle el desarrollo de los sistemas de dos DUs en CU8: GSP-Spec y OA. De esta forma se abarcarán tanto la estimación de parámetros como la

clasificación no supervisada de objetos astrofísicos, desde la perspectiva de la aplicación de técnicas de AI. El objetivo será el de encontrar los mejores algoritmos para resolver cada tarea, así como las configuraciones óptimas de los mismos. Se tendrán en cuenta diversos factores a la hora de evaluar los algoritmos implementados, tanto su precisión al realizar la tarea como su coste computacional, su facilidad de adaptación a entornos cambiantes y la complejidad de análisis de los modelos resultantes. Se mostrará el diseño completo de los sistemas, de forma que sean aplicables al contexto del procesado de los datos del catálogo Gaia.

Capítulo 2

Estimación de parámetros astrofísicos mediante ANNs

Este capítulo se centra en el diseño de un algoritmo destinado a la estimación de parámetros astrofísicos (en adelante APs), basado en la aplicación de ANNs. Dicho algoritmo formará parte del DU llamado GSP-Spec dentro de la CU8 que procesará los datos de Gaia. Este es un problema bien conocido, tanto Oen aprendizaje máquina como en estadística inferencial, denominado regresión. En este caso el objetivo es la estimación de los principales APs de una estrella a partir de su espectro RVS. Los APs a estimar son T_{eff}, $log\, g$, $[Fe/H]$ y $[\alpha/Fe]$. En primer lugar, se utilizarán redes neuronales para aprender la función que relaciona los valores del flujo del espectro con los parámetros a estimar, a través de un conjunto predefinido de entrenamiento, para luego aplicar la función aprendida para la estimación de los APs de nuevas estrellas. Se compararán las distintas técnicas buscando una estimación óptima de los distintos APs, realizando un análisis detallado de los resultados obtenidos para cada tipo de población estelar que será observado por los instrumentos del satélite Gaia.

2.1 Estado del arte

El análisis de los espectros estelares permite la extracción de los principales parámetros astrofísicos de las estrellas, como su edad, masa, luminosidad y composición química. Dichos parámetros se obtienen al comparar los espectros observados con espectros sintéticos, obtenidos en el laboratorio a través del conocimiento físico de las transiciones electrónicas que se producen en los átomos (o en las moléculas), cuando éstos son atravesadas por un haz de fotones, los cuales son absorbidos o emitidos en ciertas longitudes de onda, en función de la composición química de las moléculas atravesadas.

El proceso de comparación entre el espectro observado de una estrella (mediante un espectrógrafo situado en un telescopio) y uno sintético (obtenido con modelos físicos o con espectros plantilla) se ha realizado tradicionalmente de forma manual por expertos humanos. Sin embargo, a partir de los años 80 se comenzaron a aplicar técnicas automáticas para obtener clasificación espectral en el sistema MK. Un sistema automático de clasificación permite analizar estadísticamente grandes poblaciones de estrellas. Además, los sistemas automáticos garantizan la objetividad, homogeneidad y reproducibilidad de las clasificaciones.

En su inicio, los procesos automáticos de clasificación espectral se desarrollaron en base a dos paradigmas: evaluación de criterios sobre índices espectrales y técnicas de minimización de distancias, véase la revisión en [6]. Los sistemas evaluación de criterios estaban basados en la definición, por parte de un experto, de un conjunto de reglas y modelos que asocian índices espectrales (ancho equivalente de líneas de absorción, proporciones entre lineas, índices de color, etc.) con clases del sistema MK, véase [7]. Por otro lado, las técnicas de minimización de distancias se basan en representar los espectros como vectores y minimizar la distancia (correlación cruzada, distancia euclídea, χ^2) entre el espectro observado y los espectros de referencia, véase [8]. La ventaja de los sistemas de evaluación de criterios es su relativa rapidez de cómputo y la fácil interpretación de sus resultados. Sin embargo, requieren del conocimiento de un experto para definir los índices espectrales y las reglas de clasificación, lo cual puede resultar en sistemas de clasificación subjetivos. Por otro lado, las técnicas de minimización de distancias son capaces de solventar estos inconvenientes, ya que comparan directamente los espectros a clasificar con las plantillas mediante la métrica definida, utilizando el concepto intuitivo de similitud. Sin embargo, requieren de un largo tiempo de cómputo y dependen significativamente de la completitud del conjunto de espectros de referencia utilizados en el análisis. Además, los sistemas basados en métricas son más sensibles al ruido en los espectros, así como a problemas en la calibración de los mismos.

A partir de los años 90, se comenzaron a emplear métodos de aprendizaje máquina, principalmente ANNs, para la clasificación espectral en el sistema MK, véase el trabajo de Von Hippel [9] y el de Weaver [10]. Las ANNs ofrecen varias ventajas con respecto a los sistemas de clasificación anteriores. Son capaces de ponderar las características relevantes de entre todas las que se le presenten, por lo que no es necesario definir índices espectrales. Además, son capaces de resolver problemas no lineales y de generalizar sus soluciones, de forma que no dependen tanto de la completitud del conjunto de plantillas. Por último, su ejecución es muy rápida una vez han sido entrenadas. Desde entonces, las ANNs han sido el modelo de AI más utilizado hasta la fecha en astrofísica. Además de para parametrizar o clasificar espectros, se han utilizado para diversas aplicaciones como la clasificación morfológica de galaxias [11] o la predicción de manchas solares [12].

Al final de la década de los 90, se avanzó desde el modelo de clasificación MK hacia un modelo de regresión para la estimación de los principales APs estelares. El trabajo de Bailer-Jones et al. [13] fue pionero en este sentido al aplicar espectros sintéticos como conjunto de entrenamiento. Más tarde, el trabajo de Snider et al. [14] demostraría que al entrenar las redes con conjuntos de espectros ruidosos, con un SNR parecido al del espectro de entrada, se obtienen mejores parametrizaciones. Estos trabajos demostraron la gran capacidad que tienen las ANNs para aprender funciones altamente no lineales, que incorporan múltiples variables y donde el ruido en la señal de entrada es significativo, superando ampliamente a expertos humanos. Posteriormente, ya en el siglo XXI, Manteiga et al. [15] fueron un paso más allá al utilizar wavelets como método de extracción de características. En dicho trabajo se demostró que, en función del SNR, estos métodos pueden ofrecerle a la ANN una representación mejorada del espectro, de forma que la red es capaz de obtener parametrizaciones más precisas. A pesar de sus ventajas, las ANNs han recibido críticas debido la complejidad de los modelos que generan, de forma que resulta muy difícil comprender sus soluciones, creando un efecto de caja negra. Además, las soluciones obtenidas no aportan una medida de incertidumbre sobre las mismas, por lo que el investigador no puede evaluar de forma simple hasta qué punto los APs calculados por la ANN son aceptables o válidos.

En los últimos años, ha surgido una nueva metodología basada en modelos generativos o "hacia adelante". Estos modelos realizan la operación inversa a los modelos típicos de regresión. En lugar de estimar los APs a partir de los espectros, estiman los espectros a partir de los APs. Así, enfocándose en la parametrización de espectros, estos modelos estiman el espectro a partir de sus parámetros atmosféricos. Este espectro estimado puede entonces compararse con el observado, mediante una función de verosimilitud dada, por ejemplo el test de χ^2. Este tipo de modelado puede introducirse en un esquema bayesiano donde se haga uso del mismo para calcular la distribución de probabilidad a posteriori de cada uno de los APs, la cual puede tomarse como medida de incertidumbre. Esta metodología está siendo aplicada para la estimación de APs en dos algoritmos que forman parte de GSP-Phot: ILIUM y Aeneas, descritos en [16] y [17]. Varios algoritmos en GSP-Spec usan también esta perspectiva, como veremos en las próximas secciones.

2.2 DU-823: GSP-Spec

GSP-Spec es el paquete de trabajo en CU8 encargado de la obtención de APs a partir de espectros RVS. El instrumento RVS fue originalmente diseñado para medir velocidades radiales de las estrellas con $G < 17$. Su rango espectral coincide con el pico de emisión de las estrellas de tipo F, G y K, las más abundantes en nuestra galaxia. Estas estrellas

muestran, en dicha región del espectro electromagnético, el triplete de calcio, formado por tres bandas de absorción del calcio ionizado, situadas en 8498, 8542 y 8552 Å. Por otro lado, en el caso de estrellas más tempranas, de tipo espectral A y B, los espectros presentan líneas espectrales de la serie de Paschen del hidrógeno, así como líneas de helio. Como veremos en este capítulo, dichas líneas también permiten la estimación de APs de un gran número de estrellas.

GSP-Spec depende de la calibración realizada en CU6, unidad de la que recibe el espectro calibrado y normalizado. Además, se ha definido una retroalimentación entre GSP-Spec y CU6, de forma que los APs estimados por GSP-Spec servirán de ayuda para mejorar la normalización del espectro realizada por CU6, lo cual a su vez ayuda a mejorar la calidad en la estimación de nuevos APs. GSP-Spec incorpora varios algoritmos que se encargan de estimar sus APs de forma independiente. Así, el módulo aportará varias estimaciones para cada estrella, de forma que los análisis posteriores pueden escoger de qué estimación partir, normalmente en función de los resultados obtenidos por cada algoritmo para cada combinación de SNR, parámetro a estimar y tipo de estrella. En la actualidad han sido implementados hasta 5 algoritmos, desarrollados por diferentes grupos de investigación, dedicados a la estimación de APs en GSP-Spec:

- *DEGAS* es un algoritmo basado en un árbol de decisión oblicuo. El árbol es binario, siendo la mediana multidimensional de los valores de flujo la norma de decisión entre las dos ramas. Una vez el espectro alcanza una hoja del árbol, los APs estelares se estiman mediante una media ponderada de los mismos en los espectros de referencia que componen dicha hoja, teniendo en cuenta la distancia entre el espectro de entrada y los de referencia. DEGAS se describe en [18].

- *MATISSE* es un algoritmo basado en álgebra lineal. En la fase de aprendizaje, el algoritmo estima una serie de funciones $B(\lambda)$, las cuales definen un producto escalar que obtiene el valor de un AP dados los valores de flujo del espectro. Las funciones se calculan de forma local, cubriendo una pequeña región del espacio de parámetros. Así, se define una aproximación lineal local a un modelo de regresión no lineal. Como estimación inicial, el algoritmo se inicializa con APs obtenidos por otros algoritmos como DEGAS. Más información puede ser obtenida de [19].

- *ANN* se basa en la aplicación de ANNs para la estimación no lineal de los parámetros. La transformada wavelet se utiliza para obtener representaciones del espectro más adecuadas para la estimación de APs, en función de la SNR y del parámetro a estimar. El funcionamiento de *ANN* ha sido descrito en [15].

- *GAUGUIN* realiza un minimización de distancia entre el espectro a parametrizar y la grilla de referencia. A partir de una estimación inicial del conjunto de

APs, normalmente proporcionada por otros algoritmos como DEGAS, GAUGUIN realiza una interpolación lineal de la grilla a través de las derivadas del flujo con respecto a los parámetros. El algoritmo converge cuando se alcanza la mínima distancia entre el espectro de referencia interpolado y el espectro a parametrizar, el cual recibe como estimación los APs correspondientes al espectro interpolado. Más información se puede encontrar en [18].

- *FERRE* tiene un funcionamiento similar a GAUGUIN. El conjunto completo de espectros teóricos se guarda en memoria. Entonces, a la hora de parametrizar un espectro concreto, los espectros teóricos son interpolados, hasta que se alcanza la mejor concordancia entre dicha interpolación y el espectro observado. La interpolación se realiza por defecto mediante el algoritmo Nelder-Mead [20]. Para ahorrar espacio en memoria y acelerar el proceso de interpolación se comprimen los espectros teóricos mediante un PCA.

A continuación, se profundiza en el diseño del módulo ANN, detallando nuevos experimentos y diversas mejoras que se han realizado posteriormente a [15].

2.3 Simulación de espectros RVS

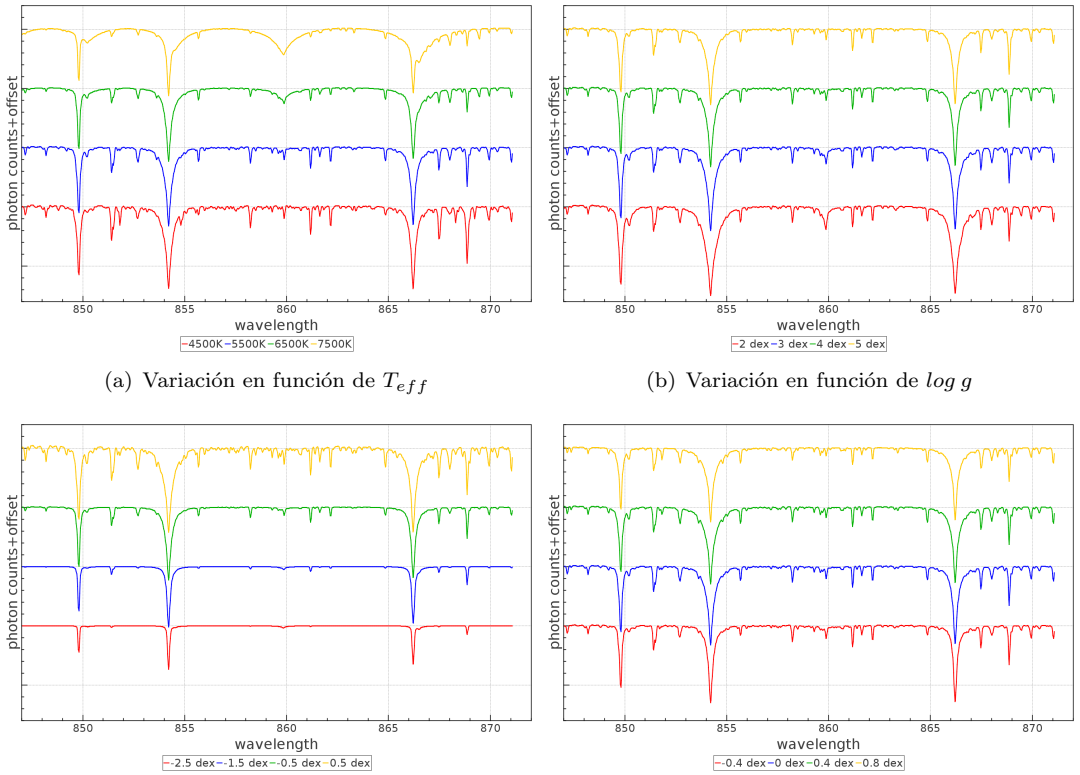
Con el objetivo de preparar los algoritmos de GSP-Spec, se ha generado una simulación que abarca todo el espacio de APs a estimar. Se han generado simulaciones para estrellas A, F, G y K, utilizando el modelo de atmósferas estelares de Kurucz [21]. En el caso de estrellas A, no se estimará el parámetro $[\alpha/Fe]$, ya que dichas estrellas apenas muestran líneas de absorción relacionadas con este tipo de elementos metálicos. Por lo tanto, se han generado dos grillas diferentes, una para estrellas F, G y K y otra para estrellas A, con el objetivo de *entrenar* los algoritmos de estimación. La cobertura y resolución de ambas grillas se especifica en la tabla 2.1. En el caso de las estrellas F, G y K, dicha resolución en 4 APs conlleva un total de 5831 espectros en la grilla. Además, el instrumento RVS tiene dos modos de observación: en alta resolución (en adelante HR) para estrellas con $G < 10$ y en baja resolución (en adelante LR) para estrellas con $10 < G < 16$. Así, los espectros RVS tienen 1039 píxeles en HR y 346 en LR. Todo esto provoca que los tiempos de entrenamiento sean un factor a tener en cuenta a la hora de diseñar algoritmos de regresión. Este problema se denomina comúnmente en el campo del aprendizaje máquina como *la maldición de la dimensionalidad*.

Las figuras 2.1 y 2.2 muestran la variación en el flujo espectral cuando se modifican cada uno de los APs mientras los demás se mantienen fijos. Se puede observar que los espectros de estrellas FGK son especialmente sensibles a la temperatura y la metalicidad,

TABLA 2.1: Grillas simuladas para el entrenamiento de algoritmos en GSP-Spec.

Clase espectral	**Parámetro**	**Rango**	**Resolución**
A	T_{eff}	[7500,11500]	500K
A	$log\ g$	[2,5]	0.5dex
A	$[Fe/H]$	[-2.5,0.5]	0.5dex
FGK	T_{eff}	[4000,8000]	250K
FGK	$log\ g$	[2,5]	0.5dex
FGK	$[Fe/H]$	[-2.5,0.5]	0.5dex
FGK	$[\alpha/Fe]$	[-0.4,0.8]	0.2dex

mientras que los espectros de estrellas A están dominados por la gravedad superficial. Por lo tanto, es esperable que las estimaciones de los APs que ejercen más impacto sobre el espectro sean más robustas al ruido que las de aquellos cuya influencia sea menor.



(a) Variación en función de T_{eff}

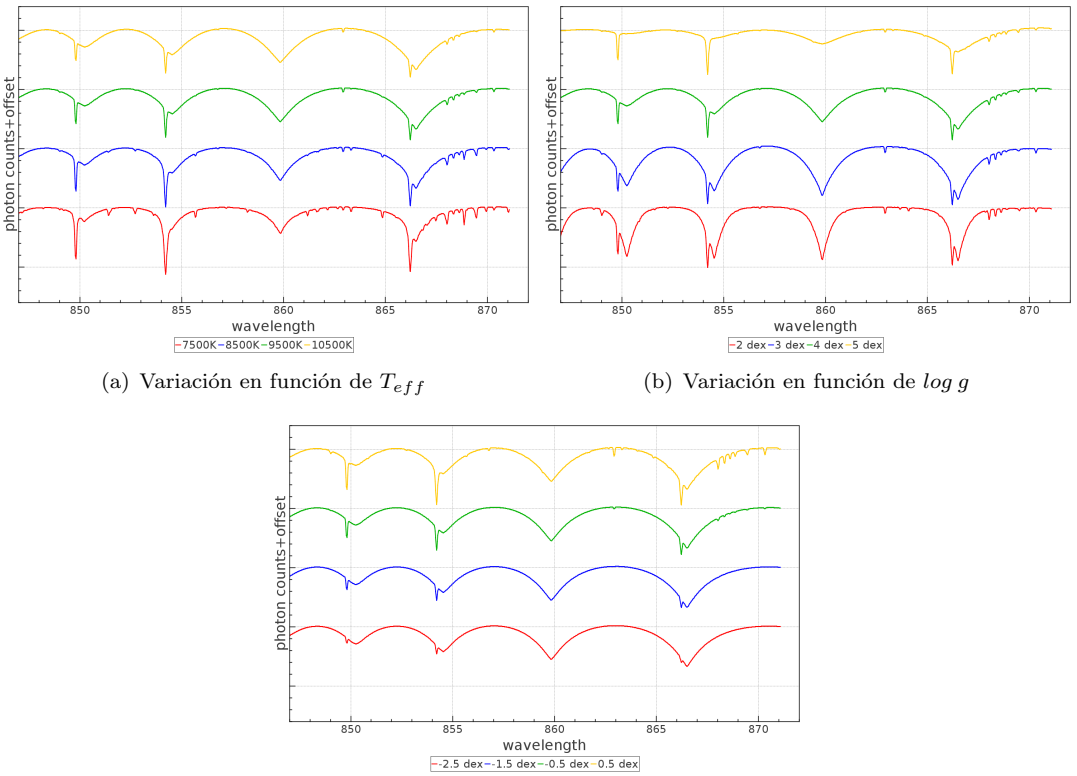
(b) Variación en función de $log\ g$

(c) Variación en función de $[Fe/H]$

(d) Variación en función de $[\alpha/Fe]$

FIGURA 2.1: Variación del espectro RVS (sin ruido) en función de sus APs para estrellas FGK. La variación de un parámetro se muestra fijando los valores del resto de APs. Los valores fijados son $T_{eff} = 5500K$, $log\ g = 4$ dex, $[Fe/H] = 0$ dex y $[\alpha/Fe] = 0$ dex.

Los algoritmos de GSP-Spec tienen como objetivo la predicción de los parámetros de nuevas estrellas observadas. Con el objetivo de evaluar la capacidad de los algoritmos para generalizar sus soluciones a nuevas estrellas, y de comparar su efectividad, se han generado un conjunto de simulaciones de test. Este conjunto, llamado conjunto *Random*,



(a) Variación en función de T_{eff}

(b) Variación en función de $log\ g$

(c) Variación en función de $[Fe/H]$

FIGURA 2.2: Variación del espectro RVS en función de sus APs para estrellas A. La variación de un parámetro se muestra fijando los valores del resto de APs. Los valores fijados son $T_{eff} = 9500K$, $log\ g = 4$ dex y $[Fe/H] = 0$ dex

.

se basa en la generación aleatoria de un conjunto de APs, a partir de los cuales se obtiene un espectro mediante la interpolación de los espectros con APs similares en la correspondiente grilla. Entonces, al espectro interpolado, se le añaden los efectos de ruido esperados en operación real, en concreto el ruido de Poisson inherente a la llegada de fotones y el ruido de Gauss debido al error introducido por el CCD (ruido termal, ruido de lectura, conversión a digital, etc.). Descartando los errores de calibración, el ruido en los espectros RVS depende principalmente de la magnitud del objeto en la banda observada, G_{rvs} y del número de tránsitos, $ntran$. En las simulaciones, se asume $ntran = 72$ en todos los casos. La figura 2.3 muestra el SNR del espectro RVS en función de la magnitud de la estrella, mientras que la figura 2.4 muestra la variación del espectro RVS en función de la magnitud. Otra fuente de errores podría ser el error en la determinación de la velocidad radial en las estrellas, ya que esto conllevaría un fallo en la corrección del desplazamiento al rojo (o al azul) del espectro. Sin embargo, los errores estimados en la corrección de velocidades radiales en operación real del satélite son extremadamente pequeños, de modo que su efecto en comparación con el ruido en la lectura de fotones es ínfimo y por lo tanto se ha descartado su consideración en las simulaciones. Por último, se esperan errores de normalización y calibración en

los espectros, así como otro tipo de desviaciones entre los espectros simulados y los observados por Gaia. Aquí se muestra el rendimiento de los algoritmos sin tener en cuenta dichos efectos, los cuales son impredecibles. Su impacto será evaluado en el futuro mediante observaciones de Gaia sobre un conjunto de estrellas de referencia.

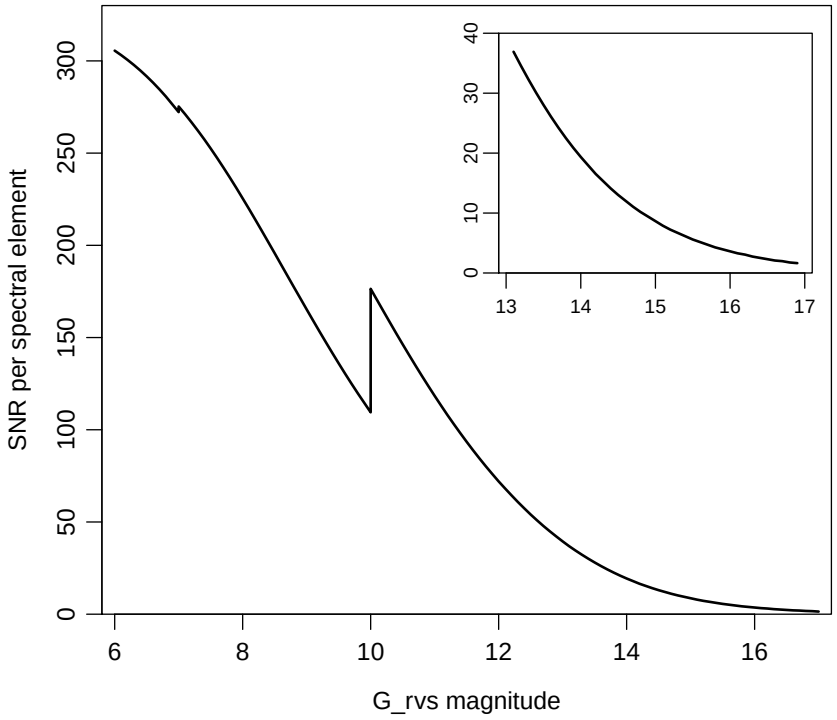
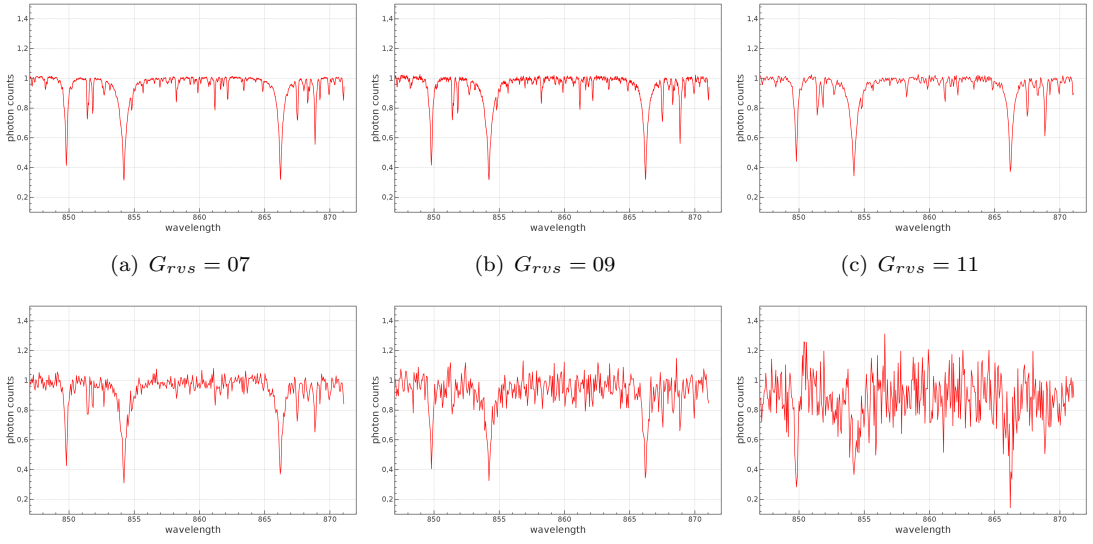


FIGURA 2.3: Relación de señal a ruido del espectro RVS en función de la magnitud de la estrella. El salto en $G_{rvs} = 10$ se debe al cambio de resolución del instrumento.



(a) $G_{rvs} = 07$ (b) $G_{rvs} = 09$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

FIGURA 2.4: Variación de un espectro RVS en función de la magnitud G_{rvs}.

Se han generado dos conjuntos independientes de espectros *Random*, uno de validación con 100 espectros y otro de test con 10,000 espectros. El de validación se utiliza de modo interno para el entrenamiento de *ANN*, mientras que el de test será utilizado como base para la comparación de los distintos algoritmos.

2.4 El módulo ANN en GSP-Spec

2.4.1 Objetivos de ANN

El objetivo del módulo *ANN* es el de proporcionar estimaciones precisas para los APs del mayor número de estrellas posible. En este sentido, complementará a otros algoritmos ya integrados en GSP-Spec, ejerciendo por un lado como un método de validación de los mismos y, por otro lado, como alternativa cuando ofrezca mejores resultados, especialmente en casos de extrema dificultad. Los resultados se valorarán en función de los residuos entre los valores esperados y los obtenidos por el algoritmo cuando se le presentan los espectros del conjunto de test. Se buscará obtener unos residuos de la menor magnitud posible, que además no presenten un sesgo sistemático debido al mal funcionamiento de los algoritmos.

En el campo de las ANNs se han definido multitud de arquitecturas de organización de las neuronas, así como algoritmos para entrenarlas. Nos centraremos en la arquitectura *feed-forward* con tres capas, que es la más utilizada en problemas de regresión. Se estudiarán varios métodos de entrenamiento de las redes, valorando su velocidad de convergencia y la calidad del modelo obtenido. Además, se estudiará el efecto de cambiar la representación del espectro de entrada en los resultados obtenidos. Todo ello en función del SNR y el tipo de estrella a parametrizar. Por último, se desarrollará un método bayesiano para la estimación de la probabilidad de distribución a posteriori de los APs, utilizando para ello un modelo generativo basado en ANNs.

2.4.2 Configuración de la ANN

Establecer la configuración de una ANN de cara a resolver un problema concreto no es trivial y requiere de cierta experimentación. Para ello, deben definirse la arquitectura de la red y la metodología del aprendizaje de la misma.

2.4.2.1 Arquitectura de la ANN

Existen multitud de posibles arquitecturas para una ANN, dependiendo de cómo se definan las conexiones entre las neuronas de la misma. Sin embargo, en regresión la arquitectura más utilizada es la llamada *feed-forward*, comúnmente con tres capas de neuronas: la capa de entrada, la capa oculta y la capa de salida, como se muestra en la figura 2.5. La capa de entrada tiene tantas neuronas como el número de variables predictoras X, mientras que la capa de salida tiene tantas neuronas como el número de variables a predecir Y. La capa oculta conecta las capas de entrada y salida mediante una función no lineal. De esta forma, dado que la red está completamente conectada, y siempre que el número de neuronas en la capa oculta sea suficiente, las redes *feed-forward* pueden aproximar cualquier función real. Esto, junto con su sencillez, es la razón por la cual nos hemos decantado por esta arquitectura en este trabajo.

Con el objetivo de adaptar la arquitectura *feed-forward* a este problema, se ha seleccionado como función de activación de la capa oculta una función logística, mientras que las neuronas de la capa de salida se activan mediante una función lineal. Además, por conveniencia, las entradas y salidas de la red se normalizan al intervalo [0,1] antes de ser presentadas a la ANN, mediante el cálculo de los máximos y mínimos para cada píxel de entrada y cada parámetro astrofísico de salida en las mallas de entrenamiento. Por último, se debe establecer el número de neuronas en la capa oculta. Sin embargo, dicho parámetro depende de la función a aproximar, por lo que se determinará junto al resto de parámetros de aprendizaje de la ANN en la siguiente sección.

2.4.2.2 Método de aprendizaje para la ANN

Una vez se ha definido la arquitectura de la ANN, el proceso de aprendizaje se basa en la optimización de una función de coste en función de los pesos de las conexiones de la red. Generalmente, la función de coste a minimizar es el error cuadrático medio (MSE, por sus siglas en inglés) entre las estimaciones proporcionadas por la red neuronal $f(x_{ij})$ y las salidas deseadas para los patrones en el conjunto de entrenamiento y_{ij}, siguiendo la ecuación 2.1. Minimizar el MSE es equivalente a maximizar la función de verosimilitud, cuando el error en las salidas es de tipo gaussiano e independientemente distribuido. En caso de que los errores no sigan dicho régimen, es más conveniente maximizar una función de verosimilitud adecuada para el tipo de ruido en cuestión. Aquí, asumiremos el ruido como gaussiano y minimizaremos el error cuadrático. Para ello, se evaluarán diversos métodos de entrenamiento de redes neuronales.

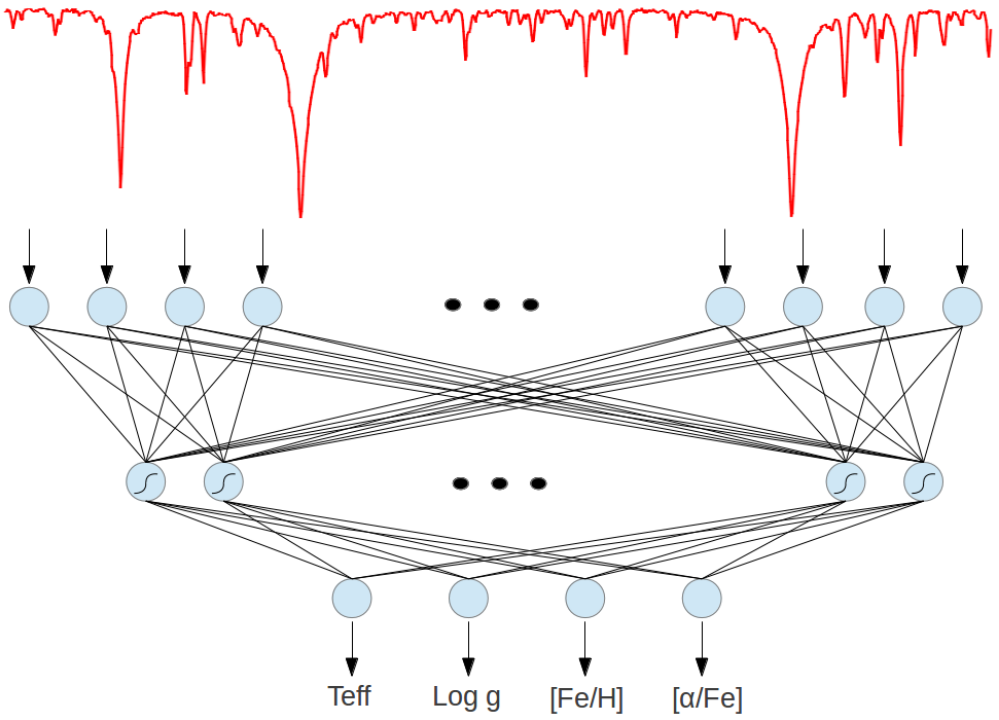


FIGURA 2.5: Arquitectura de una ANN feed-forward de tres capas para la estimación de APs

.

$$MSE = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{K} (y_{ij} - f(x_{ij}))^2 \tag{2.1}$$

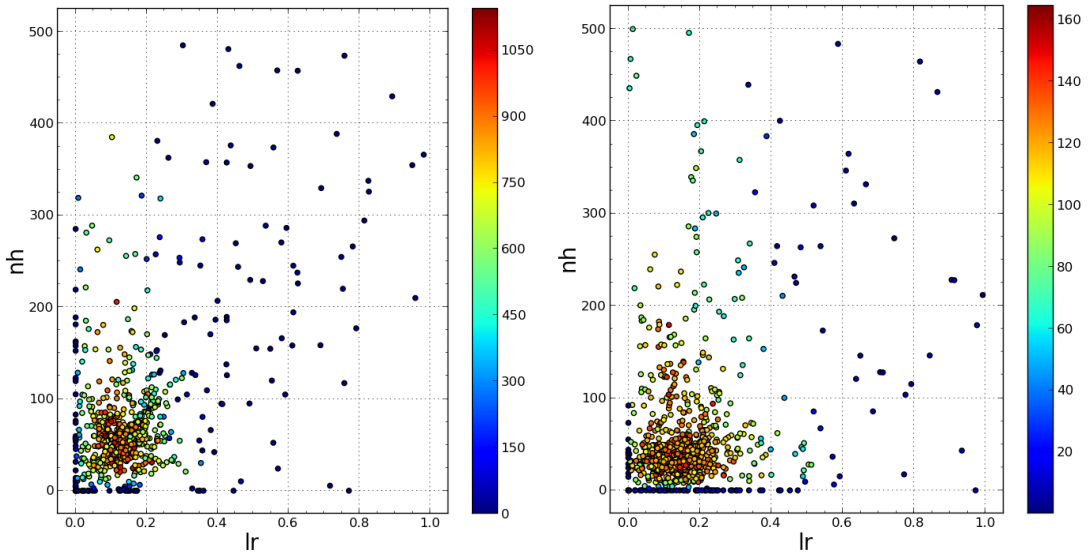
En los procesos de entrenamiento que se realizan en el módulo ANN se establece la misma magnitud G_{rvs}, tanto en el conjunto de entrenamiento como en el de test. La razón fundamental es la de ajustar el SNR de los espectros en ambos conjuntos. Esto proporciona mejores resultados de parametrización, como ha sido demostrado en otros estudios anteriores, véase [14]. La explicación a dicho comportamiento se basa en que el SNR en los espectros varía con la longitud de onda (heteroscedasticidad), debido a las diferencias en la sensibilidad del instrumento RVS. Al aplicar un ruido realista en el conjunto de entrenamiento, la red es capaz de tener dichas variaciones en cuenta, focalizándose en los valores menos ruidosos, siguiendo así un modelo de regresión ponderada por el ruido en las entradas. Además, se utiliza el método de *parada temprana* para buscar el mejor punto de generalización de la ANN en cada caso. La parada temprana es un método en el cual se pausa el entrenamiento de la ANN cada cierto número de iteraciones, con el objetivo de evaluar el rendimiento de la red, cuando ésta se aplica al conjunto de validación. De esta forma, es posible seleccionar la iteración en la cual la red neuronal proporciona mejores resultados, al aplicarse a conjuntos de

datos distintos al de entrenamiento, es decir, se guarda el estado de la red que mejor generaliza sus soluciones.

Adoptaremos el paradigma de entrenamiento de ANNs basado en la retropropagación del error. Nos centraremos en los métodos de primer orden, ya que los de segundo orden no serían viables con un número de pesos tan elevado, véase [22]. La inicialización de los pesos de la red es aleatoria, restringida en el intervalo [-0.2,0.2], para inicializar así la función logística cerca de su punto lineal, lo cual evita un sobreentrenamiento temprano, véase [23]. El algoritmo de retropropagación sufre de caída en mínimos locales, por lo que se proporcionarán hasta 10 inicializaciones independientes en cada caso. Entre ellas, se escoge la que ofrezca un mejor ajuste (menor MSE en el conjunto de validación) a lo largo de 1000 iteraciones. Asumiendo una ANN completamente conectada, faltarían tres parámetros de configuración por establecer: el factor de aprendizaje lr, el número de neuronas en la capa oculta nh y el modo de entrenamiento: *online* o *batch*.

Con el objetivo de optimizar los parámetros de entrenamiento de las ANNs se ha utilizado un algoritmo de optimización evolutivo. En este caso, se ha optado por el algoritmo llamado optimización por enjambre de partículas (PSO, por sus siglas en inglés), el cual se ha descrito como un método poderoso para la exploración de espacios de búsqueda continuos multivariantes, véase [24]. El algoritmo PSO se basa en la competición entre una serie de individuos o partículas, que exploran conjuntamente el espacio de soluciones, buscando maximizar una función dada de bondad, llamada función de $fitness$, cuyo valor es mayor cuanto mejor sea la solución aportada por la partícula. En el algoritmo PSO, cada partícula recuerda la posición donde encontró un mejor $fitness$. Además, el colectivo de partículas sabe cuál es el lugar del espacio donde se obtuvo el mejor valor de la función. Las partículas se mueven arbitrariamente entre estos dos puntos, recorriendo el espacio de soluciones. En este caso, la función de $fitness$ para las partículas se define como la inversa del error cuadrático medio producido por el conjunto de validación, i.e. $1/MSE$, cuando se aplica a una ANN entrenada con los parámetros de aprendizaje seleccionados por la partícula: lr y nh. La configuración del algoritmo PSO se ha establecido siguiendo las recomendaciones dadas en [25], dónde se describen un conjunto de valores para los parámetros de PSO (inercia, constantes de aceleración), los cuales permiten obtener buenas optimizaciones en un amplio rango de problemas no lineales. La figura 2.6 muestra el resultado del proceso de optimización de parámetros de aprendizaje de ANNs, dónde se ha escogido la versión *online* del algoritmo de entrenamiento. Como se puede observar, ciertas regiones en el espacio de parámetros ofrecen mejores parametrizaciones de forma significativa. El factor de aprendizaje lr óptimo se sitúa entre 0.1 y 0.2, en ningún caso por encima de 0.3. Por otro lado, se observa que el número de neuronas ocultas nh debe estar entre 20 y 80. Además, se nota una tendencia a favorecer ANNs con más neuronas ocultas en el caso

de espectros en HR con respecto a los mismos en LR. Dicho resultado es esperable, ya que los espectros en HR permiten definir funciones de regresión más complejas, sin perder generalidad en sus soluciones. Este proceso de optimización se ha repetido para comprobar su variabilidad cuando se aplica al entrenamiento bajo distintos valores de SNR y tipo de estrella. Como resultado, se ha determinado que los valores $lr = 0.12$ y $nh = 60$ ofrecen parametrizaciones cercanas al óptimo en todos los casos, y por lo tanto se han adoptado los mismos para el entrenamiento de ANNs.



(a) *Espectros HR* (b) *Espectros LR*

FIGURA 2.6: Evolución del *fitness* del algoritmo PSO en función del factor de aprendizaje (lr) y el número de neuronas en la capa oculta (nh), para ANNs entrenadas con espectros de alta y baja resolución. La barra de colores indica el fitness obtenido para cada posición en el espacio de parámetros.

Una vez establecidos los parámetros óptimos para el entrenamiento de ANNs en modo *online*, se han realizado experimentos con el método de entrenamiento por retropropagación en modo *batch*. Como resultado, se ha determinado que la velocidad de aprendizaje del modo *batch* es hasta 10 veces más lenta que la versión *online*. Además, con el objetivo de realizar un estudio más completo, se han implementado diversas variantes del método de retropropagación de errores, como el método llamado *steepest descent* y el método de retropropagación con momentos, véase [26] y [27]. Ninguno de ellos ha mejorado significativamente los resultados obtenidos con el método de entrenamiento de retropropagación *online*, por lo que éste será el utilizado para la parametrización de espectros RVS.

2.4.3 Selección y extracción de características

2.4.3.1 Extracción de características mediante procesado wavelet

Las técnicas de procesado digital permiten realizar estudios en diferentes espacios de representación de las señales, utilizando para ello las transformadas de dominio pertinentes. Un ejemplo es la representación de una señal en el dominio de la frecuencia, mediante la transformada de Fourier. Sin embargo, en los últimos años, la representación en el dominio del tiempo-frecuencia se ha popularizado ampliamente a través de la transformada *wavelet*, debido a sus ventajas con respecto a la representación única en frecuencia, ya que mediante las wavelet es posible analizar sistemas con frecuencias variables en el tiempo (representación en tiempo-frecuencia). Dicha técnica ha sido aplicada en campos como el reconocimiento de patrones y la minería de datos, véase [28], así como en astronomía, en la que se ha utilizado tanto para el procesamiento de imágenes como de espectros, véase [29] y [30]. En el caso del análisis espectral, se asume la longitud de onda del espectro como variable independiente y el espectro como variable dependiente de la longitud de onda. De esta forma, el procesado wavelet puede ser aplicado a un espectro dado, con el objetivo de ensalzar líneas espectrales de interés, para disminuir el ruido en los espectros o como método de compresión. En esta tesis, utilizaremos el procesado wavelet como método de preprocesado previo a la estimación de APs, con el objetivo de mejorar las estimaciones obtenidas por las ANNs.

El campo del procesamiento wavelet contiene una amplia variedad de técnicas, cada una de las cuales puede ser configurable para una aplicación en concreto. En primer lugar, es necesario escoger una función wavelet madre adecuada para la aplicación en cuestión. Basándonos en [15], aplicamos una wavelet madre de tipo Daubechies con 5 momentos de desvanecimiento (véase [31]). Esta elección se debe al gran número de problemas en los cuales las wavelets de tipo Daubechies han sido aplicadas con éxito. Por otro lado, se utilizará una transformada wavelet rápida (FWT por sus siglas en inglés), la cual fue propuesta por Mallat en 1988 [32]. La FWT es no redundante, lo cual evita problemas por alta dimensionalidad, y además es computacionalmente eficiente. Se basa en un esquema jerárquico de filtrado, donde en cada nivel se aplica un filtro de paso bajo $g[n]$ y un filtro de paso alto $h[n]$, acompañados por un proceso de submuestreo por un factor de dos. El filtro de paso bajo produce una nueva señal denominada "Aproximación", mientras que el filtro de paso alto genera una señal denominada "Detalle". Este proceso puede entonces repetirse filtrando la Aproximación, obteniendo así la Aproximación y el Detalle del siguiente nivel, siguiendo el esquema mostrado en la figura 2.7. De esta forma, es posible descomponer una señal en varias subseñales que representan distintos componentes de frecuencia de la misma. La figura 2.8 muestra un ejemplo de

transformación wavelet de un espectro RVS, en la cual se generan tres niveles de filtros. Como puede observarse, se generan 3 Aproximaciones (A1, A2 y A3) y 3 Detalles (D1, D2 y D3). De acuerdo con [15], utilizaremos las 3 Aproximaciones para la estimación de APs, ya que en dicho trabajo se han descrito como las representaciones más eficientes para la estimación con ANNs.

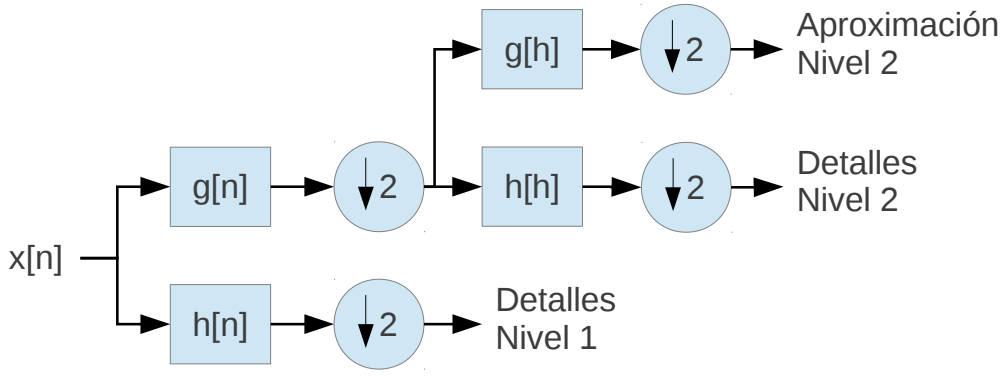
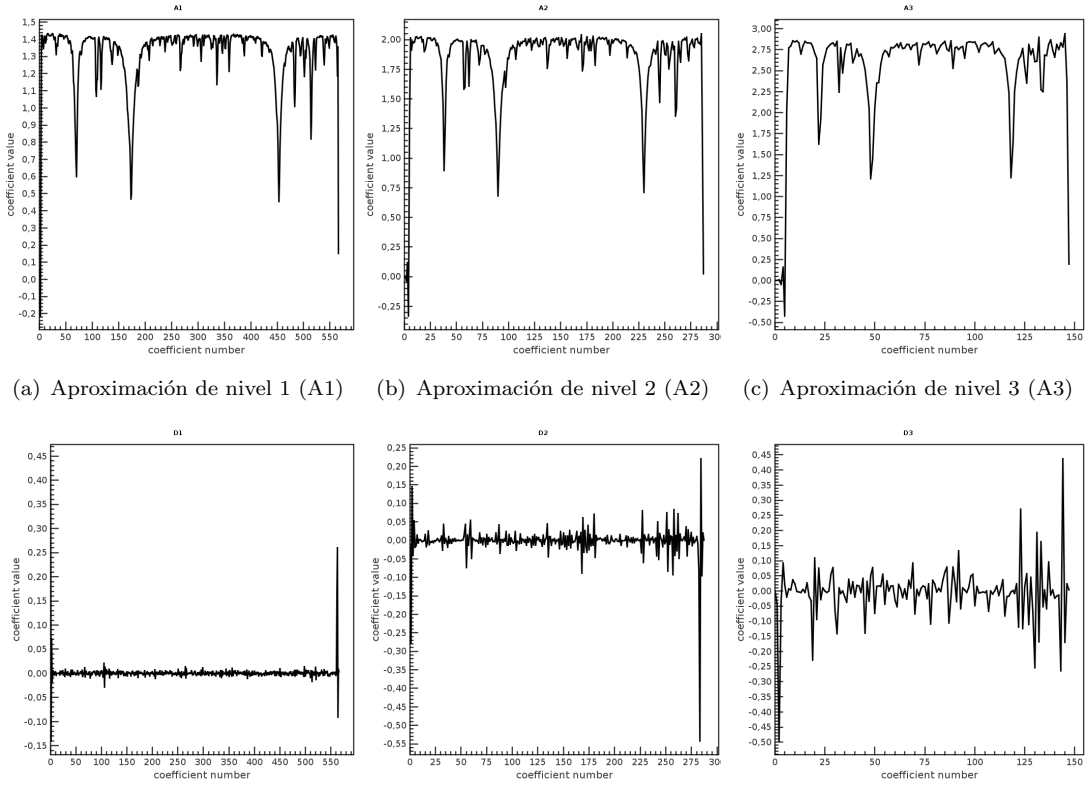


FIGURA 2.7: Esquema de filtrado para obtener una representación wavelet multiresolución de una señal dada.



(a) Aproximación de nivel 1 (A1) (b) Aproximación de nivel 2 (A2) (c) Aproximación de nivel 3 (A3)

(d) Detalles de nivel 1 (D1) (e) Detalles de nivel 2 (D2) (f) Detalles de nivel 3 (D3)

FIGURA 2.8: Descomposición wavelet piramidal en tres niveles de un espectro RVS.

Con los coeficientes de cada una de estas Aproximaciones, construimos una representación independiente de los conjuntos de entrenamiento y validación. Entonces, le presentamos a la ANN cada una de las representaciones, obteniendo los errores de validación correspondientes. Dichos errores se comparan con los obtenidos por medio de los espectros en formato original (de aquí en adelante formato lambda) en la figura 2.10, para estrellas de tipo A y en la figura 2.9, para estrellas tipo F, G y K. Los resultados dependen de varios factores, incluyendo tanto la resolución y SNR de los espectros como el AP a estimar. Se puede observar que, mediante las Aproximaciones wavelet, se obtienen estimaciones competitivas para estrellas de tipo F, G y K, siendo estas similares o mejores que las obtenidas con el espectro original. En concreto, la representación wavelet proporciona mejores resultados para estrellas muy débiles, con $Grvs >= 13$. Por otro lado, en el caso de estrellas de tipo A, las Aproximaciones obtienen residuos significativamente menores a los obtenidos con el espectro original. Esto se atribuye a la presencia de las líneas de Paschen, las cuales son relativamente anchas, de forma que las Aproximaciones, cuyas escalas se corresponden a bajas frecuencias, son capaces de obtener una mejor representación de las mismas, especialmente en presencia de ruido.

Cabe destacar que las ANNs, entrenadas con Aproximaciones wavelet, ofrecen mejores de resultados de validación que las entrenadas con el espectro original, pero no de entrenamiento. De hecho, el MSE de entrenamiento obtenido en el caso de espectros originales es considerablemente menor. Por lo tanto, se puede concluir que la representación wavelet mejora la capacidad de generalización de la ANN pero no su capacidad para ajustar el conjunto de entrenamiento. Es decir, mediante la representación wavelet, se obtiene un modelo más general, que obtiene mejores estimaciones cuando se aplica a espectros distintos a los que contiene el conjunto de entrenamiento, a pesar de ajustar peor el conjunto de entrenamiento en sí. Los resultados que nos importan son los de generalización, ya que éstos son los aplicables a la operación en entornos reales, por lo que nos decantaremos por la representación wavelet cuando ésta aporte mejores generalizaciones.

2.4.3.2 Selección de características mediante algoritmos genéticos

Debido a la alta resolución del espectro RVS, éste contiene un número de píxeles muy alto, sobre todo en el caso de espectros en HR. Para acelerar el proceso de estimación de parámetros en operación real, es posible realizar una selección de las características más relevantes del espectro, las cuales serán las presentadas a la ANN. Además, en ocasiones, los métodos de selección de características consiguen mejorar los resultados proporcionados por las ANNs. Un método típico para la selección de características es el análisis de componentes principales (PCA). Sin embargo, dicho método es no

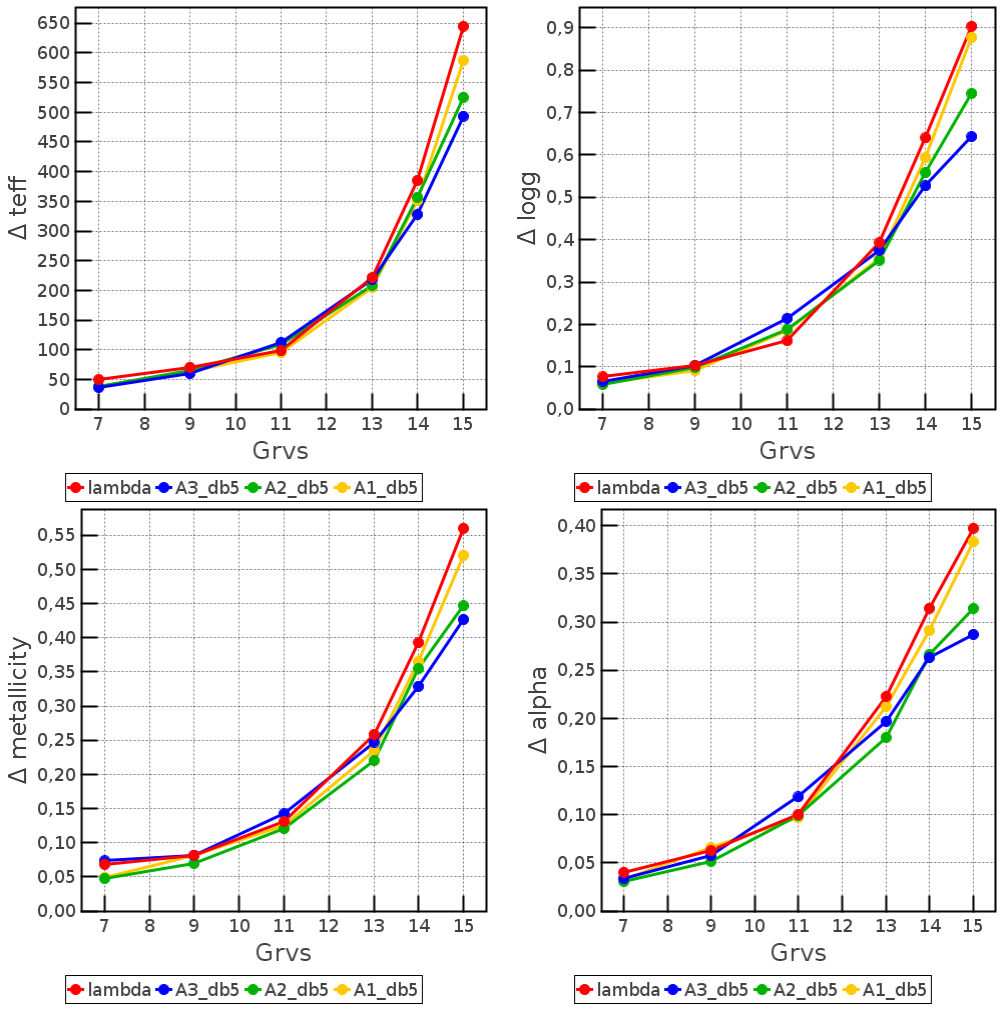


FIGURA 2.9: Percentil 70 de los residuos absolutos para estrellas de tipo FGK, para distintos dominios de entrada.

supervisado, en el sentido de que no tiene en cuenta el parámetro a estimar, si no que selecciona combinaciones lineales de entradas que acumulen la mayor cantidad de varianza posible. En [18] se demuestra que la PCA proporciona resultados similares a los obtenidos con el espectro completo, cuando éstas se aplican a la estimación de APs, y se retiene más del 99% de la varianza. Aquí, se utilizarán algoritmos genéticos clásicos para realizar la selección de características. La utilización de algoritmos genéticos permite realizar una selección de características no lineal y dirigida por los parámetros a estimar, en lugar de por la varianza en las entradas.

El algoritmo genético propuesto para la selección de características se basa en un alfabeto binario, de forma que cada uno de los píxeles del espectro puede ser escogido (valor 1) o no (valor 0). De esta forma, cada individuo en la población contendrá una selección de características, codificada en su cromosoma con el alfabeto binario, la cual competirá con la selección de los demás individuos. La competición entre individuos para optimizar la selección de características se divide en diversas fases, mostradas en la figura 2.11.

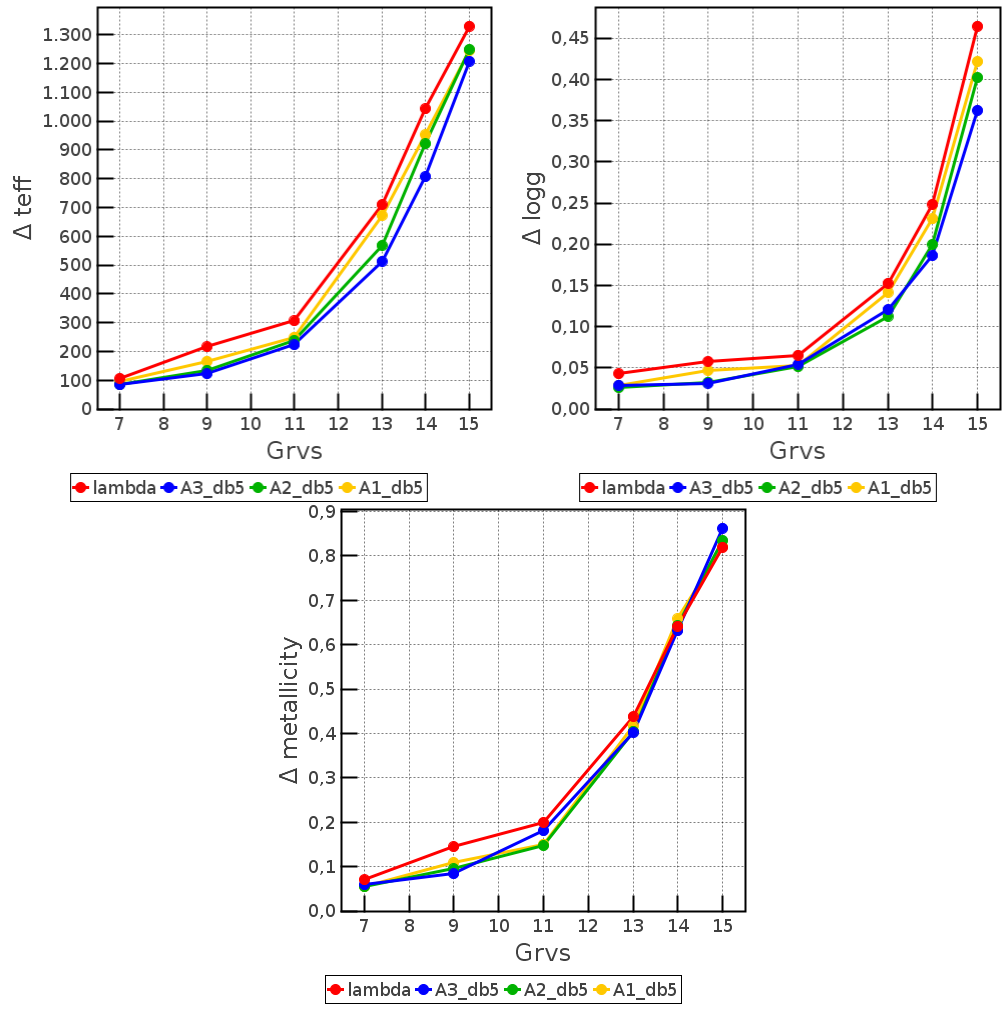


FIGURA 2.10: Percentil 70 de los residuos absolutos para estrellas de tipo A, para distintos dominios de entrada.

En primer lugar, se inicializa la población con selecciones de características aleatorias, las cuales contienen una selección de aproximadamente el 30% del número total de características. Entonces, por cada iteración:

1. Se evalúa la idoneidad (*fitness*) de cada individuo. En este caso, el *fitness* se calcula como la inversa del error de test $1/MSE$ producido por una ANN, entrenada y validada con los espectros RVS, reducidos con la selección de características codificada por el cromosoma del individuo que está siendo evaluado.

2. Se seleccionan la mitad de los individuos en la población para ser reproducidos. Para ello, se utiliza un algoritmo de ruleta, en el cual cada individuo tiene una probabilidad de ser seleccionado, la cual es proporcional al *fitness* del mismo.

3. Los individuos seleccionados se cruzan en parejas, de forma que cada pareja genera dos nuevos individuos. El cruce se realiza estableciendo dos puntos de

corte equidistantes en el cromosoma, de forma que las tres tiras resultantes son intercambiadas entre los individuos progenitores para generar dos descendientes.

4. Cada nuevo individuo, resultante del cruce de dos progenitores, es susceptible de sufrir ciertas mutaciones en su cromosoma, de acuerdo a una probabilidad prefijada.

5. Si se ha alcanzado el número máximo de iteraciones, se finaliza el proceso. En otro caso, se vuelve al paso 1 para evaluar el $fitness$ de los nuevos individuos.

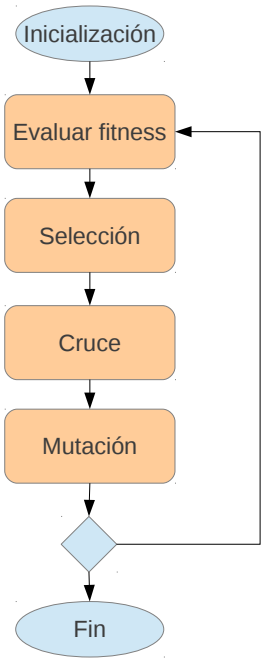


FIGURA 2.11: Esquema del algoritmo genético de selección de características en el espectro RVS para la estimación de APs.

La selección de características en un problema complejo, ya que en casos de alta dimensionalidad existe una gran cantidad de posibles combinaciones. De hecho, en el caso de espectros RVS en HR, existen 2^{1039} posibles selecciones. Los algoritmos genéticos resultan útiles para explorar de forma eficiente dicho espacio de soluciones. Seguidamente, se muestran los resultados obtenidos a través de la selección de características mediante algoritmos genéticos. Además de la selección de características del espectro en formato lambda, se realiza una selección de coeficientes wavelet de entre todos los provenientes de la descomposición piramidal de nivel 3 (A3, D3, D2 y D1). La tabla 2.2 compara los residuos obtenidos con ANNs al estimar APs de espectros completos, en formato lambda y wavelet, con espectros reducidos con la selección de características realizada mediante algoritmos genéticos. La comparación

TABLA 2.2: Comparación de resultados obtenidos en la estimación de APs mediante ANNs aplicadas a datos completos, con respecto a los resultados obtenidos mediante ANNs aplicadas a la selección de características obtenida mediante algoritmos genéticos.

Dominio	**SNR**	**MAR**(T_{eff})	**MAR**($log\ g$)	**MAR**($[Fe/H]$)	**MAR**($[\alpha/Fe]$)
lambda	200	80	0,16	0,12	0,07
lambda(genético)	200	102	0,18	0,11	0,05
lambda	10	297	0,59	0,37	0,17
lambda(genético)	10	450	1,09	0,38	0,12
wavelet	200	91	0,17	0,15	0,08
wavelet(genético)	200	125	0,19	0,13	0,06
wavelet	10	268	0,55	0,35	0,16
wavelet(genético)	10	368	0,62	0,37	0,09

se realiza mediante el residual absoluto medio (MAR). Se ha escogido dicho estimador para la comparación por ser el mismo robusto a la presencia de estimaciones atípicas, al contrario que el MSE. Como se puede observar, los errores medios obtenidos con la reducción son similares a los obtenidos con el conjunto de datos completo, ligeramente por debajo en el caso de T_{eff}, $log\ g$ y $[Fe/H]$ mientras que ligeramente por encima en el caso de $[\alpha/Fe]$. En todos los casos, el número de características (píxeles o coeficientes según el caso) ronda el 40% del número total de características.

Cabe mencionar que los resultados mostrados en esta sección han sido obtenidos para estrellas F, G y K de la simulación de espectros RVS anterior a la descrita y utilizada en este capítulo, por lo que no son directamente comparables con el resto de resultados presentados en el mismo. La principal limitación del método de selección explicado aquí es su gran complejidad computacional, ya que es necesaria una población de 50 individuos, compitiendo durante 25 iteraciones, para alcanzar un $fitness$ aceptable. Se han probado otras proporciones entre el número de iteraciones y el número de individuos sin que se observase ninguna mejoría. Los resultados mostrados aquí han sido obtenidos mediante la ejecución de una versión distribuida del algoritmo genético, implementada mediante el paradigma MPI, la cual conlleva varias semanas de cómputo. Por este motivo, no se ha continuado con esta línea de procesado en los ciclos posteriores. Más información sobre este experimento puede encontrarse en [33]. En el futuro, se planea una segunda ejecución del experimento completo, efectuada sobre espectros RVS reales, provenientes de las observaciones de Gaia.

2.4.4 Obtención de APs y medidas de incertidumbre mediante ANNs generativas (GANNs)

A la hora de estimar APs, es importante proporcionar no sólo un valor óptimo para los mismos, sino también un valor de incertidumbre sobre dicha estimación. De esta forma, los análisis posteriores pueden realizar inferencia a través de dicha estimación de

incertidumbre. Esta inferencia puede consistir en la detección de casos atípicos o mal modelados, en la calibración de los conjuntos de entrenamiento para ajustarse mejor a las observaciones (aprendizaje activo), o bien como herramienta de selección y análisis para el usuario explotador de los APs obtenidos.

La estimación de incertidumbre en las salidas es un campo de investigación activo en el ámbito de las ANNs, dada la complejidad de los cálculos implicados. El error de estimación de una ANN proviene de varias fuentes. En primer lugar, está el error en el modelo, provocado por la (falta de) densidad del conjunto de entrenamiento, por el error en las salidas deseadas $Y*$ y por el ajuste imperfecto de los pesos. Este error depende de la entrada X, de forma que puede ser mayor o menor en función de la misma. Los trabajos de MacKay [34] y Williams [35] constituyeron un primer paso hacia la obtención de estimaciones de incertidumbre que contemplasen las mencionadas fuentes de error. El método propuesto para ello se enmarca en un esquema bayesiano, en el cual primero se estima la incertidumbre de los pesos de la red dados los datos $P(W|X)$, para luego estimar la incertidumbre de las salidas de la red dada la entrada y los pesos óptimos $P(Y|X, W*)$. Ya que la ANN es una función no lineal, dicho cálculo de incertidumbre se aproxima mediante un desarrollo en serie de Taylor de segundo orden, lo que implica el cálculo de la matriz Hessiana. Por otro lado, varios investigadores han enfrentado este problema mediante la extensión de la arquitectura de las ANNs con neuronas adicionales en la capa oculta y en la de salida, véase [36].

En Astrofísica, y en Física en general, los datos provienen de sensores cuyas mediciones están sujetas errores. Este es el caso de los espectros observacionales, cuyos errores asociados vienen usualmente dados junto con las mediciones. Los trabajos previos no incorporaban dichos errores de entrada en su medición de incertidumbre en la salida de la ANN. Posteriormente, algunos autores incorporaron dichos errores al esquema bayesiano, véase [37]. Sin embargo, el cálculo de la matriz Hessiana resulta inviable en el caso que nos ocupa, ya que se trata de una ANN con unos 60.000 pesos, lo que conllevaría una matriz Hessiana de 60.000 filas por 60.000 columnas con datos en coma flotante de doble precisión, lo cual se traduce a 27GB. Tan sólo el almacenaje en memoria de dicha matriz resultaría impracticable, cuanto más el cálculo de su inversa. Otra opción, dada la estimación de incertidumbre en la entrada, sería muestrear la misma mediante una simulación de Montecarlo y evaluar la ANN para cada muestra. De nuevo, dicho método resultaría inviable debido a la alta dimensionalidad de los espectros.

En los últimos años, varios investigadores han propuesto la obtención de modelos generativos, que resuelven el problema directo, i.e. obtener la observación a través de los APs a estimar, en lugar de la aproximación tradicional, basada en resolver el problema inverso, i.e. estimar los parámetros a partir de la observación. Este tipo de

modelado se puede utilizar para estimar un conjunto de APs óptimo, o bien se puede combinar con el esquema de inferencia bayesiana, para estimar así la distribución de probabilidad a posteriori de los parámetros, dados los espectros observados $P(AP|S)$.

De cara a estimar los APs óptimos con un modelo generativo, es necesario definir una función de verosimilitud $P(S|AP)$. Asumiendo que el error estimado θ_i en cada valor del espectro RVS tiene una distribución normal, y que los valores del espectro son independientes, dicha función se define como:

$$P(S|AP) = e^{-d/2}, d = \sum_{i=1}^{N} \left(\frac{s_i - f_i(AP)}{\theta_i} \right)^2 \tag{2.2}$$

, donde s_i es un valor del espectro RVS y $f_i(AP)$ es la estimación proporcionada por el modelo generativo, a partir del conjunto de parámetros astrofísicos AP. El modelo generativo utilizado en este caso es una ANN, denominada GANN (Generative ANN), la cual ha sido entrenada con la misma metodología que las ANNs presentadas en la sección anterior, pero con las entradas y salidas en orden inverso, véase el esquema en la figura 2.12. Las GANNs aprenden la función que aplica el simulador a la hora de generar los espectros *Random*, aunque en este caso se trata tan sólo de una aproximación, cuya ejecución es notablemente más rápida. No obstante, la aplicación de GANNs conlleva varias evaluaciones de las mismas, a la hora de encontrar un conjunto de APs óptimo para un espectro dado, lo cual multiplica su complejidad en comparación con las ANNs comunes. Para acelerar dicho proceso de maximización, pueden aplicarse algoritmos como el anteriormente descrito PSO. Como punto de partida, se utiliza la estimación de parámetros obtenidos por la ANN convencional, lo cual reduce el número de iteraciones necesarias para que el algoritmo PSO converja. A continuación, se discuten los resultados obtenidos mediante GANNs . Las figuras 2.13 y 2.14, comparan el percentil 70 de los residuos obtenidos por GANNs con los residuos obtenidos por ANNs, para los conjuntos de estrellas FGK y A, respectivamente. Los resultados indican que las GANNs obtienen parametrizaciones menos precisas que las ANNs, ya que los residuos obtenidos son mayores en el caso de estrellas brillantes. Sin embargo, las GANNs parecen ser más robustas al ruido, ya que éstas obtienen mejores parametrizaciones para estrellas del conjunto FGK con $G_{rvs} >= 13$ y para el conjunto A con $G_{rvs} >= 9$. La razón por la cual las GANNs son robustas se debe a su capacidad de adaptarse iterativamente al espectro observado y al ruido que éste muestra en particular.

Como se ha mencionado anteriormente, el modelo generativo, definido aquí mediante GANNs, puede aplicarse también para calcular la probabilidad a posteriori $P(AP|S)$. Dicha probabilidad queda definida por la ecuación de Bayes:

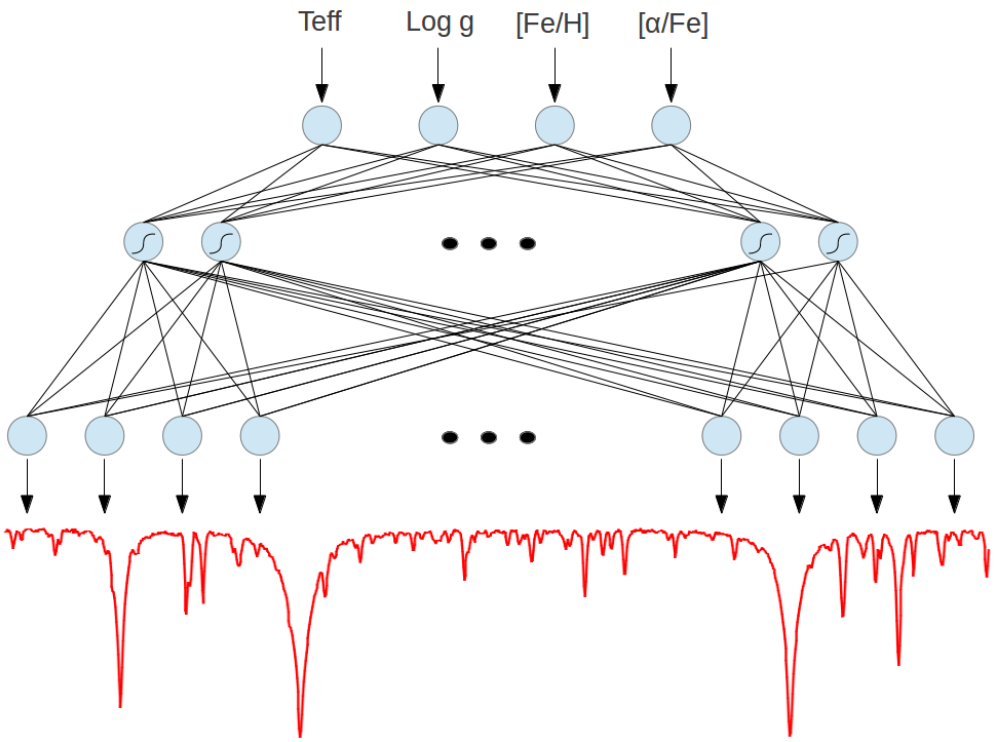


FIGURA 2.12: Arquitectura de una GANN de tres capas para la estimación del espectro RVS a partir de un conjunto de APs

.

$$P(AP|S) = \frac{P(S|AP)P(AP)}{P(S)} \tag{2.3}$$

Para aplicar la ecuación de bayes a la estimación de parámetros astrofísicos necesitamos definir de forma adecuada cada uno de sus términos. En primer lugar, definimos la probabilidad a priori $P(AP)$ como una distribución uniforme para cada parámetro a estimar, cuyos límites coinciden con los extremos del rango del parámetro correspondiente en la grilla de entrenamiento, mostrados en la tabla 2.1. Adicionalmente, es necesario definir una función de verosimilitud, $P(S|AP)$, la cual ha sido descrita anteriormente en la ecuación 2.2. Por último, $P(S)$ se trata como un factor de normalización que asegura que $\sum P(AP|S) = 1$. Este cálculo, requiere de un número de evaluaciones aún mayor de la GANN, con respecto al esquema de máxima verosimilitud. Por lo tanto, es importante la utilización de una técnica de muestreo efectiva para aproximar la distribución a posteriori. Aquí, aplicaremos el algoritmo de Metrópolis-Hastings para muestrear $P(AP|S)$, utilizando la estimación de parámetros obtenidos por la ANN como punto de partida de la cadena de Markov. A partir de la distribución a posteriori, es posible extraer intervalos de confianza a un determinado nivel para cada AP, calculando la distribución marginal correspondiente. De esta forma, puede determinarse la desviación esperada en cada uno de los APs, estimado para una

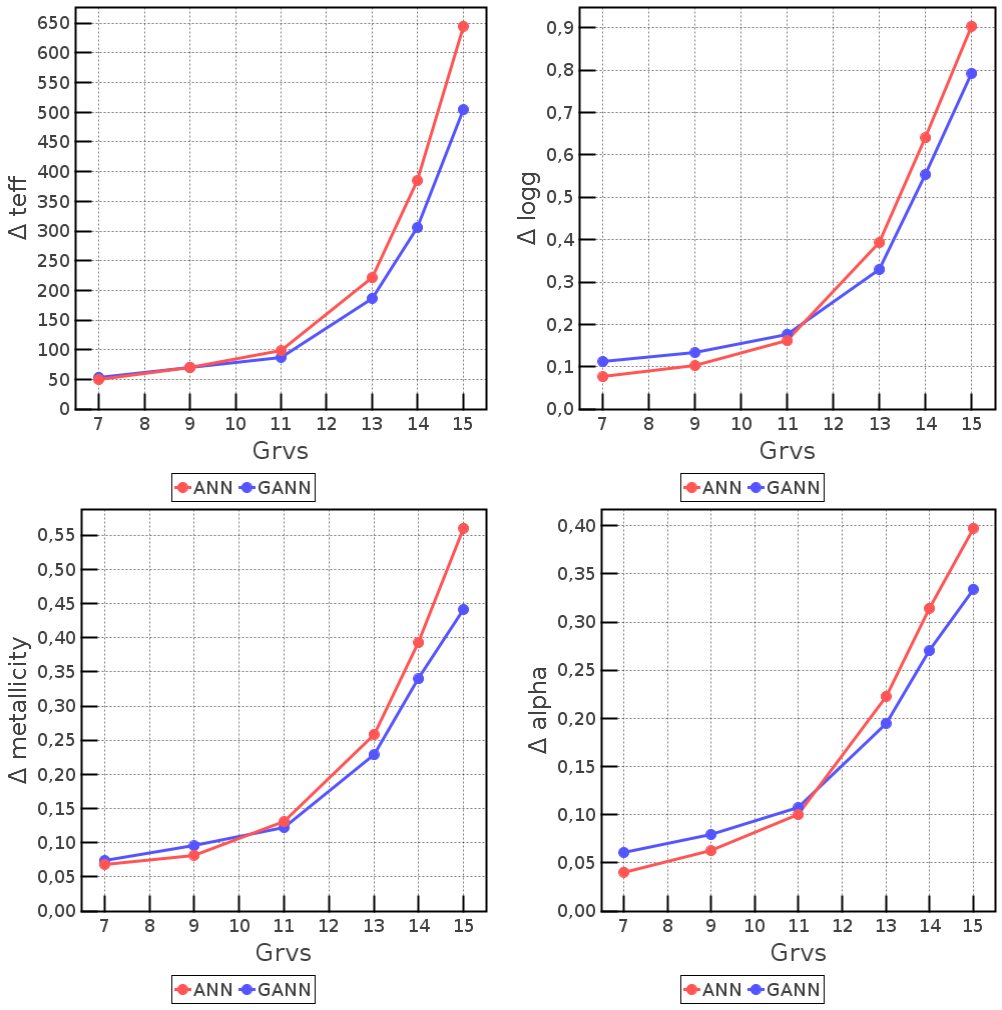


FIGURA 2.13: Percentil 70 de los residuos absolutos para estrellas del conjunto FGK, comparativa entre ANNs y GANNs.

estrella dada. La ventaja de usar una distribución a posteriori no paramétrica, en lugar de asumir una distribución normal u otra cualquiera, es que de esta forma no se fuerza a que los intervalos de confianza tengan una forma concreta, que puede no corresponderse con la realidad. Idealmente, los intervalos de confianza estimados se corresponderán con el error cometido por la ANN generativa, a la hora de estimar un conjunto de APs óptimos. Las figuras 2.15 y 2.16 muestran un ejemplo de cálculo de intervalos de confianza en la estimación de $[Fe/H]$ para cada estrella del conjunto de test de estrellas tipo A, con $G_{rvs} = 7$ y $G_{rvs} = 11$. En dicha figura, para cada estrella, primero se calcula la distribución a posteriori de los APs estimados, la cual es entonces marginalizada, para obtener la distribución a posteriori del $[Fe/H]$ estimado. A partir de la distribución marginal de $[Fe/H]$, se calculan los intervalos de confianza (asimétricos), tanto al 70% como al 90%. Lo que se observa en la figura, son el límite superior e inferior de los intervalos de confianza, colocados alrededor del valor deseado de $[Fe/H]$, con el objetivo de comparar visualmente los CIs con los errores obtenidos en la estimación mediante

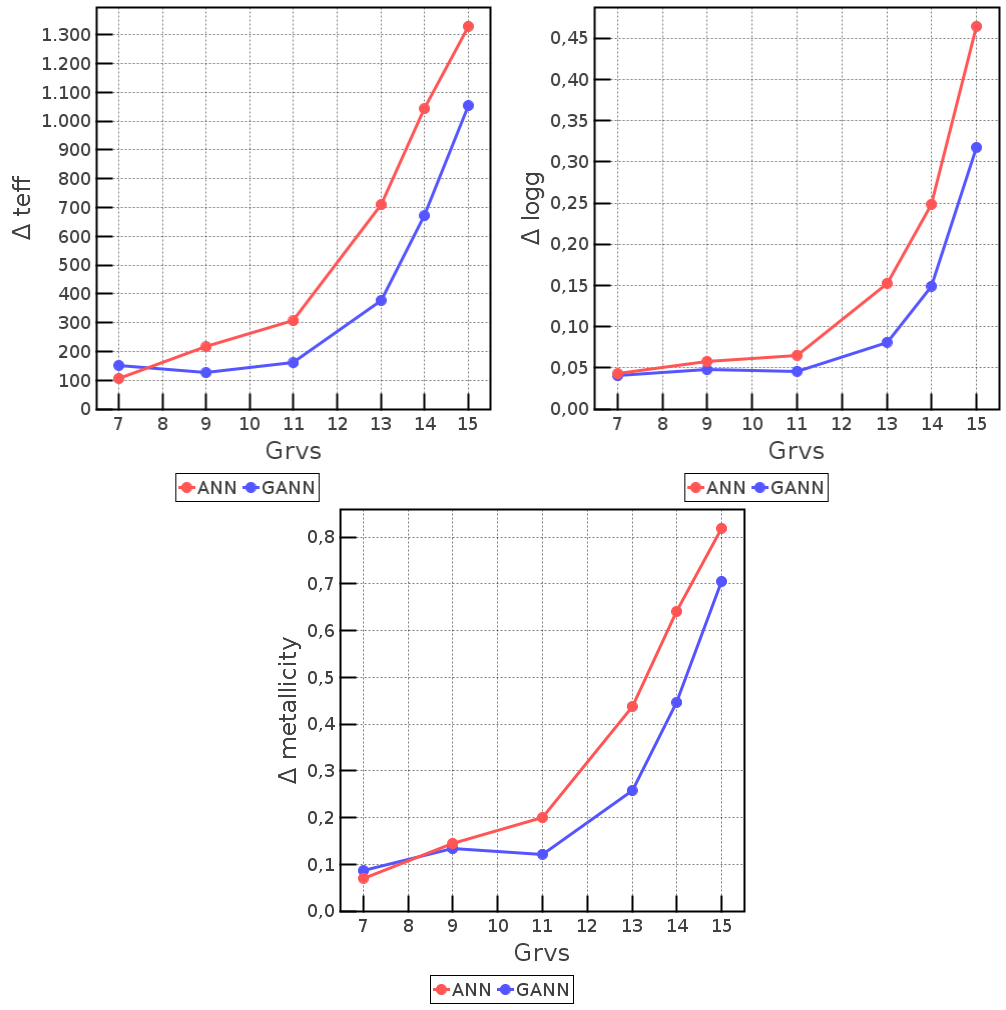
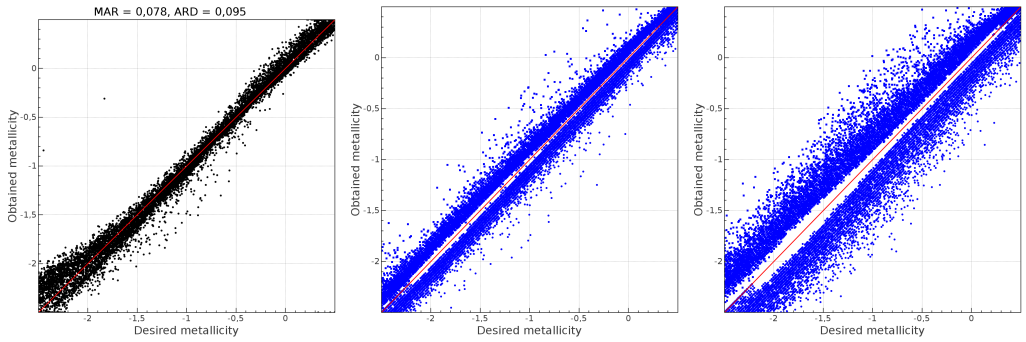


FIGURA 2.14: Percentil 70 de los residuos absolutos para estrellas del conjunto A, comparativa entre ANNs y GANNs.

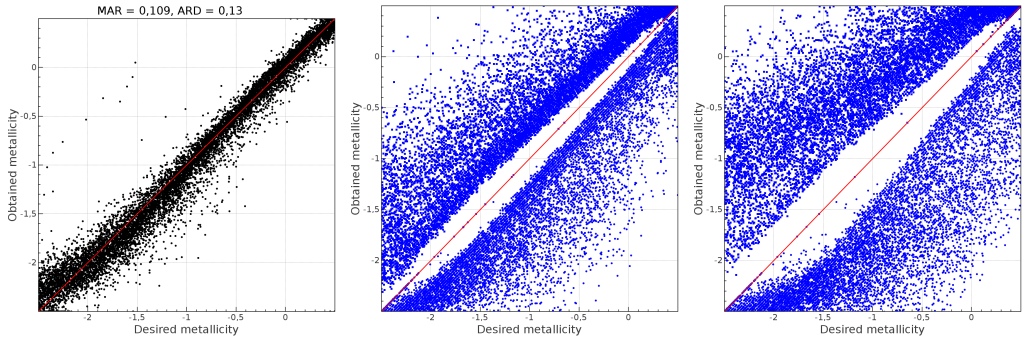
GANNs. Como se puede observar, los intervalos obtenidos se ajustan bastante bien con los errores en la estimación. Además, éstos varían con la entrada X, siendo la incertidumbre mayor para las estrellas con menor metalicidad. También se observa que la incertidumbre es mayor para estrellas más débiles, como cabría esperar debido a su menor SNR. Por último, se observa que los intervalos obtenidos no son simétricos, como tampoco lo son los errores de las estimaciones de APs para este conjunto de estrellas.

Uno de los mayores problemas a los que un sistema de aprendizaje supervisado debe enfrentarse es la gestión de valores atípicos o inesperados. En misiones astronómicas como Gaia, es muy probable que exista un desajuste entre las observaciones y las simulaciones utilizadas para caracterizar los instrumentos. Además, se espera la observación de astros de naturaleza desconocida, dado que el satélite observará todo el cielo de forma no sesgada. Por lo tanto, es necesario establecer un sistema de detección de objetos novedosos, concepto el cual será tratado con más profundidad en el capítulo 3. En el caso del módulo ANN, la función de verosimilitud $P(S|AP)$, puede utilizarse como



(a) $G_{rvs} = 7$ (b) $CI : 70\%$, $G_{rvs} = 7$ (c) $CI : 90\%$, $G_{rvs} = 7$

FIGURA 2.15: Intervalos de confianza (al 70% y 90%) para la estimación de $[Fe/H]$ de estrellas A con $G_{rvs} = 7$.



(a) $G_{rvs} = 11$ (b) $CI : 70\%$, $G_{rvs} = 11$ (c) $CI : 90\%$, $G_{rvs} = 11$

FIGURA 2.16: Intervalos de confianza (al 70% y 90%) para la estimación de $[Fe/H]$ de estrellas A con $G_{rvs} = 11$.

medida de bondad del ajuste entre la simulación y la observación. Dicho procedimiento puede resultar también de utilidad a la hora de mejorar activamente el conjunto de entrenamiento sobre el que se generan los modelos de regresión.

En los tests realizados con ANNs generativas se ha asumido una distribución a priori $P(AP)$ uniforme en los APs. Sin embargo, otras formas de distribución podrían utilizarse como priores. Por ejemplo, se podría establecer una distribución normal centrada en los parámetros aportados por GSP-Phot. También se podría hacer uso de la astrometría a tal efecto. Esta información a priori tendría la función de mejorar las estimaciones en casos donde el espectro RVS no tiene calidad suficiente para la obtención de unos APs confiables.

2.4.5 Estimaciones de APs obtenidas por el módulo ANN

En esta sección, nos centraremos en la configuración óptima para las ANNs, mediante la selección de los mejores parámetros para la ANN y de la mejor representación del

TABLA 2.3: Selección de la mejor configuración en función del tipo de estrella y la magnitud.

Conjunto	**Magnitud**	**Configuración**
A	7	ANN-A2
A	9	ANN-A3
A	11	GANN
A	13	GANN
A	14	GANN
A	15	GANN
FGK	7	ANN-A2
FGK	9	ANN-A2
FGK	11	ANN
FGK	13	ANN-A2
FGK	14	ANN-A3
FGK	15	ANN-A3

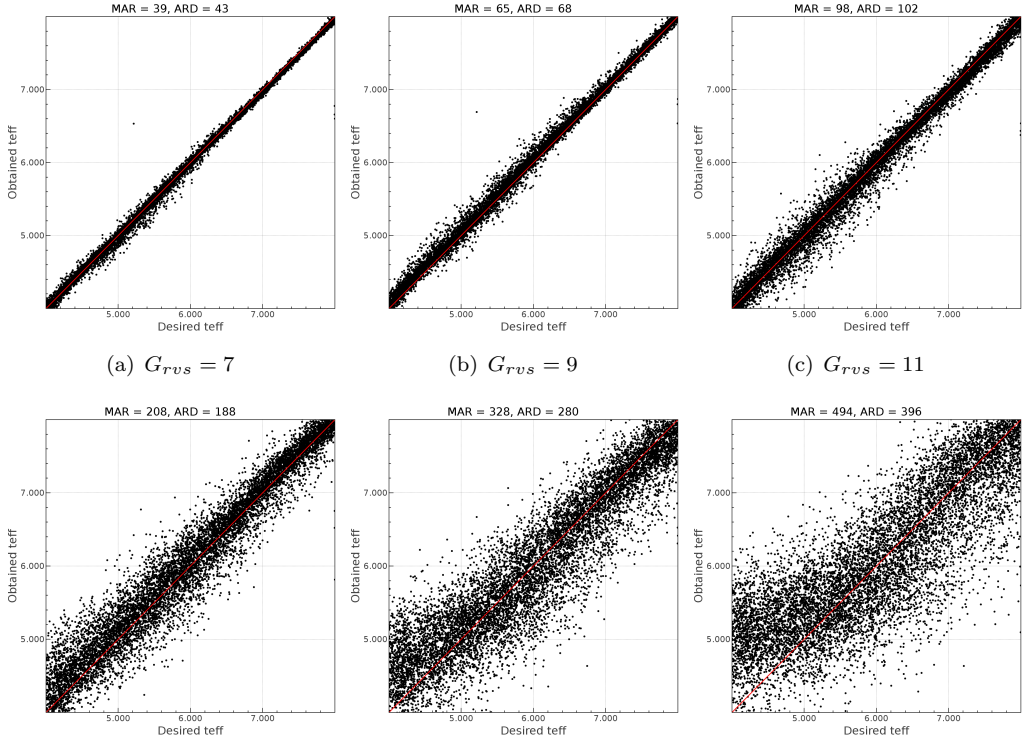
espectro de entrada en cada caso. Esto es, para cada magnitud G_{rvs} y tipo de estrella, seleccionamos la configuración de la red neuronal, en función del error obtenido cuando la misma se aplica al correspondiente al conjunto de validación. La tabla 2.3 muestra dicha selección. Se observa que la mejor representación de la red neuronal depende tanto del tipo de estrella como de la magnitud de la misma.

A continuación, se estudian en más detalle los residuos obtenidos por las configuraciones óptimas de la red neuronal en cada caso. Las figuras 2.17, 2.18, 2.19 y 2.20 muestran las estimaciones obtenidas para el conjunto de estrellas FGK, mientras que las figuras 2.21, 2.22 y 2.23 muestran las mismas para el caso del conjunto de estrellas de tipo A. Adicionalmente, en las gráficas se indica el residual absoluto medio (MAR) y la desviación del residual absoluto (ARD). Como puede observarse, las estimaciones se degradan con el aumento de magnitud de las estrellas, debido al mayor nivel de ruido en los espectros de las mismas. Se puede observar también que se obtienen estimaciones más precisas de T_{eff} y $[Fe/H]$ para estrellas del conjunto FGK en comparación con las obtenidas para estrellas del conjunto A. Sin embargo, la estimación de $log\, g$ es más precisa en el caso de estrellas tipo A, gracias a las líneas de Paschen. Podemos observar las siguientes tendencias con respecto a la distribución de los residuos de cada parámetro.

- El parámetro T_{eff} tiende a ser sobrestimado en el caso de estrellas frías de tipo K, mientras que tiende a ser subestimado en el caso de las estrellas de tipo A más calientes. Dicha tendencia es más pronunciada con magnitud $G_{rvs} > 13$.
- El parámetro $log\, g$ se subestima sistemáticamente en el extremo superior (estrellas enanas) para las estrellas FGK más débiles.

- En la estimación de $[Fe/H]$, el parámetro se sobrestima de forma sistemática en el extremo inferior en el caso de estrellas brillantes, lo cual está enmascarado por el ruido en el caso de estrellas más débiles. Esta tendencia se atribuye al bajo contenido de lineas metálicas significativas en estrellas de tipo tardío.
- El parámetro $[\alpha/Fe]$ se subestima de forma sistemática en el extremo superior para estrellas con magnitud $G_{rvs} > 13$.

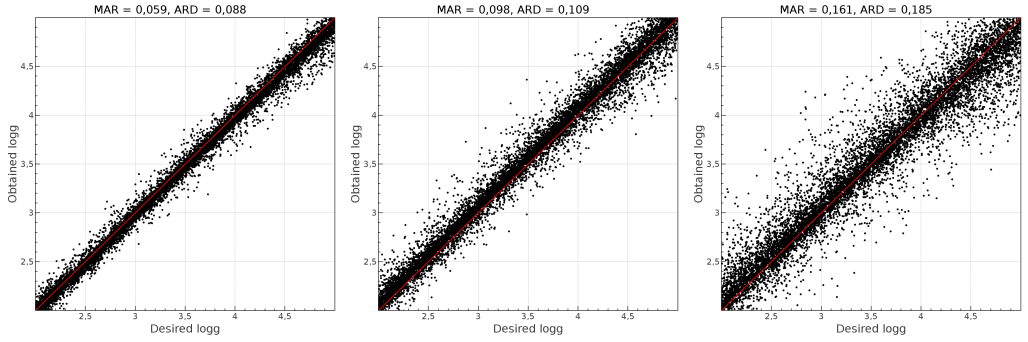
Se puede concluir que las tendencias observadas en los residuos se pueden explicar por la naturaleza de los datos y no por errores sistemáticos del algoritmo utilizado. En cuanto a la utilidad científica de las parametrizaciones obtenidas por *ANN*, ésta depende del tipo de estudio que se vaya a realizar y las metas del mismo. A pesar de ello, se ha descrito en [38] que sería deseable una precisión en los APs del orden de $\Delta T_{eff} < 200$, $\Delta log\, g < 0.3$ y $\Delta[Fe/H] < 0.2$, en el caso de estrellas FGK. Por lo tanto, podemos decir que el módulo *ANN* producirá APs útiles para todas las estrellas con $G_{rvs} < 13$, lo cual incluye un número aproximado de 10^7 estrellas.



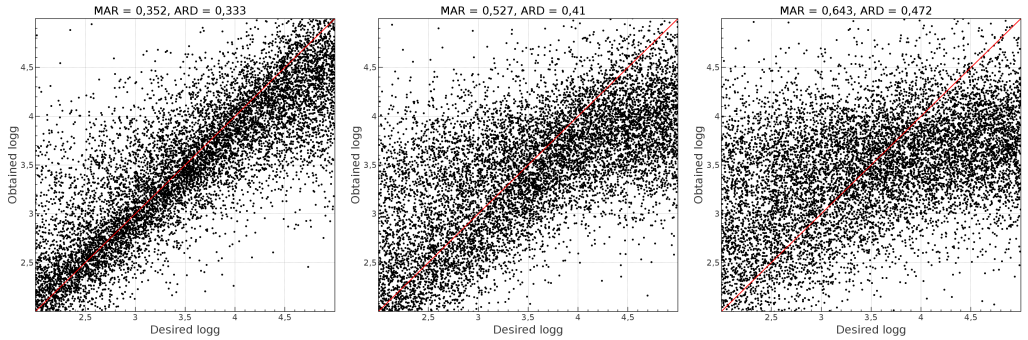
(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

FIGURA 2.17: Estimaciones de T_{eff} para estrellas FGK.



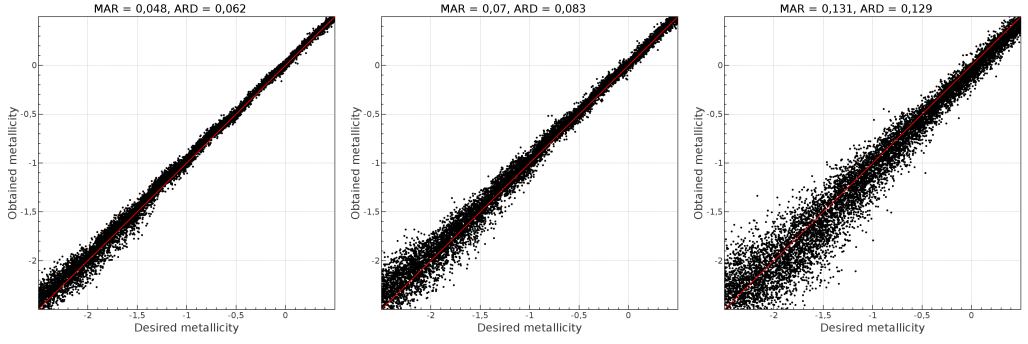
(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$



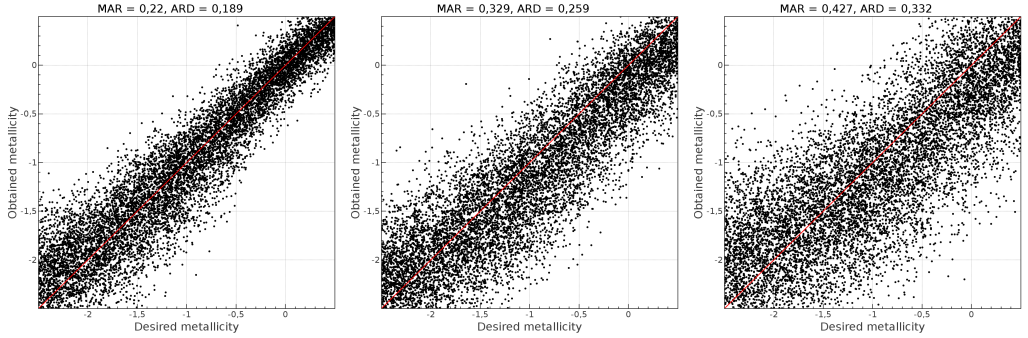
(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

FIGURA 2.18: Estimaciones de $log\, g$ para estrellas FGK.

.

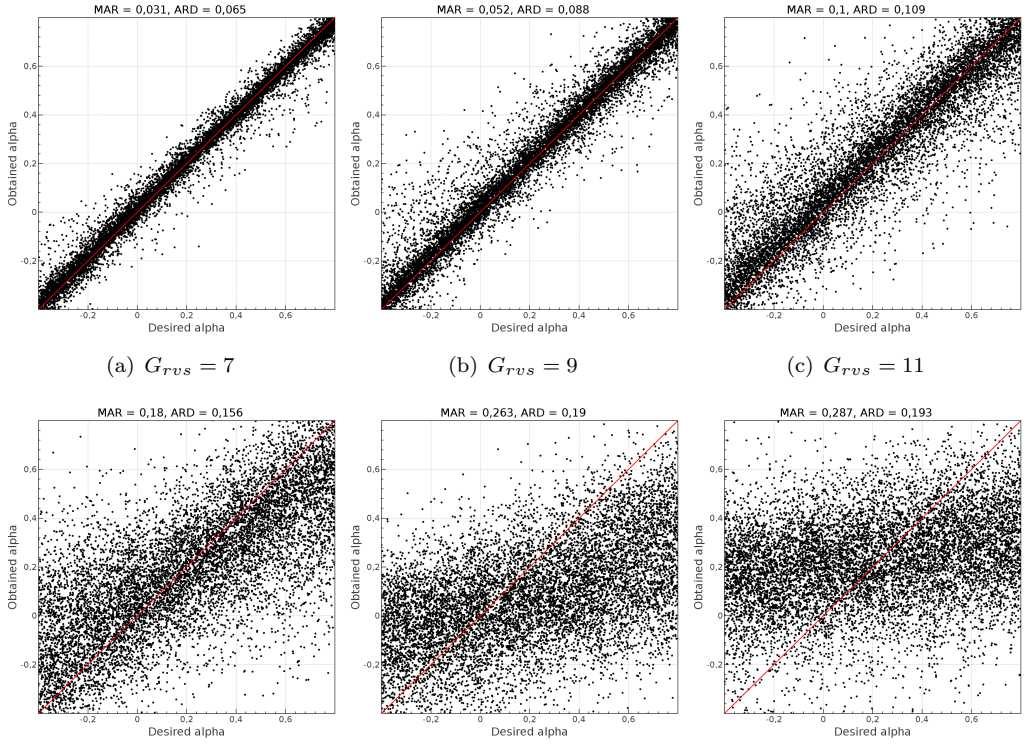


(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$



(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

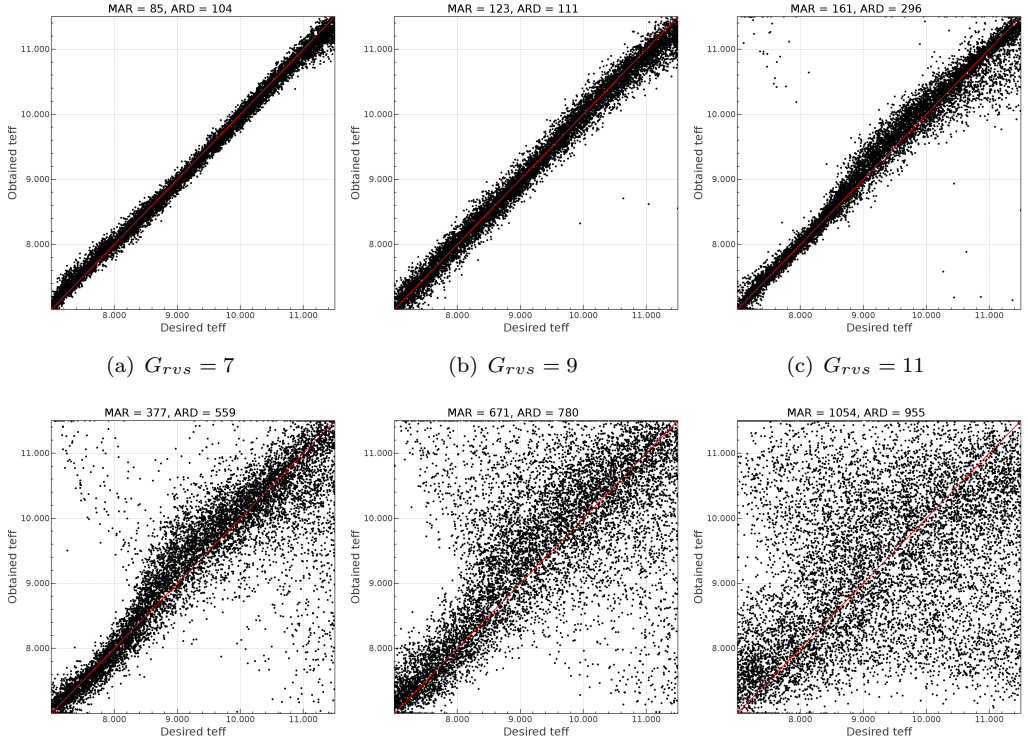
FIGURA 2.19: Estimaciones de $[Fe/H]$ para estrellas FGK.



(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

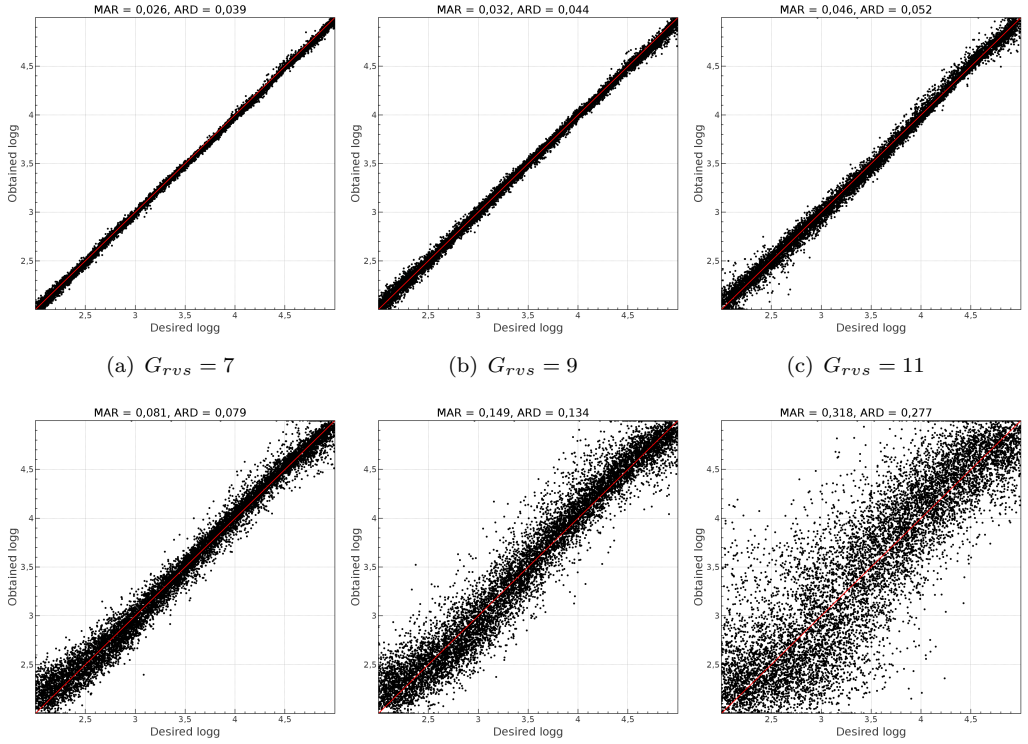
FIGURA 2.20: Estimaciones de $[\alpha/Fe]$ para estrellas FGK.



(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

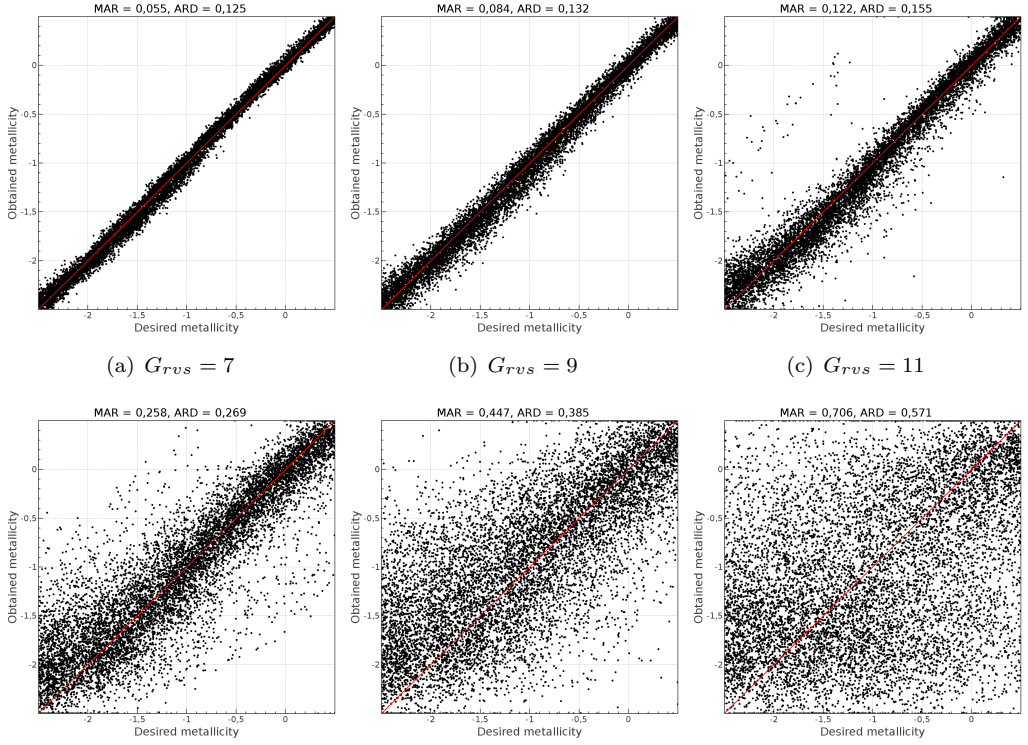
FIGURA 2.21: Estimaciones de T_{eff} para estrellas A.



(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

FIGURA 2.22: Estimaciones de $log\ g$ para estrellas A.



(a) $G_{rvs} = 7$ (b) $G_{rvs} = 9$ (c) $G_{rvs} = 11$

(d) $G_{rvs} = 13$ (e) $G_{rvs} = 14$ (f) $G_{rvs} = 15$

FIGURA 2.23: Estimaciones de $[Fe/H]$ para estrellas A.

2.4.6 Implementación

Todos los métodos expuestos aquí han sido implementados y testados en el lenguaje de programación Java, con el objetivo de realizar las comparativas en una misma plataforma estable y homogénea. El código correspondiente a la definición y entrenamiento de ANNs se encuentra en una librería denominada *NeuralToolkit*. Dicha librería incluye pruebas de unidad para comprobar el correcto funcionamiento de los componentes de la ANN, así como ejemplos visuales que permiten comprobar el poder de ajuste de las ANNs mediante varios ejemplos sencillos. Por otro lado, tanto el algoritmo PSO como el Metrópolis-Hastings se han integrado en la librería *OptimizationToolkit*, la cual también incluye tests de unidad dónde se presentan varios problemas sencillos de optimización que los algoritmos deben resolver. Por último, los tests relacionados con el manejo de espectros RVS y la estimación de APs se han implementado en la librería *GSPSpecNNTests*, la cual incorpora las dos librerías anteriores para realizar sus operaciones.

La evaluación de la función de verosilimlitud definida en 2.2 resulta especialmente problemática. La razón de ello es que la distancia D, entre el espectro RVS observado y el estimado, puede llegar a tener valores elevados, del orden de 10^4, en el caso de espectros HR. Debido a ello, al calcular la exponencial negativa de la distancia, nos encontramos con que la precisión en coma flotante de Java no es suficiente. Para paliar este problema, primero se calculan todas las distancias D implicadas en el problema, para luego normalizarlas al intervalo $[0, 1]$, de forma de las diferencias relativas se mantienen y la exponencial resulta calculable.

Actualmente, el módulo *ANN* no ha sido integrado en APSIS por motivos técnicos. Sin embargo, se prevé su integración en 2014, pasando a engrosar la lista de algoritmos integrados en GSP-Spec.

Capítulo 3

Minería de objetos atípicos mediante clasificación no supervisada

En este capítulo, trataremos el problema de clasificación de objetos astronómicos en CU8, que será llevada a cabo principalmente por los paquetes de trabajo Discrete Source Classifier (DSC) y Object Cluster Analysis (OCA). Este paquete utiliza clasificación supervisada para asignar cada fuente observada a una de entre las clases predefinidas, listadas en la tabla 1.2. Dado que Gaia realizará, por primera vez, una observación no sesgada de todo el cielo hasta magnitud veinte, se espera que un gran número de objetos no pueda ser asignado a una de las clases predefinidas, debido a su naturaleza novedosa o a problemas en su adquisición. Por lo tanto, en el diseño de DSC, se han incorporado desde el principio los conceptos de detección de objetos atípicos (*outliers* en inglés) y aprendizaje activo. DSC incorpora un sistema de detección de objetos atípicos, el cual rechaza objetos que no entran dentro de las clases de objetos esperadas. Se estima que acerca del 5% de los objetos observados por Gaia sean rechazados por dicho detector, con lo que se espera un número aproximado de 50 millones de observaciones atípicas. El análisis manual de semejante cantidad de objetos resultaría completamente inviable, por lo que se requiere de algún tipo de análisis automatizado. Este capítulo describe un sistema automático de análisis de observaciones atípicas, integrado en el paquete de trabajo llamado Outlier Analysis (OA). Dada la naturaleza del problema, se aplicarán técnicas de aprendizaje no supervisado y minería de datos, con el objetivo de ayudar a que los investigadores desvelen la naturaleza de los objetos más extravagantes que observará Gaia.

3.1 Estado del arte

3.1.1 El paquete de trabajo DSC

DSC es, junto con OCA, el principal paquete de clasificación en CU8, ya que clasificará todas las fuentes puntuales observadas por Gaia. Se basa en clasificación supervisada mediante algoritmos de aprendizaje máquina. DSC se compone de tres subclasificadores: uno basado en espectrofotometría BP/RP, otro basado en paralajes y movimientos propios y un último basado en posiciones y magnitud G de los astros. Además de proporcionar la probabilidad de pertenencia de cada astro a cada clase predefinida, DSC también proporciona un conjunto de probabilidades combinadas a partir de las obtenidas por los subclasificadores. De los tres subclasificadores existentes, el basado en espectrofotometría BP/RP es el que proporciona una información más completa sobre los objetos y, al mismo tiempo, el más complejo. Esta complejidad viene dada por un alto número de características (dimensiones) en la entrada del algoritmo, así como por la carencia de conjuntos de datos que permitan la construcción de un conjunto de entrenamiento sólido. Esta carencia se debe a que los catálogos espectroscópicos existentes están limitados a ciertos tipos de objetos y a regiones limitadas del cielo. Gaia será un catálogo único, en el sentido de que ofrecerá espectroscopía no sesgada de todo el cielo hasta magnitud veinte. Además, los instrumentos con los que se equipa Gaia tienen un conjunto de características diferenciadas, sujetas a incertidumbre en operación real, causada por el efecto de diversos factores observacionales. Por esta razón, el consorcio DPAC, ha realizado un gran esfuerzo para llevar a cabo una gran simulación, a través del simulador GOG [39], de los distintos tipos de objetos conocidos, tal y como serán observados por Gaia.

El clasificador espectrofotométrico integrado en DSC se basa en la aplicación de SVMs, especializadas para clasificar fuentes con un SNR dado. Se trata de un sistema jerárquico en varios niveles. En primer lugar, los objetos son clasificados por un clasificador entrenado con un conjunto de astros restringido, cuya física es bien conocida. Dicho clasificador incorpora un detector de objetos atípicos, el cual determina si un objeto puede ser procesado o no. En caso contrario, el objeto que está siendo procesado pasa a la siguiente capa, en la cual se aplica un clasificador entrenado con un conjunto más amplio de objetos provenientes de simulaciones con modelos astrofísicos. De nuevo, los objetos pueden ser rechazados por un detector de objetos atípicos, en cuyo caso los mismos se marcarán como outliers. Por otro lado, los objetos que sí han sido clasificados reciben un vector de probabilidades con un valor por clase predefinida. En el caso de que las probabilidades de pertenencia sean bajas, resulta complicado asignar una clase

concreta. Este último tipo de objetos también será objeto de análisis por parte del paquete OA.

3.1.2 Detección de anomalías y aprendizaje activo

Las técnicas de clasificación supervisada requieren la especificación de un conjunto de entrenamiento sobre el cual aprenden un modelo que servirá para predecir la clase de nuevas observaciones. Sin embargo, cualquier dominio de aplicación real está sometido a cierta incertidumbre, provocada por entornos cambiantes y por problemas en la adquisición de los datos. Dicha incertidumbre puede provocar que la clasificación asignada por el modelo resulte espuria, ya que el mismo ha sido computado en un entorno con condiciones controladas. Por lo tanto, es necesario detectar casos en los cuales la clasificación automática no es fiable. Para ello, surgieron los llamados métodos automáticos de detección de anomalías, que permiten seleccionar objetos que no pueden ser clasificados con fiabilidad debido a su carácter novedoso. Estos objetos atípicos son entonces analizados por un humano experto, el cual puede tomar diversas decisiones, por ejemplo marcar un objeto como erróneo o bien sugerir una actualización del modelo de clasificación. Este modelo adaptativo de clasificación se denomina como "aprendizaje activo" y ha sido objeto de estudio en los últimos años, surgiendo varias estrategias y guías para su puesta en funcionamiento. Siguiendo la taxonomía descrita en [40], existen tres tipos diferentes de detección de objetos atípicos:

- El tipo 1 se basa en el modelado de la distribución de las observaciones sin conocimiento a priori, mediante un proceso de clasificación no supervisada o agrupamiento. A partir de los grupos obtenidos, se identifican los objetos que sobresalen de la distribución normal. Dicho proceso de agrupamiento puede ser efectuado mediante algoritmos como los mapas auto-organizativos (Self-Organizing Maps, SOM) o las mezclas de gaussianas (Gaussian Mixture Model, GMM), véase [41]. La desventaja de este tipo de detección de objetos atípicos es la gran cantidad de cómputo requerida en el caso de grandes conjuntos de datos. Además, requiere que todos lo datos estén disponibles antes de que el método sea aplicado.

- El tipo 2 se basa en un sistema de clasificación supervisada que distingue entre lo que se considera normal y lo que se considera atípico. Dentro de lo que se considera normal, pueden existir varias clases, con lo que se trataría de un sistema de clasificación multiclase, siendo una de ellas la correspondiente a los objetos atípicos. Desgraciadamente, la definición de un conjunto de entrenamiento que represente lo atípico es muy complicado en una situación realista, por lo que la aplicación de este tipo de detectores es muy compleja.

- El tipo 3 modela tan sólo el comportamiento normal del sistema, mediante el establecimiento de una envoltura, de forma que los objetos que caen fuera de ella se marcan como atípicos. Estos sistemas se suelen denominar en la literatura como "sistemas de detección de objetos novedosos" ya que detectan casos no contemplados en el conjunto de entrenamiento. Existen varias técnicas para el modelado del comportamiento normal, principalmente técnicas estadísticas o basadas en redes neuronales. Una técnica muy popular para generar la envoltura del conjunto de entrenamiento se basa en la aplicación de SVMs uniclase, véase [42]. Dicha técnica es la que ha sido seleccionada por DSC para realizar su proceso de detección de anomalías.

Los sistemas de detección de objetos atípicos creados en los años 90 fueron los precursores del paradigma llamado "aprendizaje activo", en el cual el conjunto de entrenamiento de un clasificador supervisado se construye dinámicamente. Para construir el conjunto de entrenamiento, el sistema de aprendizaje detecta patrones novedosos, entre los patrones sin etiquetar presentados al mismo, y lanza una consulta a un *oráculo* (normalmente un humano experto en el dominio) para que etiquete dicho patrón, el cual pasará a formar parte del conjunto de entrenamiento. Este tipo de sistemas permite el aprendizaje con un conjunto de entrenamiento reducido, además de permitir la adaptabilidad del sistema de clasificación en entornos cambiantes e incluso mejorar la precisión del clasificador. Son especialmente recomendables para casos en los que la adquisición de etiquetas es un proceso costoso. Una excelente revisión de los métodos y estrategias relacionadas con el aprendizaje activo puede encontrarse en [43] y las referencias que ahí se encuentran.

3.1.3 Aprendizaje no supervisado

Aprendizaje no supervisado se refiere a la generación de un modelo que explique un conjunto de observaciones sin ningún conocimiento a priori. Se trata de un concepto relacionado con la estimación de densidad estadística, aunque abarca un abanico más amplio de técnicas, muchas de ellas frecuentemente utilizadas en la minería de datos. Las dos subramas principales de este tipo de aprendizaje son la clasificación no supervisada y la reducción de dimensionalidad. A continuación, se revisan los trabajos realizados hasta la fecha en los que se aplica aprendizaje no supervisado para resolver problemas en astrofísica:

- El problema de agrupamiento ha sido ampliamente estudiado, tanto desde el aspecto teórico como desde su implementación práctica. En dicho estudio, convergen autores provenientes de diferentes disciplinas como las ciencias de la

computación, la estadística y el álgebra lineal. La gran mayoría de los algoritmos desarrollados para resolver este problema coinciden en la explotación del concepto de similitud entre objetos. De esta forma, los algoritmos buscan identificar un conjunto de grupos de objetos que forman una partición disjunta del conjunto de datos de entrada. Dichos grupos se generan con el objetivo de maximizar la similitud entre los objetos pertenecientes a un mismo grupo y, al mismo tiempo, minimizar la similitud entre objetos pertenecientes a diferentes grupos. Existen cientos de algoritmos de agrupamiento en la bibliografía, adoptando distintas perspectivas. Una revisión de dichos algoritmos puede encontrarse en [44] y las referencias que allí se encuentran. En el campo de la astronomía y la astrofísica, la aplicación de algoritmos de agrupamiento es un concepto relativamente nuevo, pese a tratarse de un campo en el que la estadística tiene un gran número de aplicaciones. Los trabajos pioneros en este sentido fueron los presentados por Sánchez Almeida et al. en esta misma década [45] y [46], donde se abarca la clasificación no supervisada, respectivamente, de espectros de galaxias y estrellas provenientes del catálogo SDSS. En dichos trabajos, los autores utilizan el algoritmo *k-means*, debido a su sencillez y rapidez de cómputo, para clasificar cientos de miles de espectros en media resolución. Como resultado, se obtiene que el conjunto de espectros puede representarse mediante un conjunto reducido de grupos, de forma que los objetos pertenecientes a cada grupo tienen propiedades físicas similares.

- La reducción de dimensionalidad se define como el proceso de disminuir el número de variables involucradas en la resolución de un problema en concreto, de forma que el mismo puede ser abarcado con mayor sencillez, evitando los problemas provocados por la alta dimensionalidad. Dicha reducción puede realizarse mediante la selección de características o mediante la extracción de características. La selección de características se apoya normalmente en la supervisión, por lo que en el campo del aprendizaje no supervisado se aplican técnicas de extracción de características. Una técnica de extracción de características clásica es el análisis de componentes principales (PCA, por sus siglas en inglés), el cual ha sido ampliamente utilizado en diversos campos, incluyendo la astrofísica. La técnica PCA es una transformación ortogonal que proyecta un conjunto de observaciones multivariantes en un conjunto reducido de componentes principales, los cuales son linealmente independientes. Estos componentes se computan de forma que representan la mayor cantidad de variabilidad posible del conjunto de observaciones. El PCA puede aplicarse con diferentes metas de análisis de datos, como la búsqueda de correlaciones en los datos, la visualización en un espacio bidimensional o tridimensional o la compresión de datos. Sin embargo,

el PCA carece que la capacidad de captar relaciones no lineales entre variables del espacio de entrada. En los últimos años, han surgido diversas técnicas que intentan solventar este problema, como por ejemplo las PCA con Kernel, la llamada *Local Linear Embedding (LLE)* o las técnicas de *Multidmensional Scaling (MDS)*. En el campo de la clasificación espectral, LLE ha sido aplicado por Vanderplat et al. en [47], donde se presenta como una técnica más adecuada que PCA a la hora de representar espectros de SDSS en un espacio tridimensional, el cual puede ser visualizado fácilmente por expertos en el dominio. En dicho trabajo, se discute la utilidad de este tipo de técnicas para detectar y analizar errores de clasificación de un modo intuitivo.

Las redes neuronales también contemplan el aprendizaje de tipo no supervisado. Las principales redes neuronales no supervisadas son las redes ART, Neural Gas y los mapas auto-organizados o SOM. Dichas redes neuronales no supervisadas siguen un proceso de aprendizaje similar a los algoritmos de agrupamiento típicos, basándose en el ajuste de modelos en función de una medida de similitud. No obstante, los SOM proporcionan una característica adicional, y es la proyección no lineal del conjunto de observaciones en un espacio de dimensionalidad reducida, de forma similar a los algoritmos LLE y MDS. Por lo tanto, los SOM aúnan las dos principales ramas del aprendizaje no supervisado: agrupamiento y reducción de dimensionalidad. Esta propiedad de los SOM ha guiado los trabajos que se describen en este capítulo, realizados en análisis de objetos atípicos de clasificación por Ordóñez et al. en [48] y Fustes et al. en [49] y [50], en los que se demuestra su capacidad para desvelar las propiedades de un conjunto de espectros cuando se tiene poco o ningún conocimiento a priori. En este capítulo se profundiza en la aplicación de SOMs, y otros algoritmos de aprendizaje no supervisado, para el análisis de objetos atípicos en bases de datos astronómicas, en concreto en la que se obtendrá con la misión Gaia.

3.2 DU-836: OA

El paquete de trabajo OA se encarga de analizar las observaciones atípicas de la misión Gaia. En este caso, se define una observación como atípica cuando ésta no puede ser clasificada con fiabilidad como una de las clases de objetos esperadas. Debido a las características de la misión, es esperable una gran cantidad de objetos de este tipo, por lo que se requiere un análisis automatizado. OA puede utilizar cualquier tipo de información para realizar su análisis. No obstante, la información astrofísica más completa (y compleja) de los astros en Gaia se obtiene a través de la espectrofotometría BP/RP, la cual estará disponible para todos los objetos observados

por Gaia. Adicionalmente, la astrometría y la variabilidad de las fuentes pueden aportar información sobre las mismas, aunque no son tan determinantes para su clasificación. Por otro lado, los espectros RVS tan sólo serán obtenidos para los astros más brillantes, los cuales se conocen mejor y serán clasificados con mayor facilidad. Por lo tanto, la operación de OA depende principalmente de la calibración fotométrica de CU5 y de los resultados obtenidos por los paquetes de clasificación en CU8, cuyos principales representantes son DSC y OCA.

Los algoritmos integrados en OA son del tipo no supervisado, ya que cualquier modelo basado en conocimiento previo sería incapaz de procesar datos que por definición no siguen las reglas establecidas. De este modo, el objetivo de OA no es la reclasificación de las fuentes, sino el cómputo de un conjunto de modelos que describan los datos, de forma que los mismos puedan ser estudiados por expertos en el dominio. Estos modelos deben ser capaces de sintetizar la información de la forma más concisa y descriptiva posible. Además, el paquete OA debe desarrollar algoritmos que sean viables computacionalmente, teniendo en cuenta que se esperan millones de observaciones atípicas y las restricciones del hardware disponible. A continuación, se estudiarán una amplia variedad de algoritmos no supervisados, los cuales forman parte del estado del arte, con el objetivo de seleccionar los más adecuados para ser integrados en OA. En la evaluación de los algoritmos, se tendrá en cuenta su capacidad para sintetizar las observaciones, la comprensibilidad de los modelos generados por los mismos y su complejidad computacional.

3.3 Simulación de espectrofotometría BP/RP

Con el objetivo de desarrollar los algoritmos de CU8, se han generado un conjunto de simulaciones espectrofotométricas que cubre un amplio abanico de objetos astrofísicos. Dicha simulación se ha desarrollado en colaboración con CU2. Las librerías que forman parte de la simulación han sido generadas tanto mediante modelos físicos como a través de catálogos observacionales, convertidos al formato BP/RP usando el simulador GOG. Cada librería contiene varias versiones de los espectros BP/RP, escalados en magnitud y con un cierto nivel de ruido añadido, para simular lo que se obtendría al observar con Gaia los astros, si éstos tuvieran una determinada magnitud aparente en la banda G. Además, las librerías se dividen en un conjunto "NOMINAL" y otro "RANDOM". El conjunto "NOMINAL" contiene objetos con un conjunto regular de APs, constituyendo el conjunto de entrenamiento para algoritmos como GSP-Phot (el principal algoritmo de parametrización de estrellas mediante espectros BP/RP), mientras que el conjunto "RANDOM" contiene objetos simulados con un conjunto aleatorio de APs, con el

objetivo de ejercer como conjunto de test. Por último, aparte de los APs intrínsecos de los astros, se ha tenido en cuenta la extinción interestelar o enrojecimiento como un AP a mayores, ya que la misma no será corregida previamente y tendrá que ser estimada por los paquetes de CU8 junto con los APs intrínsecos. A continuación, se describen las principales librerías que forman parte de la simulación del ciclo 8:

- *Main Stellar Library:* Esta es la principal librería de espectros estelares sintéticos, compuesta por espectros simulados mediante los modelos de atmósferas estelares PHOENIX y MARCS, descritos en [51] y [52]. Las estrellas contempladas en la librería cubren las clases O, B, A, F, G, K y M según el esquema de clasificación MK, y se han tenido en cuenta los APs: T_{eff}, $log\, g$ y $[Fe/H]$, a la hora de simular sus espectros. Existen dos versiones de la librería: $RAN1$, dónde se incluye el parámetro de extinción A_0 como AP adicional y $RAN2$, dónde se establece $A_0 = 0$. Más información sobre esta librería y el resto de librerías sintéticas estelares de Gaia puede encontrarse en [53].

- *UCD:* Aquí se contemplan simulaciones de estrellas enanas ultra frías, de tipos L, T e Y. Esta librería ha sido generada con una estructura parecida a las incluidas en *Main Stellar Library*. Más información sobre esta librería puede encontrarse en [54].

- *WD:* Enanas blancas, estrellas de masa baja o media en su último estado de evolución, en el cual prácticamente han agotado su combustible nuclear. La librería contempla tanto del subtipo WDA como del subtipo WDB. Puede encontrarse información más detallada en [55].

- *C_STARS:* Librería de estrellas de clase espectral C o estrellas de carbono. Las estrellas de carbono son estrellas de tipo tardío, cuya atmósfera contiene más carbono que oxígeno. Esta librería se ha descrito con más detalle en [52].

- *Be:* En esta librería se contemplan estrellas azules muy calientes, cuyos espectros se caracterizan por mostrar líneas de emisión bien marcadas correspondientes al carbono, nitrógeno y oxígeno. Más información puede encontrarse en [56].

- *WR:* Estrellas Wolf-Rayet. Son estrellas muy poco usuales ya que sus atmósferas contienen helio en el lugar de hidrógeno. Se consideran supergigantes moribundas cuyo hidrógeno está siendo expulsado por vientos estelares. Más información puede encontrarse en [56].

- *PHY_BIN:* Sistemas binarios formados por dos estrellas ligadas gravitacionalmente. Esta es la librería utilizada por el paquete MSC para desarrollar sus algoritmos. Los objetos pertenecientes a PHY_BIN se han creando

por medio de la combinación de pares de estrellas pertenecientes a la librería *Main Stellar Library*. Más información se puede encontrar en [57].

- *QSO:* Se trata de una librería de quásares sintéticos, simulados en función de varios parámetros: La pendiente del continuo, el ancho equivalente de las líneas de emisión, los parámetros de absorción interestelar (enrojecimiento) A_0, R_0 y el corrimiento al rojo z.

- *GALAXY:* Esta librería se compone de galaxias sintéticas obtenidas mediante el modelo galáctico denominado PEGASE. Se contemplan un gran número de parámetros, como el tipo morfológico de galaxia (temprana, espiral, irregular, starbust), los parámetros de extinción A_0 y A_g, la tasa de formación estelar, el tiempo de acreción y el corrimiento al rojo z. Más información puede encontrarse en [58].

- *NONPHY:* Objetos superpuestos en la misma posición de la bóveda celeste. Los sistemas pueden estar constituidos por una pareja estrella-estrella, estrella-galaxia o estrella-quásar. El espectro combinado se ha construido mediante las librerías *Main Stellar Library*, *GALAXY* y *QSO*.

- *SDSS_STARS:* Librería empírica, compuesta por un conjunto de espectros estelares extraídos del catálogo SDSS DR7 y convertidos al formato BP/RP mediante el simulador GOG.

- *SDSS_QSO:* Al igual que la anterior, se trata de una librería empírica, pero en este caso compuesta por quásares extraídos del catálogo SDSS DR7.

- *GALAXIES_SE:* Esta librería es semi-empírica, ya que se ha construido mediante un ajuste de χ^2 entre una serie de plantillas extraídas del modelo PEGASE y espectros de galaxias pertenecientes al catálogo SDSS DR7. Una vez obtenidos los espectros con buen ajuste, se han convertido al formato BP/RP. Para más información, véase [59].

- *PN*: Se trata de una librería empírica de espectros pertenecientes a nebulosas planetarias, extraídos del catálogo MASH y después convertidos al formato BP/RP. Más información puede encontrarse en [60].

3.4 Técnicas de agrupamiento para simulaciones de Gaia

Como se ha descrito anteriormente, el paquete OA tiene como objetivo la clasificación no supervisada de los espectros BP/RP que se obtendrán mediante la observación

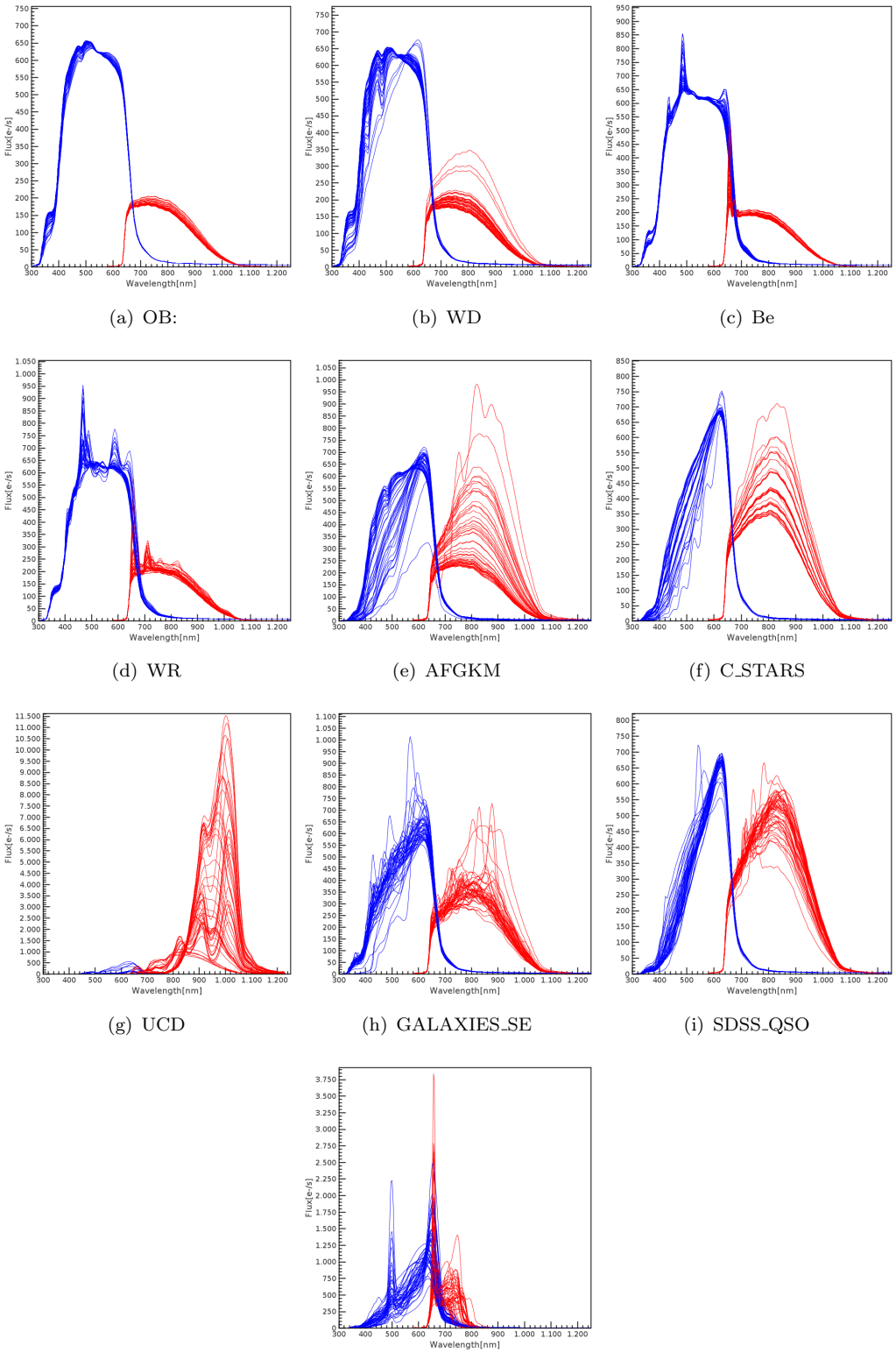
TABLA 3.1: Número de objetos por cada librería seleccionada del conjunto de simulaciones de Gaia.

Librería	**Número de objetos**
OB	9999
WD	20556
Be	522
WR	129
AFGKM	5000
C_STARS	428
UCD	9890
GALAXIES_SE	33670
SDSS_QSO	70554
PN	748

del satélite Gaia. En esta sección, utilizaremos las diferentes librerías que componen la simulación de objetos de Gaia para evaluar la efectividad de distintas técnicas de agrupamiento, con el objetivo de determinar cuál o cuáles son más adecuadas para su integración en OA. Se tendrán en cuenta diversos parámetros en la evaluación, como son la efectividad a la hora de computar grupos homogéneos, la complejidad computacional de los algoritmos y la interpretabilidad de las particiones obtenidas.

Para la evaluación de los métodos de agrupamiento se han seleccionado las siguientes librerías: *MainStellarLibrary* (modelo PHOENIX), *UCD*, *WD*, *C_STARS*, *Be*, *WR*, *WD*, *SDSS_QSO*, *GALAXIES_SE* y *PN*. Dicha selección cubre la mayoría de las clases de objetos astronómicos conocidos. Se han seleccionado librerías observacionales en lugar de sintéticas, con la intención de crear un conjunto de datos lo más realista posible. Sin embargo, en el caso de estrellas normales, se ha preferido la versión sintética, ya que con Gaia se espera obtener una muestra más completa en el espacio de APs que la presente en los catálogos existentes hoy en día. Además, se selecciona el conjunto "RAN" en cada librería sintética, y se escogen espectros sin enrojecimiento añadido, ya que éste perjudicaría el rendimiento de los algoritmos. Esta presunción será utilizada tan solo de cara a evaluar los diferentes métodos. En operación real, el enrojecimiento afectará a los resultados en cierta medida. La figura 3.1 muestra 50 espectros BP/RP seleccionados aleatoriamente de cada una de las librerías que se han escogido para la evaluación, mientras que la tabla 3.1 muestra el número de objetos pertenecientes a cada librería.

Para que la evaluación de algoritmos sea posible, es necesario realizar un preprocesado adecuado de los espectros antes de comenzar el agrupamiento. Además, es preciso definir un conjunto de medidas de evaluación de las particiones obtenidas, lo cual se acomete en las siguientes secciones.



(a) OB: (b) WD (c) Be

(d) WR (e) AFGKM (f) C_STARS

(g) UCD (h) GALAXIES_SE (i) SDSS_QSO

(j) PN

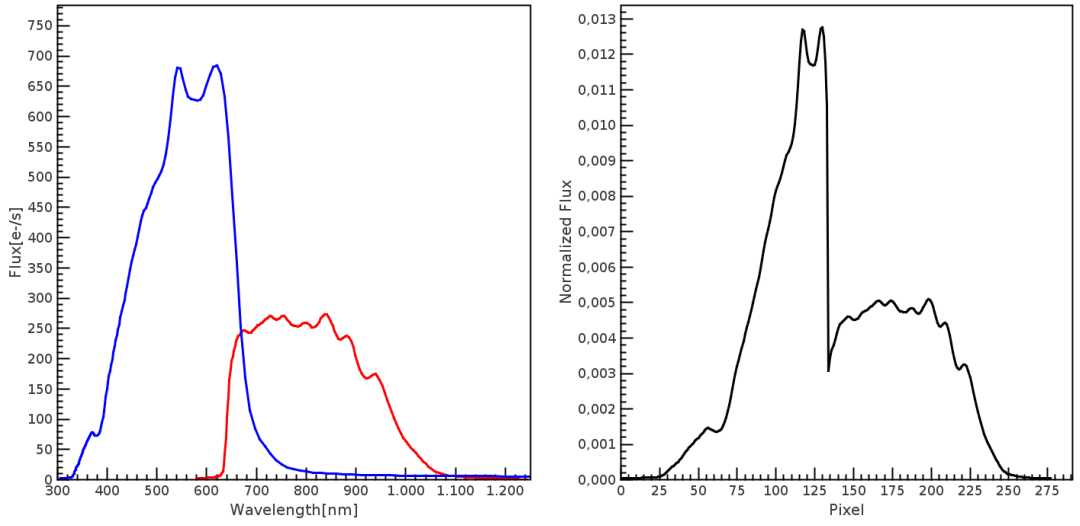
FIGURA 3.1: Muestra de 50 espectros BP/RP para cada librería. Los espectros están escalados a magnitud $G = 15$

.

3.4.1 Preprocesado de datos BP/RP

La entrada de OA está compuesta principalmente por datos provenientes de CU5, en forma de espectrofotometría BP/RP. Los espectros BP y RP se proporcionan por separado, de forma que es necesario construir el espectro combinado de algún modo. Además, los espectros están calibrados de forma interna, de modo que los mismos sufren de diversos efectos no físicos, provenientes del instrumento. Uno de ellos es la función de transmisión, que le proporciona a los espectros su aspecto característico en forma de campana. Más información sobre la calibración interna de BP/RP puede encontrarse en [61]. Se prevé una calibración externa de los espectros, aunque la misma está en desarrollo y necesitará datos observacionales de Gaia para ser completada.

El preprocesado de espectros BP/RP en el algoritmo de OA consiste en unir sin perder información los espectros BP y RP. El punto de corte donde terminan los píxeles pertenecientes al espectro BP y en el que comienzan los del espectro RP se ha determinado como el píxel en el cual la sensibilidad del detector BP cae por debajo de la del detector RP, el cual se corresponde con aproximadamente 644 nanómetros. Además, los espectros se normalizan para tener el mismo flujo integrado, evitando efectos de escala indeseados en el cálculo de distancia durante el proceso de entrenamiento. En caso contrario, los algoritmos clasificarían los espectros concentrándose en la magnitud aparente de los objetos, la cual es una propiedad no intrínseca de los astros, ya que depende de la distancia a la que éstos se sitúan del observador y de la absorción del medio interestelar. La figura 3.2 muestra un espectro BP/RP antes y después de ser preprocesado.



(a) Espectro BP/RP.

(b) Espectro preprocesado.

FIGURA 3.2: Gráfico de un espectro BP/RP, correspondiente a un quásar con magnitud $G = 15$ (izquierda) y el resultado del preprocesado realizado por OA (derecha).

Al contrario que algunos algoritmos de clasificación supervisada como los MLPs, los algoritmos de clasificación no supervisada, fundamentados en una función de distancia, no incorporan ningún proceso de selección o pesado de variables de entrada, con lo que la influencia de cada variable en el algoritmo queda determinada por la varianza de la misma. Esto puede suponer un problema cuando las distintas variables se han obtenido mediante medidas en diferente escala. Afortunadamente, en el caso de los espectros BP/RP todas las variables, que se corresponden con los píxeles del espectro BP/RP, están medidas con las mismas unidades de flujo, en este caso electrones por segundo ($e-/s$). Sin embargo, la función de transmisión del instrumento, mostrada en la figura 1.7, actúa como factor multiplicativo de la varianza de cada píxel. Una posible solución a este problema es la estandarización de las variables para que todas ellas tengan media 0 y varianza 1. En la sección 3.4.6 se discuten las diferencias entre utilizar espectros crudos y usar espectros estandarizados, así como el efecto de diversos procesos de selección y extracción de características en los resultados proporcionados por los algoritmos de agrupamiento.

3.4.2 Medidas de evaluación de algoritmos de agrupamiento

Antes de comenzar la evaluación, es necesario establecer un conjunto de medidas de calidad de las particiones computadas. En este caso utilizaremos medidas de evaluación externas, basadas en el conocimiento de la clase a la que pertenece cada objeto del conjunto de entrada. En este sentido, las "matrices de confusión" ofrecen una información muy completa para evaluar la precisión de un clasificador dado, cuando se aplica a un conjunto de datos. En el caso de clasificación no supervisada, el cómputo de esta matriz requiere el etiquetado previo de los grupos obtenidos. Para ello, se establece la clase de cada uno de los grupos como la más frecuente entre los objetos pertenecientes a dicho grupo. A continuación, se determina, para cada clase, qué porcentaje de sus objetos caen en grupos cuya clase coincida con cada una de las clases existentes. Dicho de otra forma, primero se crea un clasificador con los grupos computados y luego éste se aplica al mismo conjunto de entrada, obteniendo la matriz de confusión de la misma forma que en el caso de un clasificador supervisado. De la matriz de confusión pueden extraerse dos medidas de calidad del agrupamiento, llamadas *precision* y *accuracy*, que se definen como:

$$accuracy = \frac{1}{|X|} \sum_{x \in X} I[c(x) = p(x)] \tag{3.1}$$

$$precision = \frac{1}{|T|} \sum_{t \in T} \frac{\sum_{x \in X} I[p(x) = c(x) = t]}{\sum_{x \in X} I[c(x) = t]} \tag{3.2}$$

, donde X es el conjunto de objetos a clasificar, T el conjunto de clases existentes, $c(x)$ es la clase real del objeto x, $p(x)$ es la clase estimada por el clasificador e I es la función indicador.

La principal diferencia entre *precision* y *accuracy* es que *accuracy* tiene en cuenta el balanceo entre las clases en el conjunto de entrada, de forma que premia clasificadores que acierten más en clases más frecuentes, mientras que *precision* pondera todas las clases por igual, independientemente de qué porcentaje del total de objetos representen. Por lo tanto, *precision* proporciona una medida más robusta en el caso de conjuntos de datos desbalanceados, cuando las clases menos frecuentes son de especial interés. Este es el caso de Gaia, en el que la mayoría de objetos son estrellas, pero estamos interesados en aislar objetos raros cono estrellas binarias y quásares. Estas medidas se utilizarán para comparar los resultados obtenidos con distintos algoritmos de agrupamiento en las siguientes secciones.

3.4.3 Algoritmo k-means

El algoritmo k-means fue propuesto por Lloyd en 1957 [62]. Se trata de un algoritmo extremadamente sencillo, pero que sin embargo consigue realizar particiones coherentes, con una complejidad computacional considerablemente baja. Hoy en día existen algoritmos de agrupamiento más evolucionados que k-means, pero éste sigue siendo ampliamente utilizado en problemas que requieren un gran cantidad de cómputo, debido al gran número de observaciones o a la alta dimensionalidad de las mismas. El algoritmo se basa en minimizar una función objetivo, que se define como:

$$J = \arg\min_{S} \sum_{j=1}^{k} \sum_{x_i \in S_j} ||x_i - \mu_j||^2 \tag{3.3}$$

, donde x_i representa una observación perteneciente a un grupo dado y μ_j es la media de las observaciones en dicho grupo. Por lo tanto, el algoritmo trata de encontrar las k medias de grupo (también llamados prototipos) que minimizan la distancia euclídea cuadrática a los objetos que pertenecen a dicho grupo, donde la pertenencia de objetos al grupo se determina por mínima distancia. Dicho problema es NP-Hard, por lo que el proceso de minimización se realiza mediante un algoritmo iterativo, tal como se describe a continuación:

1. Inicializar k medias de forma aleatoria.

2. Asignar cada observación x_i al grupo más cercano según:

$$\arg\min_j ||x_i - \mu_j||^2 \tag{3.4}$$

3. Recalcular la media de cada grupo en función de las observaciones asignadas al mismo mediante:

$$\frac{1}{|S_j|} \sum_{x_i \in S_j} x_i. \tag{3.5}$$

4. Si convergencia (las medias no se mueven) entonces fin. En caso contrario, volver a 2.

Dicho algoritmo asegura la convergencia después de un número finito de iteraciones. Sin embargo, dicha convergencia puede ser debida a un mínimo local, lo cual constituye uno de los principales problemas que presenta el algoritmo. Para paliar dicho problema, normalmente se efectúan varias inicializaciones del algoritmo y se escoge la que ha obtenido una mejor minimización. También se han presentado inicializaciones robustas que aseguran un mínimo local más aceptable, véase el algoritmo k-means++ [63]. Además, el algoritmo k-means es sensible a la existencia de valores atípicos en el conjunto de datos. Por último, el número de grupos (k) no se conoce a priori en muchas aplicaciones, por lo que debe realizarse una estimación de k mediante la experimentación, lo cual puede resultar un proceso complejo en muchos casos.

A continuación, se muestran los resultados obtenidos cuando se particiona el conjunto de datos descrito en la sección 3.4, preprocesado tal y como se ha descrito en la sección 3.4.1, con el algoritmo k-means. La medida de evaluación utilizada para calificar las particiones es *precision* (ver ecuación 3.2), variándose el número de grupos k. Para cada valor de k, se ejecutan 5 instancias del algoritmo con inicialización aleatoria. Como puede observarse, los resultados se estabilizan para $k > 100$, con lo que el algoritmo es incapaz de sacar ventaja de un número mayor de grupos a partir de tal punto. El valor máximo de *precision* es 65%, un valor que podemos definir como poco satisfactorio. A continuación, se exploran otros algoritmos de agrupamiento con el objetivo de mejorar dicho valor.

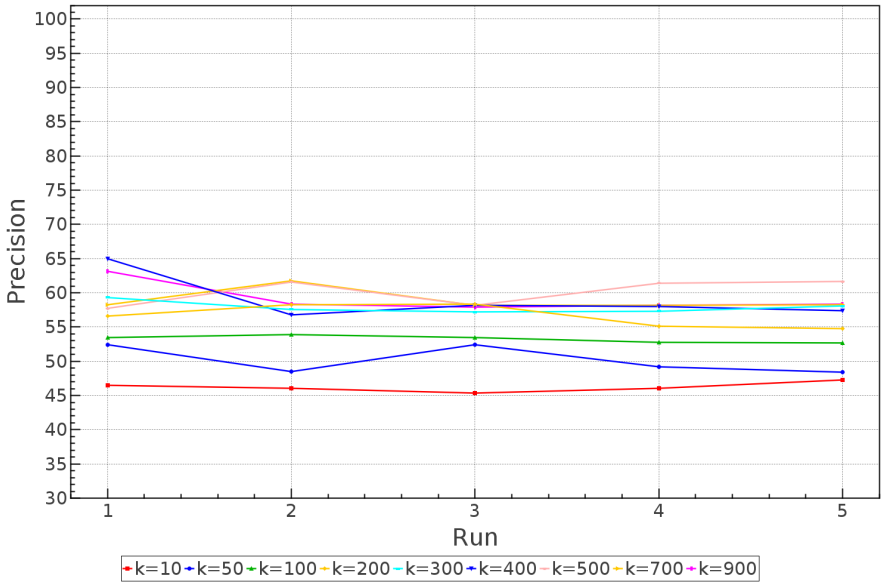
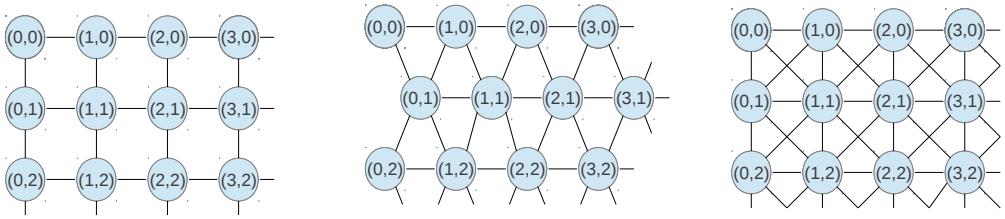


FIGURA 3.3: Evaluación del agrupamiento obtenido mediante el algoritmo k-means, con diversos valores de k, al presentar espectros BP/RP sin ruido.

3.4.4 Mapas auto-organizativos

Los mapas auto-organizativos (SOM) provienen de la rama de las redes neuronales de tipo competitivo. Fueron propuestos por Kohonen en 1988 [64]. Desde entonces, se mantienen como la ANN no supervisada por excelencia. De hecho, hasta la fecha se han publicado más de 5.000 artículos relacionados con los SOMs. Los mapas auto-organizativos se obtienen mediante la proyección de un espacio continuo multidimensional de entrada en un espacio discreto bidimensional de salida. Esto es, se proyecta el conjunto de datos de entrada en un conjunto de neuronas, las cuales están organizadas topológicamente en una malla. Generalmente, la malla es 2D, por simplicidad de cómputo y posterior visualización, aunque pueden especificarse mallas en 3 o más dimensiones. La topología de la malla debe especificarse previamente al ajuste de la misma con los datos. Normalmente se utilizan mallas con topología rectangular, hexagonal u octogonal, las cuales se muestran en la figura 3.4. Cada neurona tiene un conjunto de pesos que la relacionan con el espacio de entrada. Dicho conjunto de pesos puede pensarse también como un prototipo de los objetos del espacio de entrada que activan la neurona.

La proyección de los datos en la malla se realiza mediante un proceso de aprendizaje no supervisado, que se basa en la competición de las neuronas por ajustarse a los datos lo máximo posible, de forma que neuronas próximas en la malla de salida se mantienen próximas en el espacio de entrada, i.e. sus prototipos son similares. El entrenamiento de la red es un proceso iterativo, como se describe a continuación:



(a) Rectangular (b) Hexagonal (c) Octogonal

FIGURA 3.4: Diferentes tipos de topologías aplicadas frecuentemente a los mapas auto-organizativos.

1. Definir una topología para la malla, un factor de aprendizaje decreciente $\alpha(s)$ y una función de vecindad $h(s, win, k)$, la cual decrece con las iteraciones s y con la distancia en la malla entre la neurona ganadora win y la neurona k.

2. Inicializar las j neuronas de la malla con pesos w_j de forma aleatoria.

3. Obtener el factor de aprendizaje $\alpha(s)$ y la función de vecindad entre neuronas $h(s, win, k)$ para la iteración actual s.

4. Seleccionar una observación aleatoria x_i del espacio de entrada.

5. Calcular la neurona ganadora win mediante la mínima distancia entre la observación x_i y los pesos de la neurona w_j:

$$win = \arg\min_j ||x_i - w_j||^2 \tag{3.6}$$

6. Actualizar los pesos w_k de la vecindad de la neurona ganadora, compuesta por k neuronas, donde la función $h(s, win, k)$ se centra en la neurona ganadora:

$$w_k = w_k + \alpha(s)h(s, win, k)(x_i - w_k) \tag{3.7}$$

7. Si convergencia (los pesos no se mueven), entonces fin. En otro caso, volver a 3.

Como se ha mencionado anteriormente, los SOMs surgieron del campo de las redes neuronales competitivas. De ahí que, en su inicio, el algoritmo de entrenamiento propuesto se basase en una perspectiva *online*, y que hiciese uso de un factor de aprendizaje. Posteriormente, fue descrita la versión *batch* del algoritmo de entrenamiento, la cual prescinde del factor de aprendizaje. Dicho algoritmo se describe como sigue:

1. Definir una topología para la malla y una función de vecindad $h(s, win, k)$, la cual decrece con las iteraciones s y con la distancia en la malla entre la neurona ganadora win y la neurona k.

2. Inicializar las j neuronas de la malla con pesos w_j de forma aleatoria.

3. Obtener la función de vecindad entre neuronas $h(s, win, k)$ para la iteración actual s.

4. Asignar cada observación x_i a la neurona más cercana según la ecuación 3.6.

5. Actualizar los pesos de cada neurona w_j, en función de las observaciones asignadas a la misma y de las observaciones asignadas a sus vecinas, ponderadas por la función de vecindad:

$$w_j = \frac{\sum_{i=1}^{n} h(s, win, j) x_i}{\sum_{i=1}^{n} h(s, win, j)} \tag{3.8}$$

6. Si convergencia (los pesos no se mueven), entonces fin. En otro caso, volver a 3.

La versión *batch* del algoritmo de aprendizaje proporciona varias ventajas con respecto a la versión *online* cuando las SOMs se aplican como métodos de análisis no supervisado, como en el caso de aplicaciones de agrupamiento o de visualización multidimensional. En primer lugar, no es necesario definir un factor de aprendizaje. Además, el algoritmo no depende del orden en el que se le presenten las observaciones, ya que todas ellas se tienen en cuenta en la proyección final. Por último, es más sencillo especificar un criterio de convergencia para el algoritmo *batch*, como se discute más adelante. La versión *batch* resulta ser muy similar al algoritmo *k-means*, tomando las neuronas como grupos, y sus pesos como prototipo de cada grupo. De hecho, su funcionamiento es idéntico si se reduce la vecindad a únicamente la neurona ganadora. Así, podemos distinguir dos fases en el entrenamiento de un SOM: una primera fase de ordenación donde los pesos de las neuronas se colocan de forma que siguen la topología del mapa y una segunda fase de ajuste, donde se minimiza la distancia entre las observaciones y los pesos de las neuronas a las que activan.

Tanto la topología del mapa auto-organizado, como su función de vecindad, deben ser determinados para cada aplicación concreta. A continuación, estableceremos dichos parámetros con el objetivo de realizar la partición del conjunto de datos BP/RP, una vez éste ha sido preprocesado (ver sección 3.4.1). En primer lugar, se establece una malla en dos dimensiones con M filas y N columnas, en la cual las neuronas se conectan mediante una topología octogonal. La función de vecindad varía, en función de la distancia d entre las neuronas j, k y de la iteración s, siguiendo la ecuación:

$$h(s, j, k) = exp(\frac{-d^2}{2\sigma(s)^2}) \tag{3.9}$$

, donde la distancia d se define como el número de conexiones que es necesario atravesar para alcanzar la neurona k desde la neurona j, y $\sigma(s)$ es el radio de la vecindad en la iteración actual. Dicho radio decrece exponencialmente con las iteraciones, de acuerdo a:

$$\sigma(s) = \sigma(1)exp(\frac{-s}{T}) \tag{3.10}$$

, donde $\sigma(1)$ es el radio inicial de la función de vecindad y T es un factor que modula la velocidad de decrecimiento del radio con las iteraciones. El radio inicial $\sigma(1)$ debe cubrir una gran parte del mapa y puede establecerse en función del tamaño de la malla, por ejemplo como el número de filas o de columnas de la misma. Por otro lado, el factor T debe establecer un decrecimiento del radio lo suficiente suave como para permitir que el mapa pueda ordenarse adecuadamente, pero no demasiado suave, ya que esto ralentizaría la convergencia del algoritmo. La figura 3.5 muestra la evolución de $\sigma(s)$ en función de varios valores de T, estableciendo $\sigma(1) = 30$. Se ha seleccionado el valor $T = 50$, ya que proporciona un buen compromiso entre un correcto ordenamiento y una convergencia rápida, lo cual se ha determinado mediante varios conjuntos de datos bien conocidos en el ámbito del aprendizaje máquina, como por ejemplo el conjunto Iris, véase [65]. En este caso, se permite que el radio de vecindad decrezca hasta cero (sólo la neurona ganadora), ya que de este modo se asegura que el algoritmo converge a un mínimo de la función objetivo descrita en la ecuación 3.3. Esto es deseable, ya que de esta forma no es necesario predefinir el número de iteraciones del algoritmo de aprendizaje. Además, de esta forma el SOM obtiene un ajuste cercano al óptimo con los datos. Sin embargo, en ocasiones se detiene el decrecimiento del radio en un valor más alto, lo cual proporciona más robustez ante objetos atípicos en el conjunto de datos y una visualización más suave del mapa. Más adelante, se profundiza más en las implicaciones de la cota inferior del radio de vecindad.

A modo de ilustración del método de aprendizaje descrito para los mapas auto-organizativos, se ha generado un conjunto de datos ficticios, el cual se compone de una serie de puntos en un plano bidimensional, los cuales forman 9 grupos compactos. Utilizando dicho conjunto de datos se entrena una red SOM, de tamaño 10*10, y con la configuración descrita anteriormente. La figura 3.6 muestra las posiciones de los pesos de las neuronas de la red en cada una de las iteraciones. En primer lugar, se inicializan los pesos con valores aleatorios en torno a la media del conjunto de datos. Entonces, comienza la auto-organización de las neuronas. Una vez las neuronas se han ordenado,

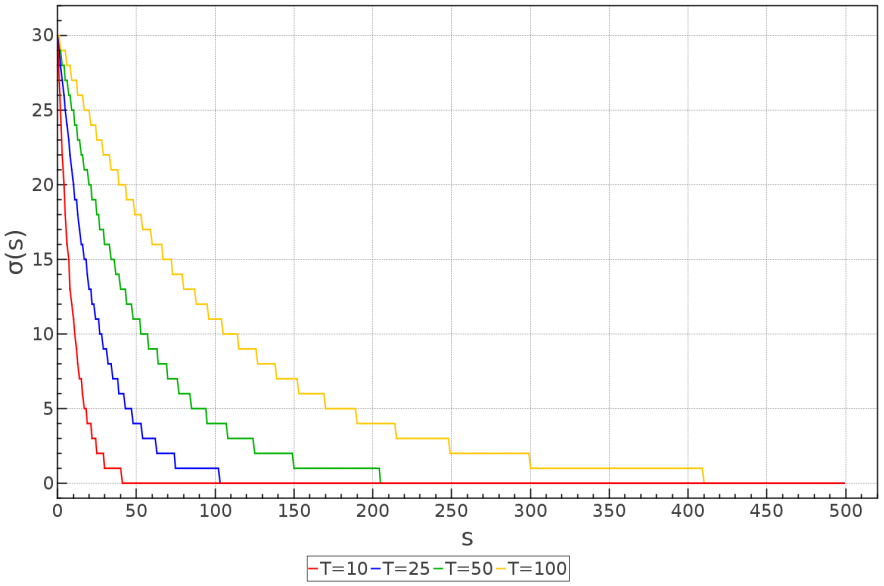
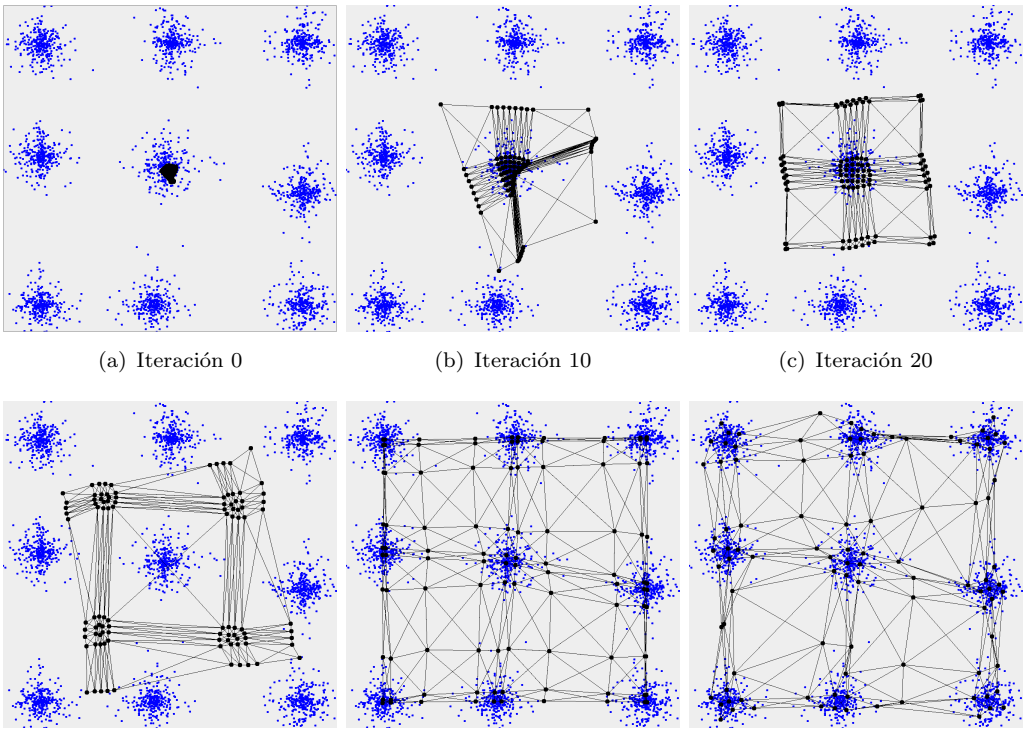


FIGURA 3.5: Evolución de $sigma(s)$ para varios valores de T, con $sigma(1) = 30$.

comienza el proceso de ajuste con los datos, hasta que finalmente el entrenamiento converge en la iteración 150.



(a) Iteración 0 (b) Iteración 10 (c) Iteración 20

(d) Iteración 40 (e) Iteración 80 (f) Iteración 150

FIGURA 3.6: Ilustración del proceso de aprendizaje de una red SOM.

La figura 3.7 muestra los resultados obtenidos mediante la evaluación de las particiones obtenidas por el algoritmo SOM. Al igual que ocurría con el parámetro k en k-means, el tamaño de la malla puede ser difícil de estimar en ciertas aplicaciones, por lo que se

muestran los resultados obtenidos para diversos tamaños del mapa. Se puede observar, que con un mapa 30*30 (900 grupos), el valor de *precision* obtenido es aproximadamente del 81%, lo cual mejora significativamente el obtenido con el algoritmo k-means. Además, la variabilidad entre inicializaciones es significativamente menor, gracias al efecto que la vecindad ejerce en el mapa-autoorganizativo. La tabla 3.2 muestra la matriz de confusión obtenida para el mapa 30*30. Se observa que las clases Be y WR se confunden parcialmente con las clases OB y WD, debido a la gran similitud entre sus espectros BP/RP. Además, las estrellas de clase C se confunden completamente con estrellas AFGKM. Esto se debe a que, con la baja resolución del espectro BP/RP, resulta inviable el distinguir las bandas de absorción propias de las estrellas de carbono. La red SOM es capaz de distinguir con alta pureza el resto de las clases. Por lo tanto, la red SOM, de tamaño 30*30, se toma como algoritmo de agrupamiento de referencia para los subsiguientes experimentos. Una red de mayor tamaño conseguiría obtener una pureza ligeramente mayor, pero a costa de un aumento en la complejidad computacional del entrenamiento y en la dificultad del análisis posterior.

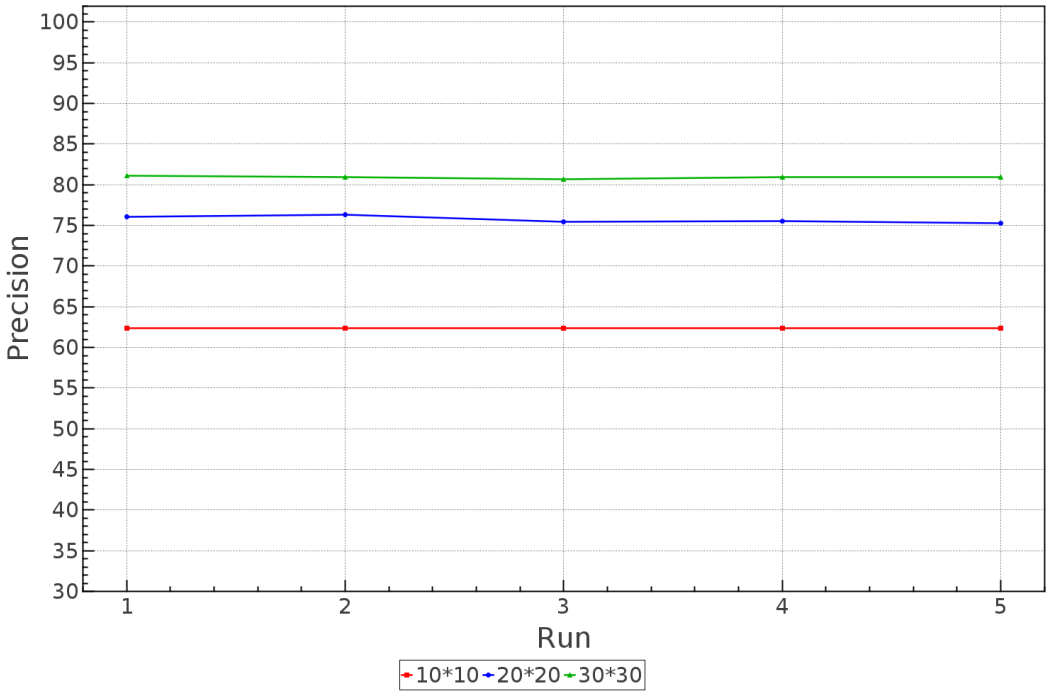


FIGURA 3.7: Evaluación del agrupamiento obtenido mediante una red SOM, de diversos tamaños, al presentar espectros BP/RP sin ruido.

3.4.4.1 Visualización de mapas auto-organizativos

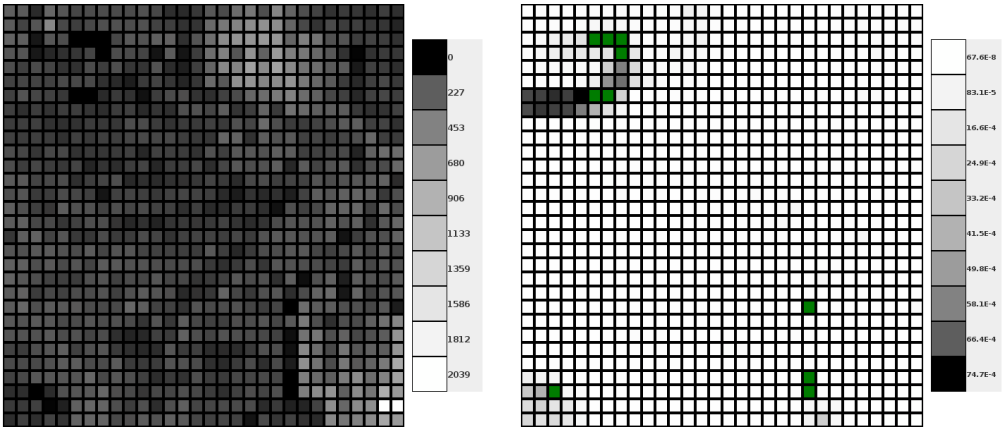
En la sección anterior, se ha mostrado que las redes SOM son capaces de proporcionar agrupamiento de calidad. Adicionalmente, los mapas auto-organizativos, una vez han sido ajustados, permiten la visualización de un conjunto de datos multidimensional

TABLA 3.2: Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia sin ruido.

	AFGKM	Be	C_STARS	GALAXY	OB	PN	QSO	UCD	WD	WR	UNDEFINED
AFGKM	96,78	0	0	0,22	0	0	0	2,94	0,06	0	0
Be	3,07	54,79	0	0	30,27	0	0	0	11,88	0	0
C_STARS	92,29	0	0	7,71	0	0	0	0	0	0	0
GALAXY	0,54	0	0	98,31	0	0	1,15	0	0	0	0
OB	0	0	0	0	99,73	0	0	0	0,27	0	0
PN	0	0	0	0	0	100	0	0	0	0	0
QSO	0,41	0	0	0,84	0	0	98,73	0,01	0	0	0
UCD	0,18	0	0	0	0	0	0	99,82	0	0	0
WD	1,65	0	0	0,01	0	0	0	0	98,33	0	0
WR	0	11,63	0	0	0	0	0	0	24,03	64,34	0

mediante las relaciones entre las neuronas y las observaciones así como entre neuronas, a través de la malla predeterminada. Esta propiedad de los mapas se debe a su capacidad para preservar la topología del espacio de entrada. Existen múltiples visualizaciones posibles de los mapas auto-organizativos, como son el diagrama de impactos, la llamada matriz unificada de distancias (matriz U) o los planos de componentes. A continuación, se describen dichas visualizaciones, y se muestran para el caso de la red SOM de tamaño 30*30, computada en la sección anterior.

- El diagrama de impactos muestra el número de observaciones que caen en cada neurona, representado por un color en una escala de grises, de forma que un color oscuro indica que la neurona contiene pocas observaciones, mientras que un color claro indica que la misma tiene un alto número de observaciones. Esta visualización resulta de utilidad para visualizar la densidad de datos en cada región de la red SOM y, por ende, del espacio de entrada. La figura 3.8(a) muestra el diagrama de impactos para la red computada con simulaciones de Gaia.



(a) Diagrama de impactos. La leyenda indica el número de objetos por nivel de gris.

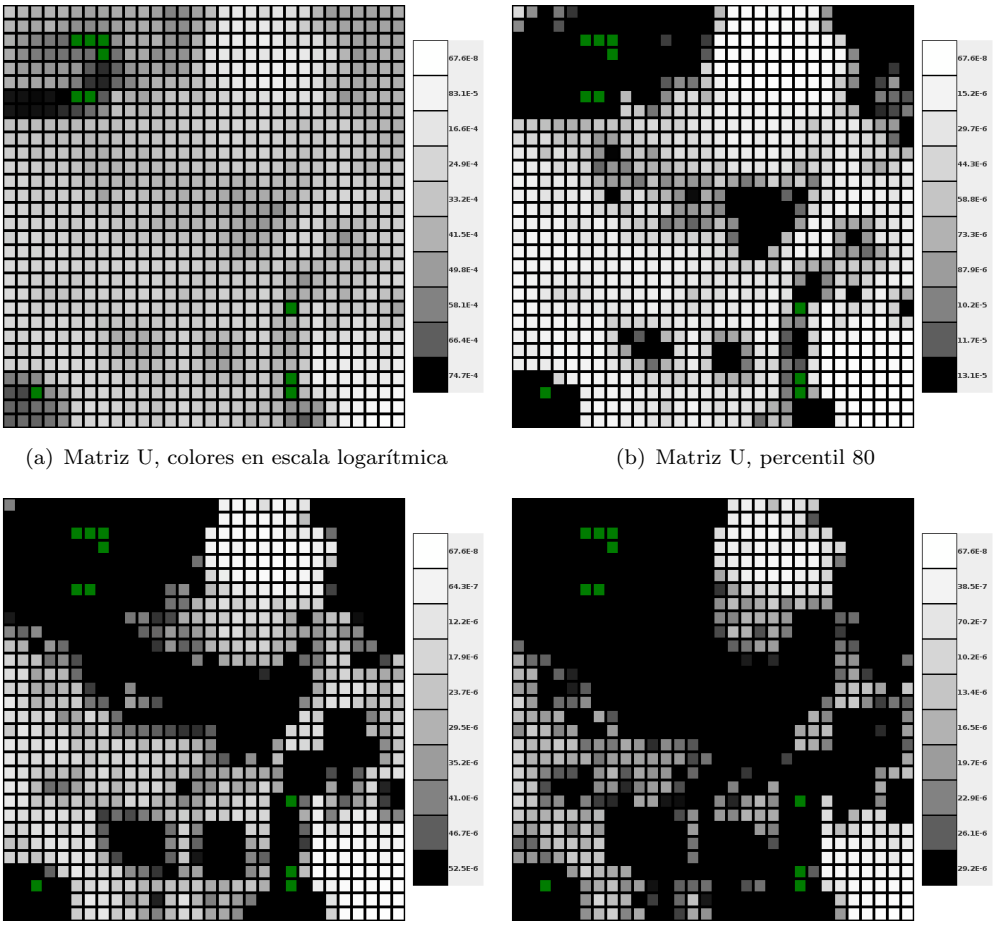
(b) Matriz U. La leyenda indica el valor de distancia por nivel de gris. Las neuronas vacías reciben el color verde.

FIGURA 3.8: Visualizaciones estándar de un mapa auto-organizativo.

- La matriz U es una visualización en la que se representan las distancias relativas entre cada neurona y sus vecinas inmediatas. A cada neurona se le asigna un color en una escala de grises, de forma que si la neurona recibe un color oscuro, ésta se sitúa a gran distancia de sus vecinas, mientras que si recibe un color claro, la neurona se sitúa a corta distancia. De esta forma, la matriz U representa cómo se distribuyen las neuronas del mapa, que a su vez se sitúan cerca de las observaciones que las activan. Las zonas claras representan regiones de alta densidad de datos, lo cual suele relacionarse con clases de observaciones típicas, mientras que las zonas oscuras pueden asociarse con zonas de transición entre clases o bien con valores de observaciones atípicas. La figura 3.8(b), muestra la matriz U obtenida a través del conjunto de simulaciones. Desafortunadamente, cuando existen neuronas, cuyo valor en la matriz U es atípico (la distancia a sus vecinas es muy grande), los colores en escala de grises no permiten visualizar la distribución de las neuronas típicas, como es el caso que nos ocupa. Para mejorar la visualización de la matriz U, es posible cambiar los tonos de gris a una escala logarítmica, o bien establecer un percentil en las distancias, de forma que los objetos que salgan del percentil reciben un color negro, mientras que los que quedan dentro del percentil reciben un color en escala de grises relativo tan sólo a las distancias de las neuronas que entran en dicho percentil. Estas visualizaciones mejoradas de la matriz U se muestran en la figura 3.9.

- Los planos de componentes permiten visualizar la distribución en el mapa de cada una de las propiedades medidas para el conjunto de observaciones de entrada. Esto permite encontrar correlaciones entre las variables de entrada, así como a modo de validación del ordenamiento realizado por la red SOM. Sin embargo, los planos de componentes no resultan demasiado útiles cuando las entradas son de alta dimensionalidad, como es el caso de los espectros astronómicos. Aun así, pueden confeccionarse planos de componentes basados en combinaciones de entradas. Por ejemplo, la figura 3.10 muestra la distribución del color $G_{rp} - G_{bp}$, el cual se obtiene restando las magnitudes correspondientes a los flujos integrados, respectivamente, del espectro RP y BP. Dicha visualización nos permite evaluar el ordenamiento de la red en función del color de los astros, el cual está directamente ligado con la temperatura de los mismos.

Las visualizaciones mencionadas hasta ahora sirven como mapas topológicos del conjunto de datos de entrada, permitiendo visualizar su distribución general. Éstas pueden ser complementadas con visualizaciones del contenido de cada una de las neuronas, ya que de este modo se obtiene una información más detallada del conjunto de datos que se está analizando. En este sentido, los pesos de cada neurona representan el prototipo



(a) Matriz U, colores en escala logarítmica

(b) Matriz U, percentil 80

(c) Matriz U, percentil 60

(d) Matriz U, precentil 40

FIGURA 3.9: Visualización mejorada de la matriz U.

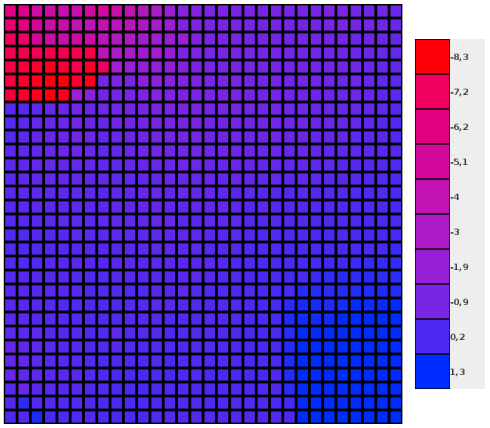


FIGURA 3.10: Distribución del color $G_{rp} - Gbp$ en el mapa auto-organizativo.

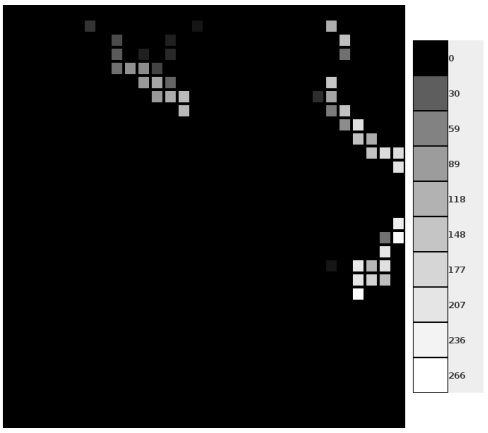
de los espectros alojados en la misma, por lo que su visualización sirve a modo de resumen del contenido del grupo. Además, puede visualizarse la desviación típica de los objetos que caen en la neurona a modo de medida de incertidumbre sobre el prototipo. Generalmente, no es necesario visualizar los prototipos de todas las neuronas, ya que la matriz U sirve a modo de guía. Las neuronas situadas en zonas más claras contienen prototipos similares, por lo que dichas regiones pueden ser exploradas con menor exhaustividad. Por el contrario, en las zonas oscuras de la matriz U, los prototipos son más dispares, con lo que la exploración debe realizarse de forma más completa.

Además de las visualizaciones típicas de la red SOM, de tipo no supervisado, el hecho de disponer de una clasificación supervisada para los objetos del conjunto de entrada, nos permite el visualizar cómo se distribuyen dichas clases en el mapa. La figura 3.10 muestra por separado los impactos de cada una de las clases de objetos en el conjunto de simulaciones utilizado en la evaluación. Puede observarse que, en general, cada una de las clases se agrupa en una región determinada del mapa. Cabe destacar que la clase QSO ocupa una región muy amplia, al igual que la clase GALAXY, aunque de un modo menos significativo. Esto se atribuye a la gran variabilidad que induce el corrimiento al rojo en los espectros. Además, la clase AFGKM se superpone completamente con la clase C_STARS, mientras que las clases correspondientes a estrellas de color azul: WD, Be y WR muestran una confusión significativa. Con la distribución de objetos por clase pueden obtenerse visualizaciones adicionales, basadas en el cálculo de estadísticas sobre las clases que pueblan cada neurona. Por ejemplo, pueden obtenerse las clases prominentes de cada neurona, lo cual se muestra en la figura 3.11(a), o bien la pureza (el porcentaje de objetos de la misma clase) de cada neurona, mostrada en la figura 3.11(b).

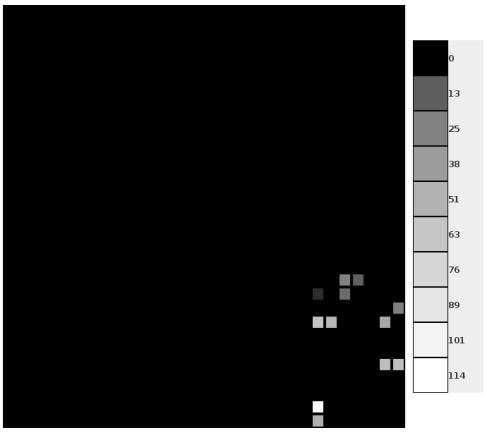
En esta sección se ha mostrado la capacidad que tienen las redes SOM, tanto para agrupar espectros BP/RP, como para proyectar los espectros en un mapa bidimensional donde se preserva la topología, de forma que distintas clases espectrales ocupan distintas regiones del mapa. Sin embargo, el ruido que afectará a los espectros BP/RP en operación real, especialmente para objetos con magnitudes altas, puede perjudicar las capacidades de las redes. En la siguiente sección, se estudia el impacto del ruido en la clasificación no supervisada.

3.4.5 Sensibilidad al ruido de las técnicas de agrupamiento

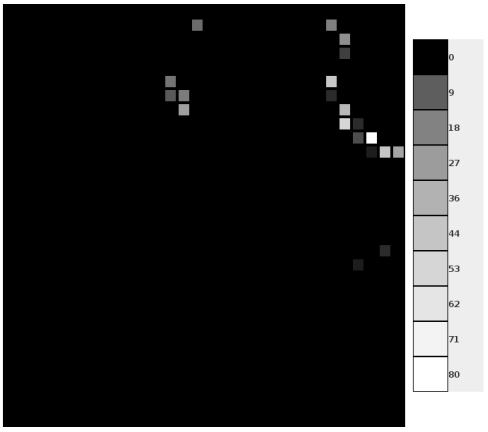
El conjunto de espectros BP/RP utilizado para la evaluación de técnicas de agrupamiento está basado en espectros con un alto SNR, provenientes tanto de simulaciones con modelos como de bases de datos astronómicas. Sin embargo, Gaia



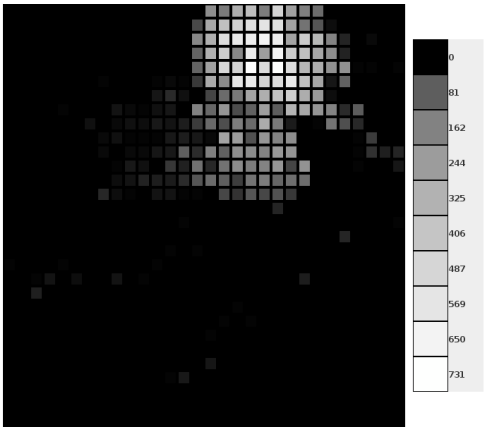
(a) AFGKM



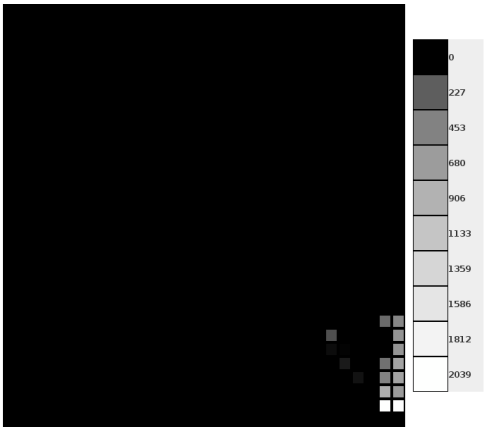
(b) Be



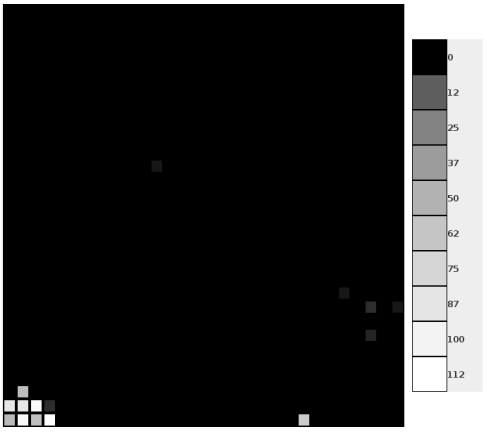
(c) C_STARS



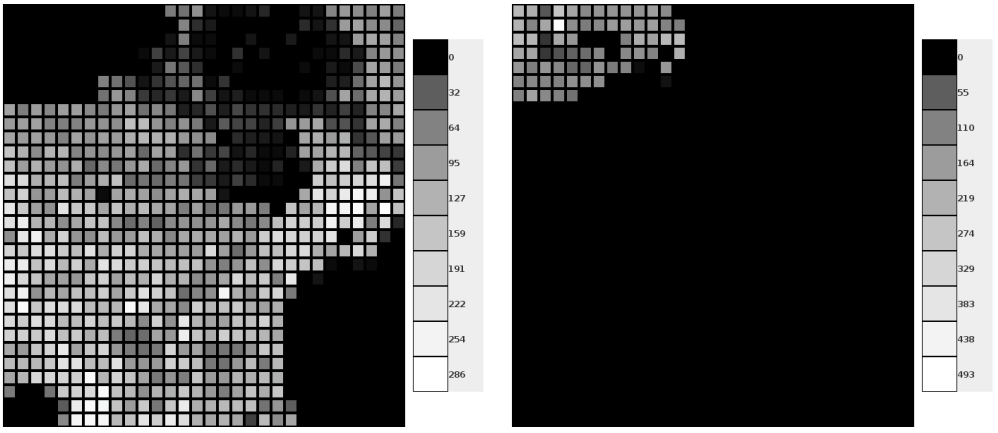
(d) GALAXY



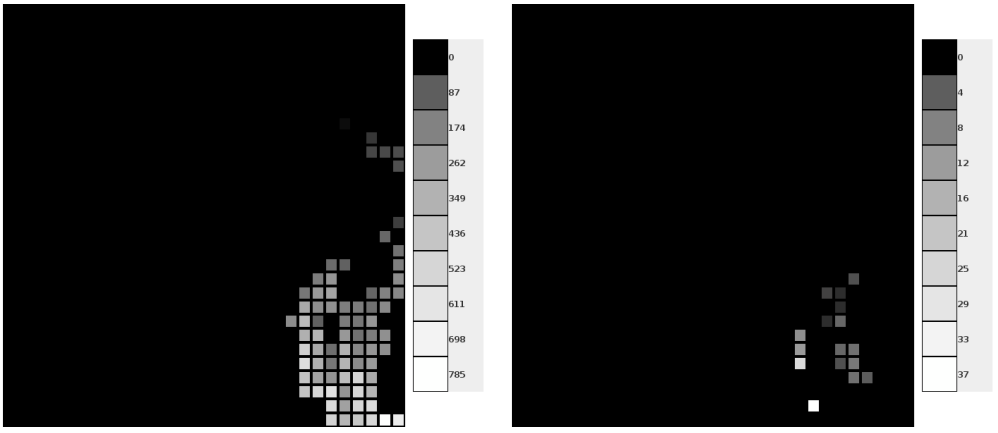
(e) OB



(f) PN

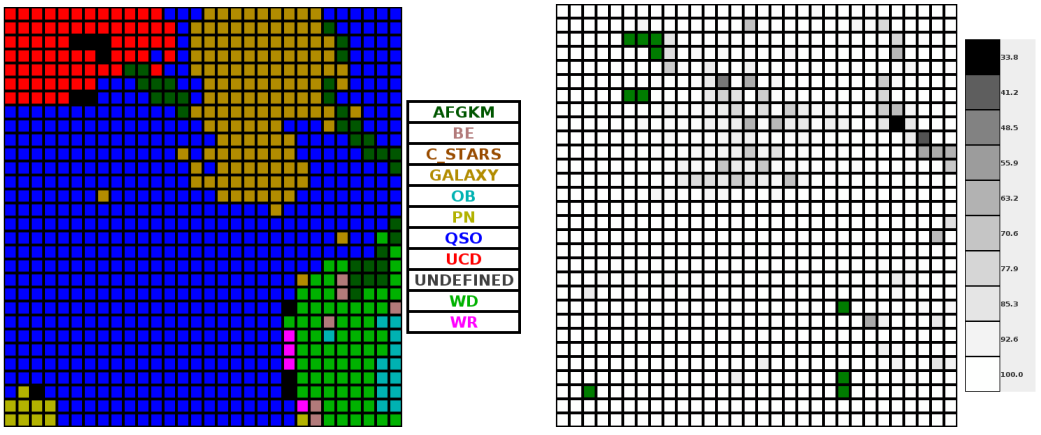


(g) QSO (h) UCD



(i) WD (j) WR

FIGURA 3.10: Número de impactos en la red SOM para cada una de las clases supervisadas.



(a) Clase predominante (b) Pureza

FIGURA 3.11: Diagramas obtenidos mediante estadísticas sobre las clases que pueblan cada neurona en la red SOM.

observará fuentes débiles, cuyo espectro BP/RP tendrá un SNR considerablemente más bajo, como se muestra en la figura 3.12, donde se muestra la relación entre la magnitud G de las fuentes y el SNR en el espectro BP/RP. Por lo tanto, es necesario evaluar el rendimiento de las técnicas utilizadas cuando se le ha añadido ruido a los espectros de entrada. La figura 3.13 muestra los resultados obtenidos al entrenar una red SOM, de tamaño 30*30, con el conjunto de datos simulados con ruido añadido en función de la magnitud G. Se puede observar cómo el rendimiento de los mapas decrece con el aumento de magnitud, bajando hasta el 48% en el caso de $G = 20$. Mediante la matriz de confusión, mostrada en la tabla 3.3, se observa que, incluso con magnitudes muy débiles, la red es capaz de distinguir correctamente entre quásares, galaxias, nebulosas planetarias, enanas blancas y enanas ultra-frías. Sin embargo, las estrellas del conjunto AFGKM o las estrellas OB se confunden con otras clases de objetos, lo cual se atribuye, además de al efecto del ruido, a que las clases en el conjunto de simulaciones tienen un número desigual de objetos. Se ha tratado de mitigar el efecto del ruido mediante un cota superior en el radio de la función de vecindad de la SOM. Sin embargo, esto no ha mejorado la robustez del algoritmo. En las próximas secciones, buscaremos encontrar métodos que nos permitan mejorar el agrupamiento de objetos, especialmente en casos con un alto nivel de ruido, como el que esperamos en los objetos procesados por OA.

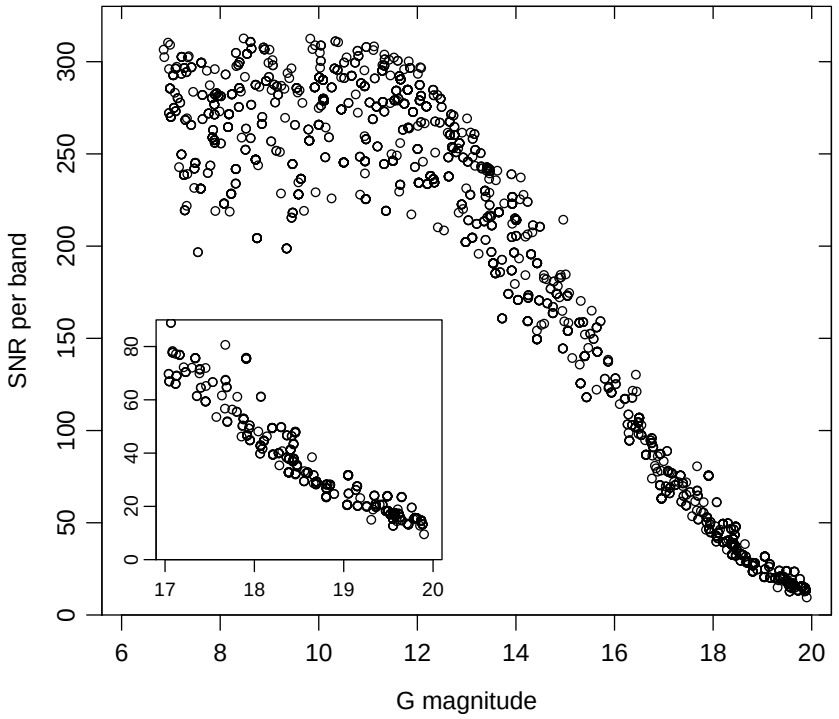


FIGURA 3.12: Relación entre la magnitud G de un objeto observado por Gaia y el SNR medio del espectro BP/RP.

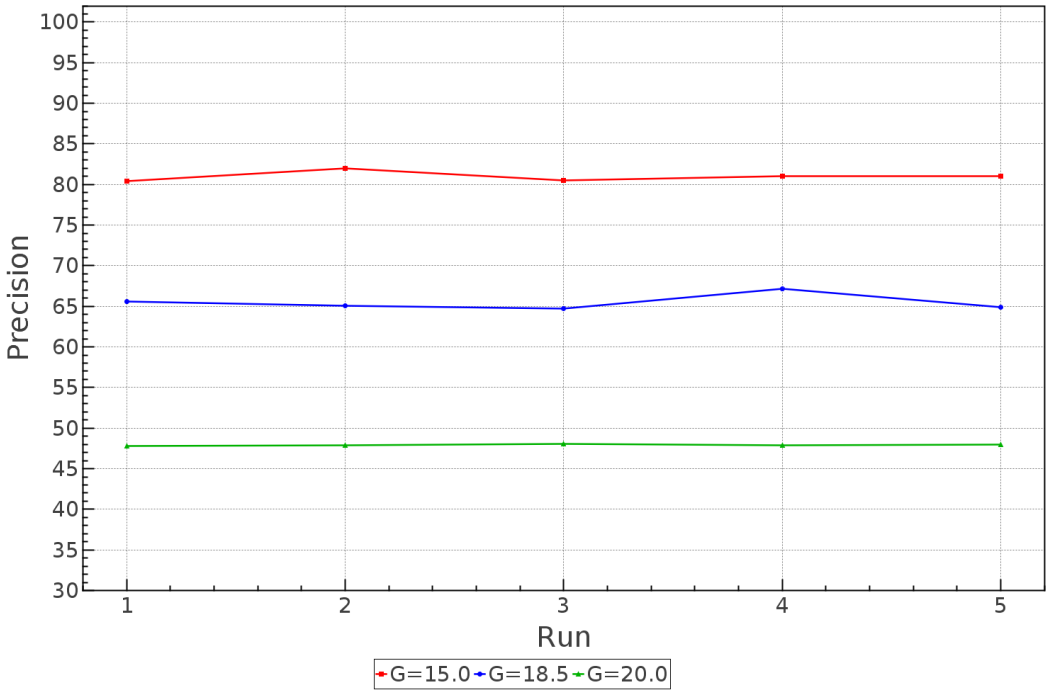


FIGURA 3.13: Evaluación del agrupamiento obtenido mediante una red SOM 30*30, al presentar espectros BP/RP con varios niveles de ruido.

TABLA 3.3: Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia con $G = 20$.

	AFGKM	Be	C_STARS	GALAXY	OB	PN	QSO	UCD	WD	WR	UNDEFINED
AFGKM	4,4	0	0	24,72	0,02	0	46,98	4,94	18,28	0	0,66
Be	0	0	0	0	3,07	0	2,68	0	94,25	0	0
C_STARS	0,93	0	0	48,36	0	0	50,23	0	0,23	0	0,23
GALAXY	0,07	0	0	89,5	0	0	9,93	0,01	0	0	0,5
OB	0	0	0	0	6,6	0	0,01	0	93,39	0	0
PN	0	0	0	0	0	99,33	0,27	0	0,4	0	0
QSO	0,07	0	0	7,99	0	0	91,08	0,08	0,45	0	0,33
UCD	0,88	0	0	0,02	0	0	0,01	99,09	0	0	0
WD	0	0	0	0,03	2,99	0	6,05	0	90,62	0	0,31
WR	0	0	0	0	2,33	0	3,88	0	93,8	0	0

3.4.6 Técnicas de extracción de características para el agrupamiento de espectros BP/RP

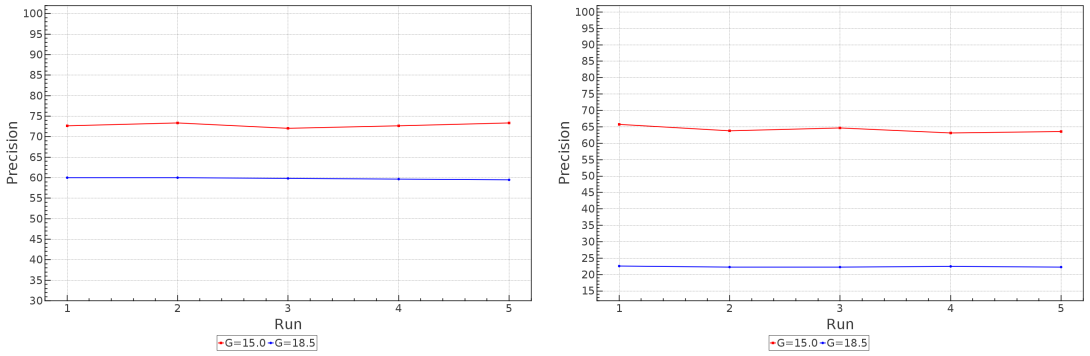
Las técnicas de clasificación no supervisada reflejan directamente la información embebida en el conjunto de datos de entrada, sin realizar ningún tipo de extracción de características implícito. Por lo tanto, es crucial la selección de un conjunto de características que representen lo mejor posible la información sobre los objetos observados, y que permitan extraer el máximo conocimiento sobre los mismos. Las características seleccionadas deben ser fácilmente interpretables, ya que de otra forma se complicaría el análisis posterior. Además, dichas características deben ser robustas en presencia de ruido en las observaciones.

En ciertas ocasiones, las observaciones están compuestas por propiedades medidas con diferentes unidades. Por ejemplo, la temperatura efectiva T_{eff} y la metalicidad $[Fe/H]$ son dos propiedades estelares con diferentes unidades de medida (kelvin y dex). Esto constituiría un problema para los métodos de clasificación no supervisada, ya que T_{eff} tiene un rango de valores mucho mayor que $[Fe/H]$, con lo que T_{eff} tomaría el protagonismo en la clasificación, mientras que $[Fe/H]$ sería prácticamente ignorado. Para solucionar este problema, existen técnicas de preprocesado, como la estandarización, que transforman los valores de cada variable, de forma que todas ellas tengan media 0 y varianza 1. De esta forma, todas las variables toman la misma importancia en la clasificación. Sin embargo, en el caso de espectros astronómicos, todos los valores del espectro están medidos con las mismas unidades y existe cierta dependencia entre los distintos valores. Además, en el caso de espectros BP/RP, los valores del espectro están afectados por la sensibilidad del detector, lo cual actúa como un factor multiplicativo que incrementa la variabilidad de los valores en longitudes de onda captadas con una mayor sensibilidad (los centrales), con respecto a aquellas captadas con una sensibilidad menor. Por lo tanto, es importante conocer el efecto que tienen, en las técnicas no supervisadas, las diferencias de variabilidad entre valores del espectro, diferenciando entre la variabilidad inherente de los espectros y la inducida por los instrumentos de medición. A continuación, se muestran los resultados obtenidos al entrenar una red SOM con el conjunto de simulaciones de Gaia, al cual se le han aplicado distintos tipos de preprocesado:

- Espectros BP/RP estandarizados: Para estandarizar los espectros BP/RP, primero se calcula la media y la desviación típica de cada uno de los píxeles en el conjunto de simulaciones. Entonces, a cada espectro se le resta la media del conjunto y se divide por la desviación típica. De esta forma, el conjunto de datos pasa a tener media 0 y varianza 1. La figura 3.14(a) muestra el resultado del agrupamiento realizado con esta aproximación. Se puede observar que los resultados están por debajo de los obtenidos con espectros originales. Esto indica que las diferencias de varianza entre los píxeles BP/RP no perjudica la clasificación, sino que todo lo contrario. Además, la estandarización rompe en cierta medida la dependencia entre píxeles.

- Espectros BP/RP, extrayendo la respuesta instrumental: Esta representación se obtiene al dividir cada espectro BP/RP en el conjunto de simulaciones por la respuesta instrumental teórica. Dicha respuesta se ha obtenido mediante la presentación al simulador GOG de un espectro completamente plano. Los resultados de agrupamiento con una red SOM se muestran en la figura 3.14(b). Se observa que con espectros con bajo ruido los resultados son similares a los

obtenidos con la representación original. Sin embargo, al aumentar el ruido, los resultados empeoran enormemente. La razón a ello es que, al eliminar la respuesta, se equilibra el impacto de todos los píxeles del espectro. Esto resulta perjudicial, ya que el SNR en los extremos del espectro es mucho menor que en las partes centrales, las cuales tomaban un mayor protagonismo en los espectros con respuesta añadida.



(a) Espectros BP/RP estandarizados

(b) Espectros BP/RP sin respuesta instrumental

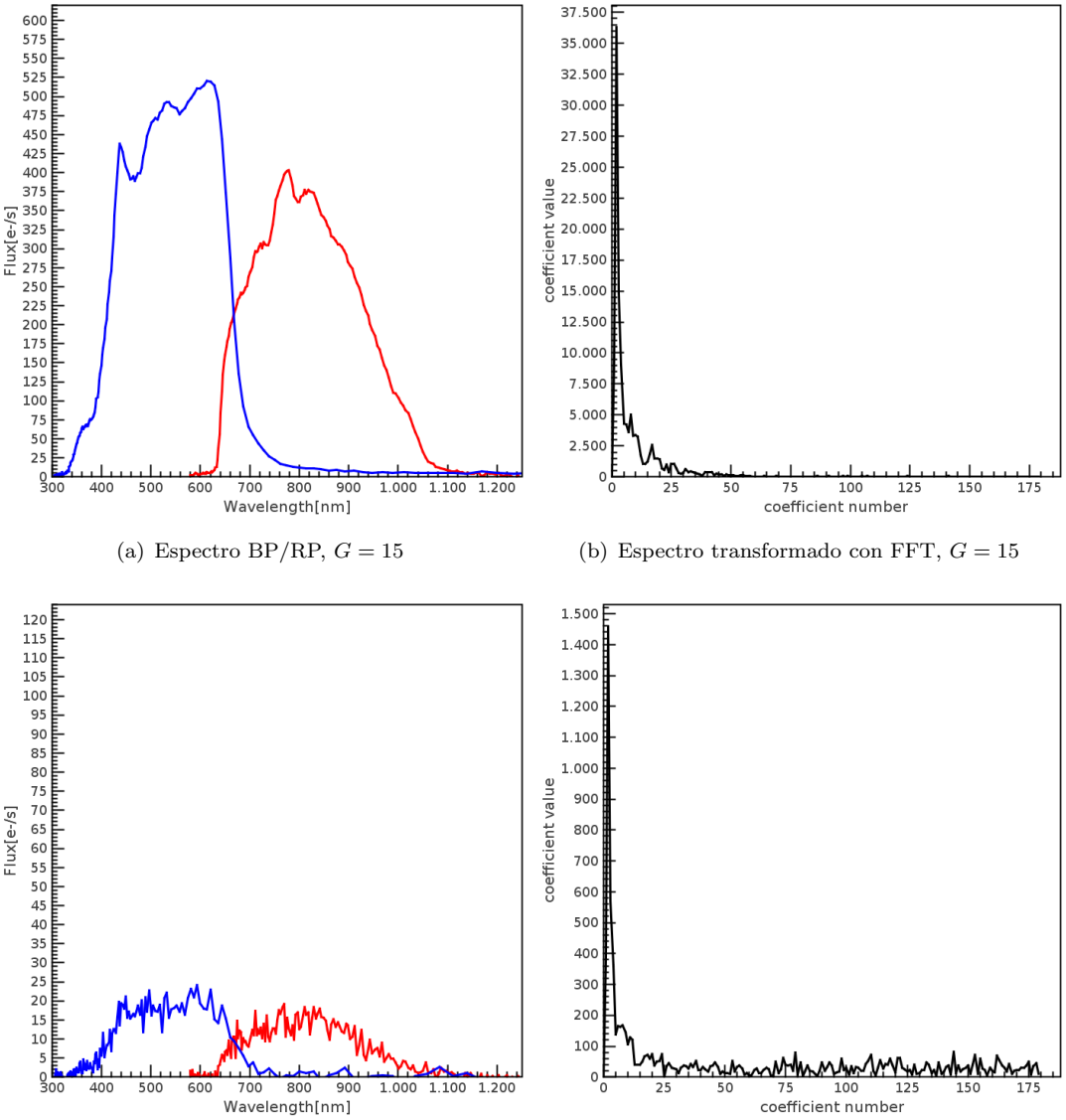
FIGURA 3.14: Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP preprocesados con distintos métodos.

Como se ha comprobado empíricamente, es beneficioso conservar, en el espectro BP/RP, tanto la varianza inherente como la inducida por los instrumentos, de cara a la clasificación no supervisada. Sería posible establecer un pesado diferente, intentando optimizar la clasificación obtenida para un conjunto de datos concreto. Sin embargo, esto introduciría la supervisión en el aprendizaje, con lo que la generalidad del método quedaría comprometida, especialmente cuando se trata el análisis de objetos atípicos.

En el capítulo 2, se ha discutido la aplicación de técnicas de procesado de señal, como son la transformada de Fourier o la transformada wavelet, para mejorar la estimación de APs. A continuación, se estudia la aplicación de dichas técnicas para realizar clasificación no supervisada. Para ello, en primer lugar, es necesario determinar cómo aplicar dichas técnicas al espectro BP/RP. En la técnica de preprocesado, descrita en la sección 3.4.1, se unen los espectros BP y RP, de forma que no exista redundancia de longitudes de onda. Sin embargo, aquí uniremos ambos espectros completos, para evitar así el salto que se produce al unir BP y RP, el cual introduciría falsos coeficientes en las representaciones.

La representación en el dominio de Fourier se obtendrá mediante la transformada rápida de Fourier (Fast Fourier Transform, FFT). Dicha transformada obtiene una serie de coeficientes complejos, de los cuales puede extraerse la amplitud de la señal y la fase de la señal. En este caso, se descarta la fase y se retiene el espectro de amplitud. Dicho espectro de amplitud es simétrico, por lo que se corta su segunda mitad, reteniendo un total de 180 coeficientes. El proceso de transformación de un espectro BP/RP para

obtener su representación en el dominio de la frecuencia se muestra en la figura 3.15. Dicha representación puede interpretarse como sigue: Las bajas frecuencias representan el continuo del espectro, las frecuencias medias las líneas de absorción/emisión y, por último, siendo el espectro limitado en banda, las frecuencias más altas representan el ruido en el espectro.



(a) Espectro BP/RP, $G = 15$

(b) Espectro transformado con FFT, $G = 15$

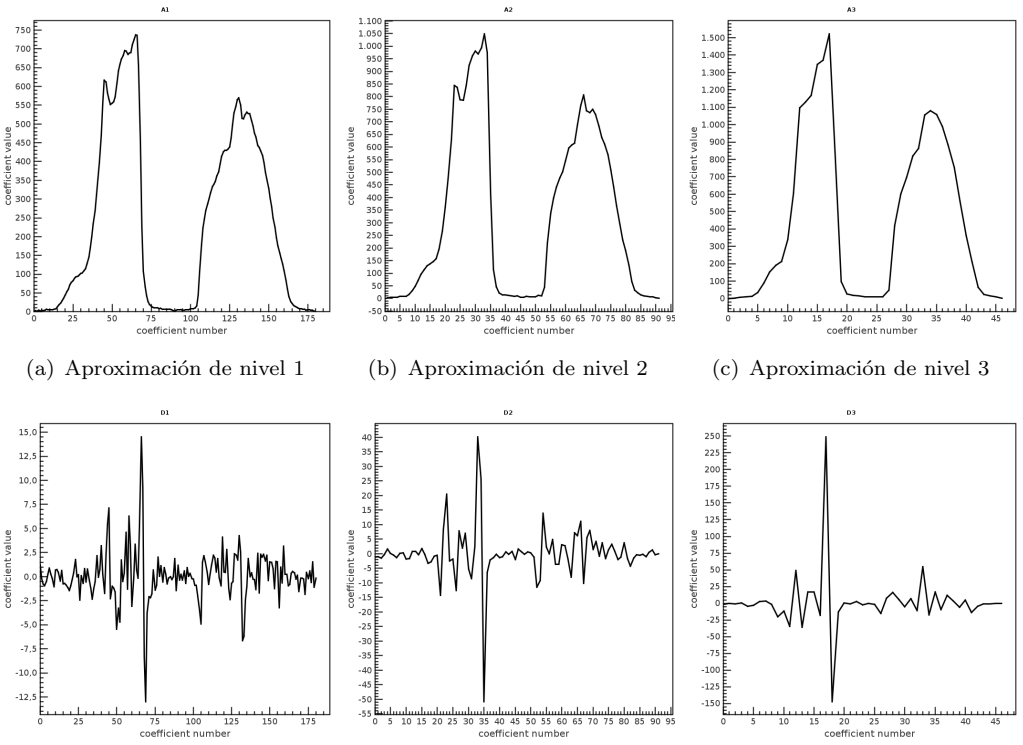
(c) Espectro BP/RP, $G = 18.5$

(d) Espectro transformado con FFT, $G = 18.5$

FIGURA 3.15: Ejemplo de representación de un espectro BP/RP, correspondiente a un quásar, en el dominio de Fourier, tanto a magnitud $G = 15$ como a magnitud $G = 18.5$.

La transformada wavelet rápida (Fast Wavelet Transform, FWT) se aplicará al espectro BP/RP para obtener la representación en el dominio del lambda-frecuencia. La metodología aplicada será la misma descrita en la sección 2.4.3.1, salvo que en este caso la wavelet madre utilizada será de tipo Daubechies con tan sólo dos momentos de desvanecimiento. Dicha elección se debe a la menor resolución que el espectro BP/RP

tiene con respecto al espectro RVS. Las figuras 3.16 y 3.17 muestran la descomposición de un espectro BP/RP en tres niveles de aproximaciones y detalles. Como puede observarse, las aproximaciones caracterizan el continuo y las líneas más anchas, mientras que los detalles caracterizan el ruido de alta frecuencia y las líneas más finas.

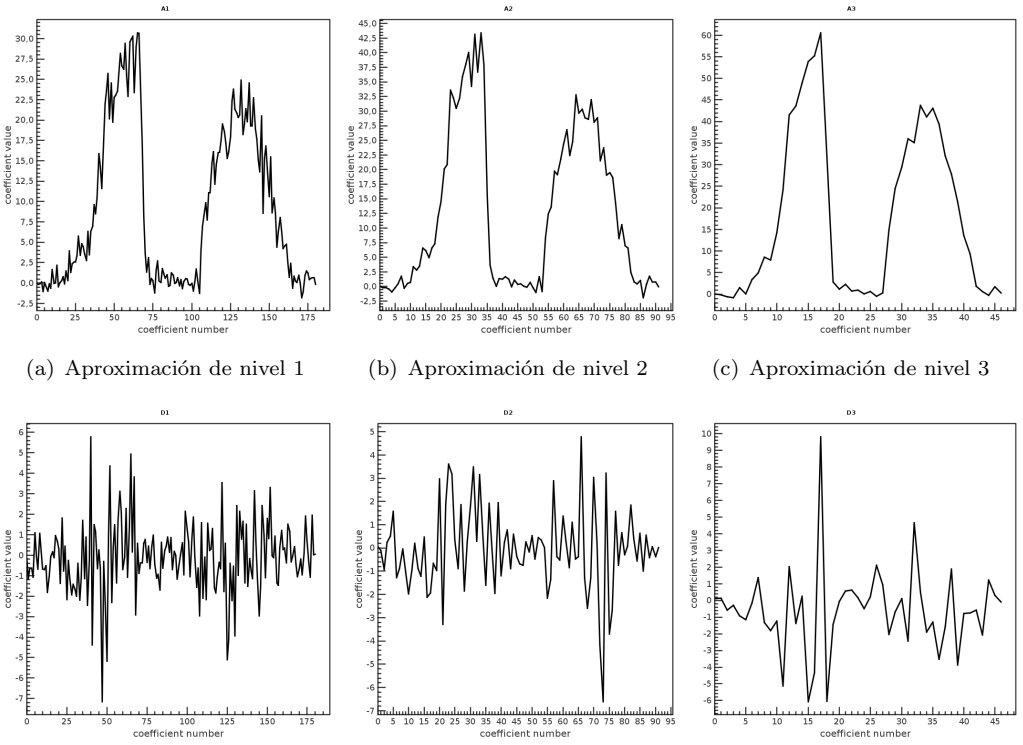


(a) Aproximación de nivel 1 (b) Aproximación de nivel 2 (c) Aproximación de nivel 3

(d) Detalles de nivel 1 (e) Detalles de nivel 2 (f) Detalles de nivel 3

FIGURA 3.16: Descomposición wavelet piramidal en tres niveles de un espectro BP/RP con $G = 15$.

A continuación, se muestran los resultados obtenidos al realizar experimentos con varios dominios de representación de los espectros BP/RP, con el objetivo de buscar la combinación más adecuada para la clasificación no supervisada. La figura 3.18 muestra los resultados obtenidos con el conjunto de datos simulados, en su versión sin ruido añadido. Puede observarse que tanto las tres aproximaciones wavelet como la representación de Fourier ofrecen resultados similares a los obtenidos con espectros originales. Sin embargo, los resultados obtenidos mediante detalles wavelet son bien diferentes. En el caso de los detalles de nivel 1, se obtiene un gran acierto de clasificación, mientras que en el caso de los detalles en niveles inferiores el acierto es muy bajo. La alta pureza que se obtiene en el caso de detalles de nivel 1 se debe a que dicha representación ensalza las líneas de absorción/emisión en los espectros BP/RP, dándoles más peso con respecto al continuo. Esto permite que el algoritmo de agrupamiento distinga mejor las características propias de objetos Be y WR, como puede observarse en la matriz de confusión mostrada en la tabla 3.4. Sin embargo, al aplicar ruido a los espectros, en



(a) Aproximación de nivel 1 (b) Aproximación de nivel 2 (c) Aproximación de nivel 3

(d) Detalles de nivel 1 (e) Detalles de nivel 2 (f) Detalles de nivel 3

FIGURA 3.17: Descomposición wavelet piramidal en tres niveles de un espectro BP/RP con $G = 18.5$.

TABLA 3.4: Matriz de confusión obtenida mediante el agrupamiento de simulaciones BP/RP de Gaia sin ruido, cuando se le presentan los detalles wavelet de nivel 1.

	AFGKM	Be	C_STARS	GALAXY	OB	PN	QSO	UCD	WD	WR	UNDEFINED
AFGKM	98,5	0	0	0	0	0	0	1,5	0	0	0
Be	1,15	80,84	0	0	17,62	0	0	0	0,38	0	0
C_STARS	91,82	0	8,18	0	0	0	0	0	0	0	0
GALAXY	0,24	0	0	97,91	0	0	1,83	0	0,01	0	0,01
OB	0	0	0	0	100	0	0	0	0	0	0
PN	0,13	0,53	0	0,27	0	96,93	2,14	0	0	0	0
QSO	0,13	0	0	1,18	0	0	98,55	0	0,13	0	0
UCD	0	0	0	0	0	0	0	100	0	0	0
WD	2,23	0	0	0	0,04	0	0	0	97,73	0	0
WR	0	6,98	0	0	1,55	0	2,33	0	1,55	87,6	0

concreto al llevarlos a magnitud $G = 18.5$, la situación cambia considerablemente, como se muestra en la figura 3.19. Los resultados obtenidos con los detalles de los 3 niveles son muy pobres. Sin embargo, las aproximaciones mejoran ligeramente los obtenidos con los espectros originales, siendo la mejor aproximación la de nivel 3. Las aproximaciones capturan los componentes de baja frecuencia de los espectros BP/RP, descartando el ruido de alta frecuencia, lo cual resulta útil cuando el nivel de ruido es significativo.

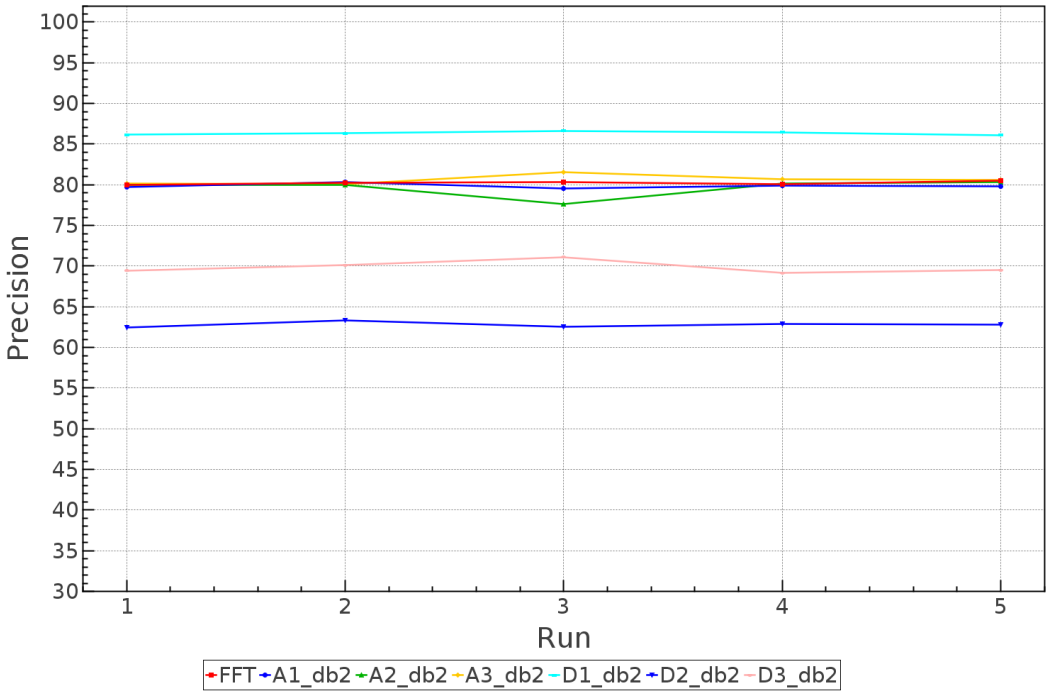


FIGURA 3.18: Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP sin ruido, transformados con diferentes técnicas de procesado de señal.

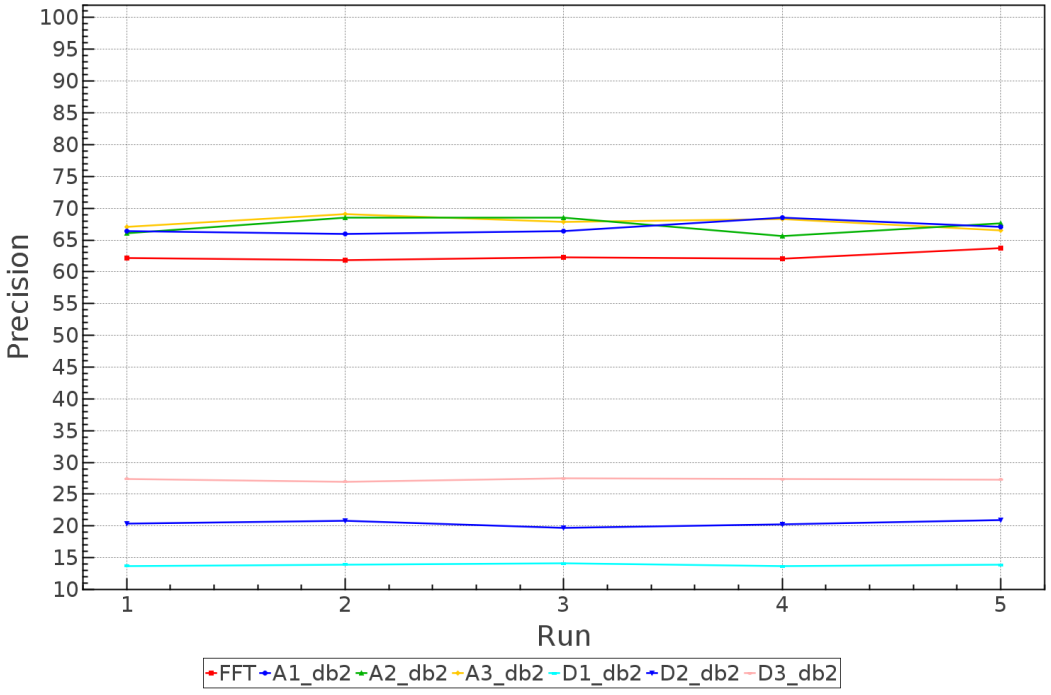


FIGURA 3.19: Evaluación del agrupamiento obtenido mediante una red SOM al presentar espectros BP/RP con $G = 18,5$, transformados con diferentes técnicas de procesado de señal.

3.4.7 Métodos de agrupamiento conjunto

Los métodos de agrupamiento conjunto (ensemble clustering methods) se basan en la premisa que dice que un conjunto de expertos, encomendados a la resolución de un determinado problema, obtendrán soluciones más robustas que las obtenidas por un único experto. A pesar de que los métodos de agrupamiento conjunto han surgido ya en el siglo XXI, se han propuesto una amplia variedad de algoritmos y aplicaciones en dicho sentido, véase [66]. En algunas de ellas se combinan instancias del mismo algoritmo de agrupamiento, sobre el mismo conjunto de datos, pero con distintas inicializaciones, o bien configurados con distintos parámetros. Por otro lado, otros métodos combinan diferentes algoritmos sobre el mismo conjunto de datos. Además, algunos métodos combinan particiones obtenidas con diferentes características. Dichas particiones pueden a su vez haber sido obtenidas mediante distintos algoritmos de agrupamiento. Aquí, con el objetivo de mejorar la clasificación no supervisada de espectros BP/RP, se aplicarán técnicas de agrupamiento conjunto que mezclan particiones obtenidas con redes SOM, a las cuales se les presentan distintas representaciones del espectro BP/RP. Dichas representaciones son las estudiadas en la sección anterior, la representación en el espacio de Fourier y la representación wavelet. De esta forma, se combinan los mapas obtenidos con cada una de ellas, obteniendo una partición en la cual los espectros en un mismo grupo comparten las mismas propiedades, tanto en el espacio original (lambda) como en el dominio de la frecuencia y del lambda-frecuencia. Para realizar dicha combinación, se ha desarrollado un algoritmo basado en votos sobre cada par de objetos, el cual se especifica en los pseudocódigos 1 y 2.

Algoritmo 1 Combinar particiones. X: conjunto de observaciones, P: conjunto de particiones.

```
function combinePartitions(X,P)
    for all (o_i, o_j) ∈ X do
        together ← true
        for all p_k ∈ P do
            s_{i,k} ← cluster(o_i)
            s_{j,k} ← cluster(o_j)
            if s_{i,k} ≠ s_{j,k} then
                together ← false
            end if
        end for
        if together then
            C ← addToEnsembledClusters(o_i, o_j, C)
        end if
    end for
     return C
end function
```

Algoritmo 2 Añadir par de objetos a los grupos combinados. o_i: primer objeto del par, o_j: segundo objeto del par, C: conjunto de grupos combinados en el momento actual.

```
function addToCombinedClusters(o_i,o_j,C)
    for all c_i ∈ C do
        if (o_i ∈ c_i) ∧ (o_j ∈ c_i) then return C
        else
            if (o_i ∈ c_i) ∧ (o_j ∉ c_i) then
                c_i ← o_j return C
            end if
            if (o_i ∉ c_i) ∧ (o_j ∈ c_i) then
                c_i ← o_i return C
            end if
        end if
    end for
    C ← createCluster(o_i, o_j) return C
end function
```

La partición conjunta que se obtiene como resultado de la combinación cumple las siguientes propiedades:

- Para cada par de objetos, si ambos caen en el mismo grupo en todas las particiones, entonces los dos caen en el mismo grupo de la partición conjunta.
- Las observaciones, que no se emparejan en el mismo grupo con ninguna de las otras observaciones, no forman parte de ningún grupo combinado.
- Si un grupo combinado contiene N objetos, entonces todos ellos caen en el mismo grupo en todas las particiones combinadas.

Este método de agrupamiento conjunto genera una gran cantidad de grupos como resultado. Por lo tanto, se realiza una selección de los grupos más poblados, que son los que se someterán al análisis posterior. La tabla 3.5 muestra los resultados obtenidos al combinar particiones, de cuyos grupos resultantes se han seleccionado los 900 más poblados. Las particiones seleccionadas para realizar la combinación son la partición obtenida con espectros originales (*lambda*), la partición obtenida con espectros transformados mediante la transformada de Fourier (FFT) y los espectros transformados con la aproximación de nivel tres de la descomposición wavelet ($A3$). Como puede observarse, los 900 grupos obtenidos mediante la combinación son menos puros, con respecto a los 900 grupos obtenidos con los espectros originales, para la magnitud $G = 15$. Por otro lado, los grupos obtenidos mediante agrupamiento conjunto son más robustos en el caso de $G = 18.5$, especialmente cuando se combinan los tres dominios (*lambda*, FFT y $A3$). Esto implica que la técnica de agrupamiento conjunto descrita tiene interés para la recuperación de grupos de objetos en conjuntos de datos ruidosos, como el que

TABLA 3.5: Resultados obtenidos por el método de agrupamiento conjunto. Particiones obtenidas mediante una red SOM 30*30 que agrupa representaciones de espectros simulados BP/RP.

Magnitud G	**Combinación**	**Número de objetos**	**Precision**
15	Lambda+FFT	95740	75,3
15	Lambda+FFT+A3	74675	76,3
18,5	Lambda+FFT	46103	67,8
18,5	Lambda+FFT+A3	26737	74,4

OA procesará con toda probabilidad. Sin embargo, además de la pérdida de objetos sufrida al realizar la selección de grupos, la técnica de combinación descrita conlleva la pérdida de la vecindad entre los grupos obtenidos, por lo que se pierde la capacidad de visualizar su distribución en un plano 2D, lo cual resulta de gran utilidad a la hora de analizar los grupos a posteriori.

Con el objetivo de paliar los defectos de la técnica de agrupamiento conjunto descrita anteriormente, se ha diseñado un segundo algoritmo de combinación de particiones que es capaz de mantener la vecindad en los grupos obtenidos, como se muestra en el algoritmo 3. Para ello, se fija una de las particiones como partición base, y se usan las otras como particiones de filtrado. Así, para cada objeto, se fija un porcentaje mínimo de objetos de su grupo en la partición base, los cuales deben caer en el mismo grupo que el objeto en las particiones de filtrado. En caso contrario, el objeto se filtra de la partición. La tabla 3.6 muestra los resultados obtenido cuando el algoritmo de combinación por filtrado se aplica al conjunto de simulaciones BP/RP, con varias magnitudes y variando el porcentaje mínimo de objetos que deben coincidir en el mismo grupo. Observando los resultados, llama la atención el bajo valor de *precision* obtenido para el caso de $G = 15$ con un porcentaje mínimo del 90%, así como el obtenido para el caso de $G = 18.5$, con cualquier porcentaje mínimo. No obstante, el valor de *accuracy* para estos casos es cercano al 100%. Esto se debe a que tan sólo algunas de las clases han pasado el filtrado. Por ejemplo, en el caso de $G = 18.5$ con el porcentaje del 60%, la matriz de confusión quedaría como la mostrada en la tabla 3.7, donde se observa que sólo han pasado el filtro objetos de tipo Be, PN, QSO y UCD. Por lo tanto, el algoritmo de filtrado de particiones se muestra efectivo, ya que selecciona los objetos que se agrupan con mayor pureza, lo cual podría ser útil para algoritmos como OA, que buscan recuperar grupos de objetos homogéneos de conjuntos de datos con mucho ruido. Además, como se ha comentado anteriormente, este algoritmo preserva la vecindad de la red SOM, por lo que es posible la visualización de la partición filtrada, tal como se muestra en la figura 3.20.

Algoritmo 3 Combinar particiones mediante filtrado. X: conjunto de observaciones, $p1$: partición base, $P2$: particiones de filtrado, t: porcentaje mínimo de objetos que deben coincidir en el mismo grupo.

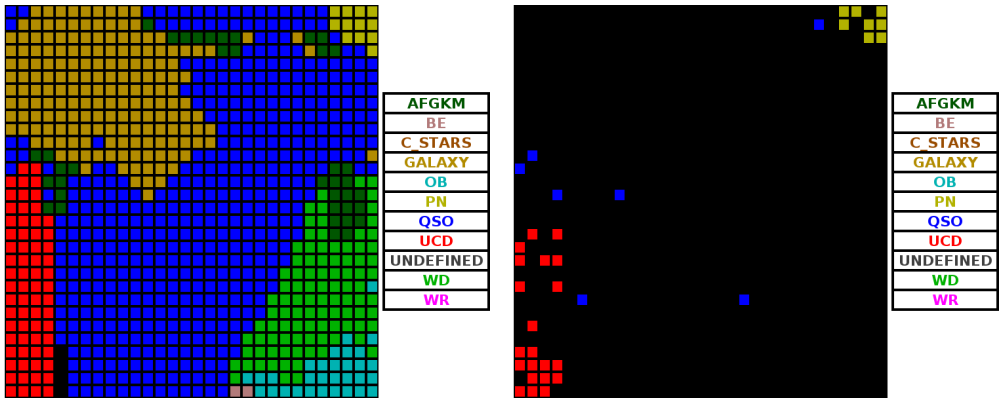
function FILTERPARTITION(X,p1,P2,t)
  for all $o_i \in X$ **do**
    $c_i \leftarrow cluster(o_i) \in p1$
    $countTogether \leftarrow 0$
    for all $o_j \in ci$ **do**
      $together \leftarrow true$
      for all $p_k \in P2$ **do**
        $c_{j,k} \leftarrow cluster(o_j) \in p_k$
        if $c_{j,k}! = c_i$ **then**
          $together \leftarrow false$
        end if
      end for
      if $together$ **then**
        $countTogether \leftarrow countTogether + 1$
      end if
    end for
    if $\frac{countTogether}{|c_i|} < t$ **then**
      $filter(o_i, p1)$
    end if
  end for
  return p1
end function

TABLA 3.6: Resultados obtenidos por el método de agrupamiento conjunto por filtrado. Particiones obtenidas mediante una red SOM 30*30 que agrupa representaciones de espectros simulados BP/RP.

Magnitud G	**Porcentage t**	**Número de objetos**	**Precision**
15	60	20572	75
15	75	9366	69,5
15	90	4785	40
18,5	60	2862	40
18,5	75	1514	30
18,5	90	505	20

TABLA 3.7: Matriz de confusión obtenida mediante el agrupamiento conjunto por filtrado, con magnitud $G = 15$ y porcentaje $t = 60$.

	AFGKM	Be	C_STARS	GALAXY	OB	PN	QSO	UCD	WD	WR	UNDEFINED
AFGKM	0	0	0	0	0	0	0	100	0	0	0
Be	0	100	0	0	0	0	0	0	0	0	0
C_STARS	-	-	-	-	-	-	-	-	-	-	-
GALAXY	0	0	0	0	0	0	100	0	0	0	0
OB	-	-	-	-	-	-	-	-	-	-	-
PN	0	0	0	0	0	100	0	0	0	0	0
QSO	0	0	0	0	0	0	100	0	0	0	0
UCD	0	0	0	0	0	0	0	100	0	0	0
WD	-	-	-	-	-	-	-	-	-	-	-
WR	0	100	0	0	0	0	0	0	0	0	0
Num. Objetos	1	25	0	7	0	323	433	2072	0	1	0



(a) Clase predominante, SOM original (b) Clase predominante, SOM filtrada

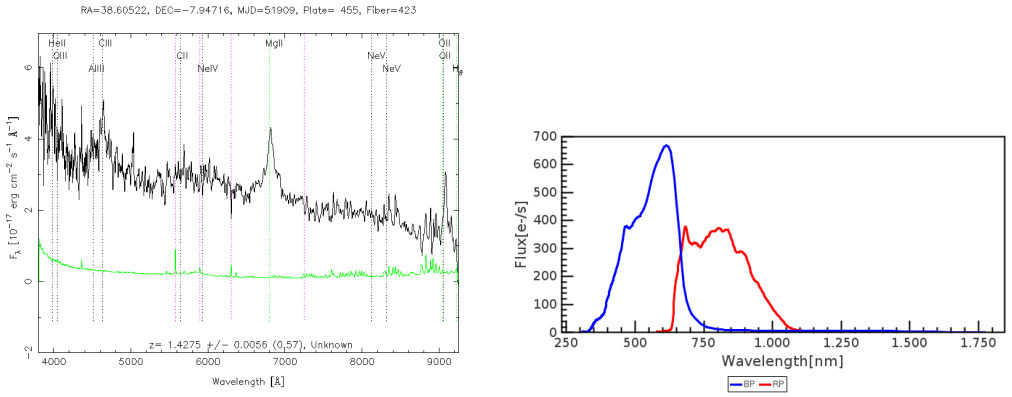
FIGURA 3.20: Distribución de los grupos de una SOM 30*30, obtenidos mediante agrupamiento conjunto por filtrado, para espectros BP/RP con magnitud 18,5 y un porcentaje $t = 60$.

3.5 Análisis de objetos atípicos en SDSS

En la sección 3.4 se han estudiado distintas técnicas de clasificación no supervisada de espectros, evaluando las mismas con el conjunto de simulaciones que está siendo utilizado para preparar la misión Gaia. Sin embargo, es muy difícil simular un conjunto de objetos atípicos como el que será procesado cuando los datos de Gaia estén disponibles. Por lo tanto, es necesario comprobar las técnicas desarrolladas para OA mediante su aplicación a bases de datos astronómicas con observaciones reales. Afortunadamente, hoy en día existen catálogos con una gran cantidad de espectros astronómicos, como el llamado Sloan Digital Sky Survey (SDSS, ver [67]). Por lo tanto, aplicaremos los algoritmos desarrollados para OA al estudio de los espectros atípicos de SDSS, aquellos que no han podido ser clasificados como una de las clases esperadas. Dicho estudio es interesante no sólo de cara a la preparación para Gaia, si no que también representa un experimento de alto interés científico.

El catálogo SDSS contiene alrededor de un millón de espectros, la gran mayoría de origen extragaláctico (galaxias y quásares). Puede considerarse como uno de los catálogos más importantes de la historia de la astronomía, ya que sus contribuciones han sido muy importantes. Por ejemplo, gracias a SDSS se han descubierto los quásares más distantes y se ha obtenido una descripción muy precisa de la población de galaxias en el Universo y su relación con la distribución de la materia oscura, así como con otras variables cosmológicas. Existen varias versiones del catálogo, las cuales son liberadas periódicamente. Aquí nos centraremos en la versión 7 (ver [1]), de la cual se han extraído los 10000 espectros para los que el software de SDSS no ha podido asignar una clasificación fiable. Dichos espectros cubren el rango [3800Å, 9200Å] y tienen una resolución media de entre R=1850 a R=2200, lo que se traduce en más de 3000 valores

por espectro. Además de los espectros en formato original, se ha utilizado el simulador GOG para obtener los espectros en formato BP/RP, para estudiar así el impacto de de dicha transformación. La figura 3.21 muestra un ejemplo de transformación de un espectro que corresponde a un quásar de SDSS.

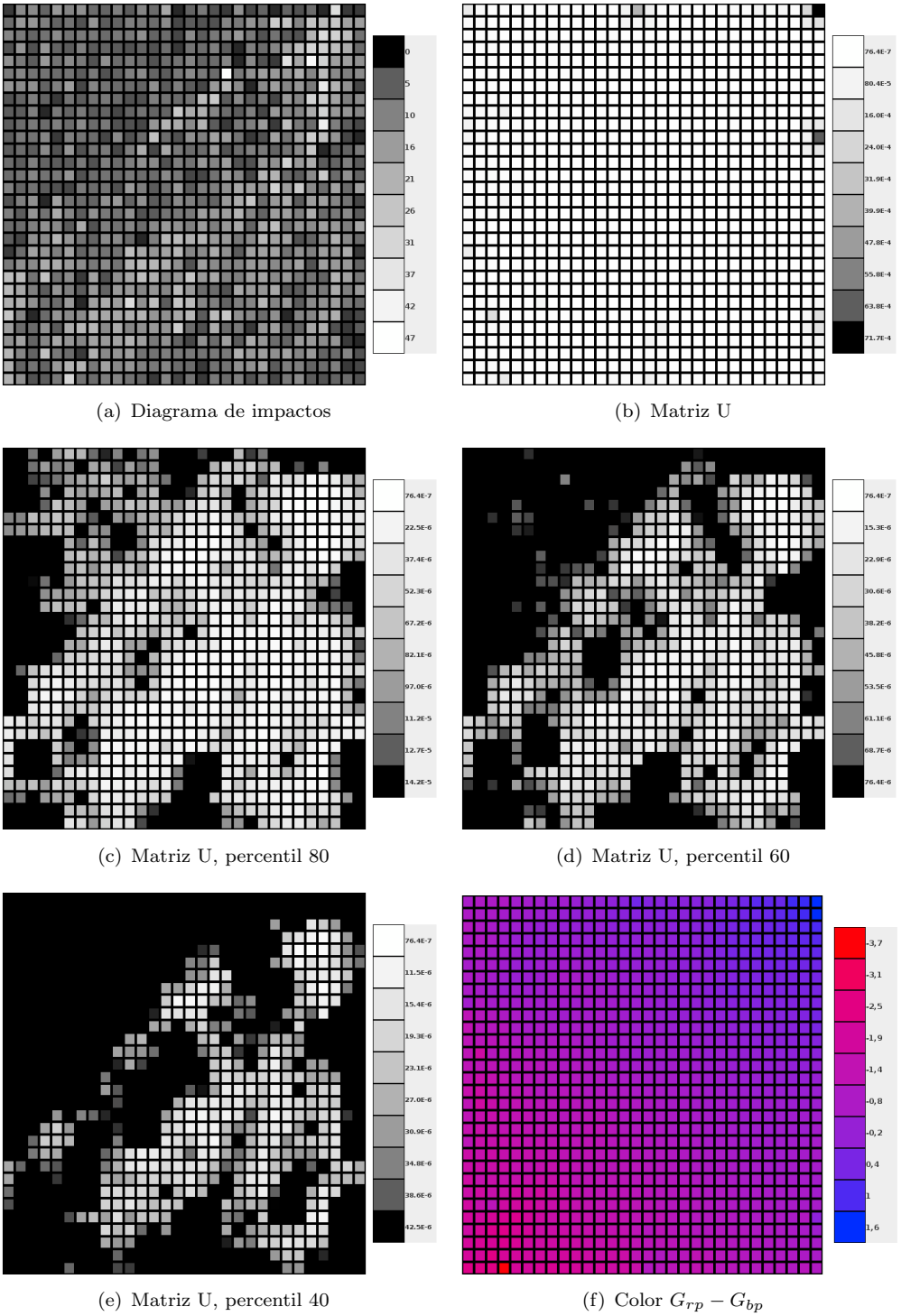


(a) Espectro original, extraído de SDSS (b) Espectro transformado al formato BP/RP

FIGURA 3.21: Transformación de un espectro de SDSS al formato BP/RP mediante el simulador GOG.

El conjunto de espectros atípicos de SDSS, convertido a formato BP/RP, y preprocesado según el método presentado en la sección 3.4.1, se ha analizado mediante el entrenamiento de una red SOM. Se ha seleccionado un mapa bidimensional con 30*30 neuronas, con los parámetros de configuración descritos en la sección 3.4.4. Tras el aprendizaje de 10 mapas con diferente inicialización, se ha escogido el que obtiene un mínimo en el error de cuantización (ecuación 3.3). La figura 3.22 muestra las visualizaciones obtenidas para dicho mapa. Puede observarse que hay una gran región en la parte superior izquierda de la matriz U cuyas neuronas quedan fueran de los percentiles, así como otras regiones más pequeñas en los bordes inferior y derecho. Por otro lado, dentro de los percentiles, pueden observarse varios grupos diferenciados con alta densidad. Mediante el estudio de los prototipos, así como el acceso a las imágenes de los objetos a través del SkyServer de SDSS [68], ha permitido descubrir que las zonas en negro, al menos hasta el percentil 60, se corresponden con espectros extraídos de forma errónea (valores perdidos, espectros superpuestos), o bien con falsas detecciones de los objetos en las imágenes, como en los ejemplos que se muestran en la figura 3.23. Sin embargo, la exploración de zonas densas, cuyas neuronas reciben un color claro, ha determinado que estas están pobladas por objetos astrofísicos, correctamente detectados en las imágenes y con espectros correctamente extraídos.

Gracias al entrenamiento de una red SOM, se han conseguido agrupar los 10000 objetos atípicos en 900 grupos homogéneos, lo cual facilita su análisis. Además, la exploración de los grupos puede guiarse con las visualizaciones del mapa bidimensional. De esta forma,



(a) Diagrama de impactos

(b) Matriz U

(c) Matriz U, percentil 80

(d) Matriz U, percentil 60

(e) Matriz U, percentil 40

(f) Color $G_{rp} - G_{bp}$

FIGURA 3.22: Visualización de una red SOM entrenada con espectros atípicos de SDSS.

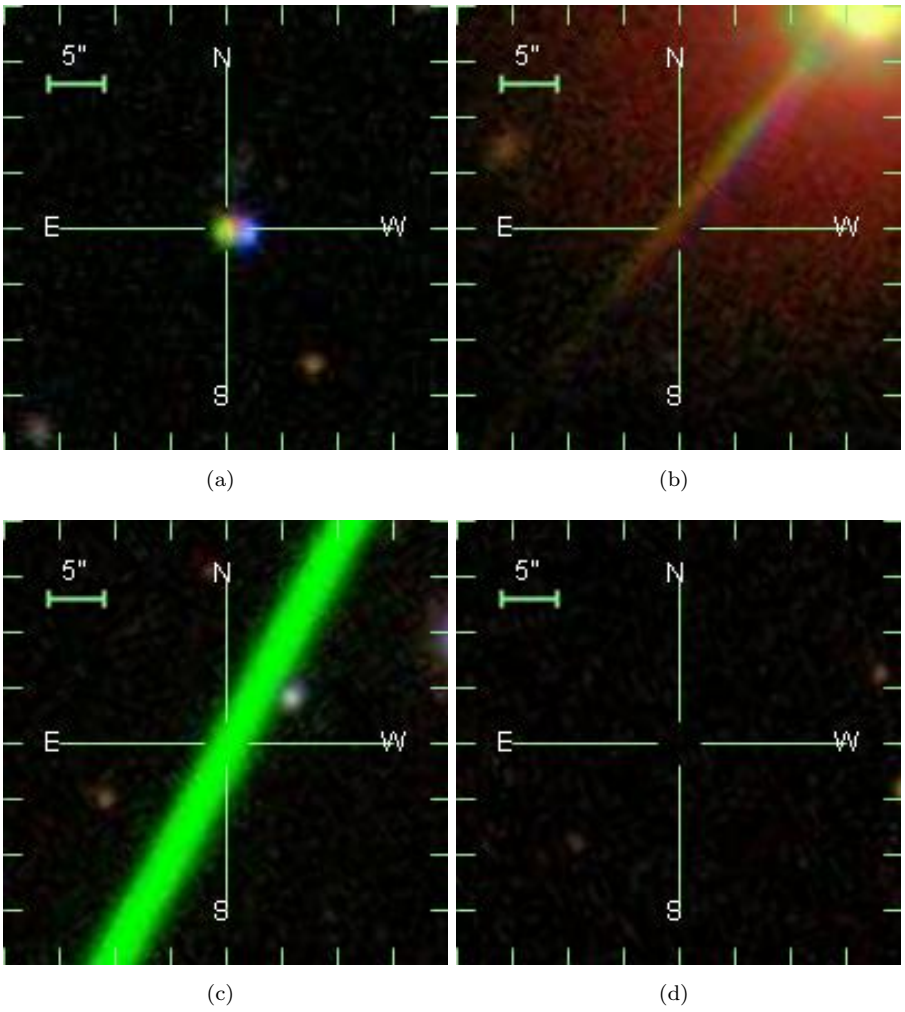


FIGURA 3.23: Detecciones erróneas de objetos en SDSS, encontradas en la región superior izquierda del mapa SOM.

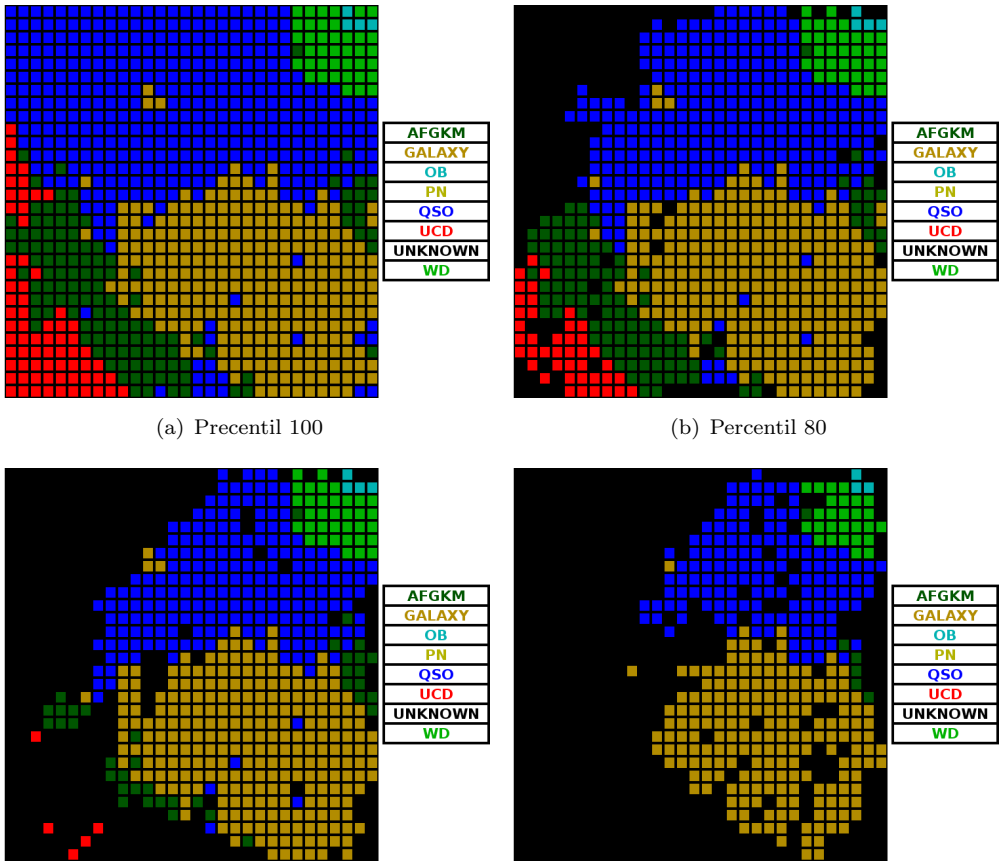
se ha conseguido separar observaciones erróneas, bien sea por problemas en la extracción del espectro o por falsas detecciones, de objetos astrofísicos de interés científico. Aunque dicha información resulta de gran valor, es posible identificar la naturaleza de los grupos de forma más completa gracias a la información existente en catálogos externos. En la siguiente sección, se profundiza en el cruce de los grupos obtenidos con catálogos externos.

3.5.1 Identificación de grupos en la red SOM mediante cruce con catálogos externos

El estudio de espectros e imágenes, así como de otras propiedades en el catálogo de estudio, en este caso SDSS, resulta de gran utilidad a la hora de identificar los grupos formados por la red SOM. Sin embargo, también es posible recurrir a información

externa. Tal información puede consistir en espectros de mejor calidad, espectros en otro rango electromagnético, clasificaciones de los objetos, etc. Con el objetivo de identificar los grupos de espectros atípicos en la red SOM, se han seguido varias estrategias. En primer lugar, se han cruzado los espectros BP/RP simulados para Gaia con los grupos de espectros atípicos, ya que el conjunto de simulaciones incluye una amplia variedad de clases espectrales, algunas de las cuales no se contemplan en el catálogo SDSS. El cruce se realiza mediante un proceso de minimización de distancia euclídea entre el prototipo del grupo de espectros atípicos y los espectros simulados. Dicho proceso es similar a un clasificador por k vecinos más cercanos (k-nearest neighbors, KNN), en el lenguaje del aprendizaje máquina. La figura 3.24 muestra el resultado de la identificación realizada con el método del vecino más cercano. Teniendo en cuenta que estamos intentado clasificar objetos atípicos, las clasificaciones obtenidas no son fiables y han de interpretarse como información de referencia. No obstante, podemos analizar la bondad del ajuste entre el objeto a clasificar y su plantilla más cercana. La figura 3.24 muestra también las identificaciones filtradas mediante el método del corte por percentiles sobre la máxima distancia. Puede observarse que se oscurecen las mismas zonas cuando se aplican los percentiles, con respecto a lo obtenido en el caso de la matriz U. Esto confirma la separación entre objetos astrofísicos y observaciones erróneas que se había obtenido en la sección anterior.

La segunda estrategia de cruce con catálogos externos implementada en este estudio se basa en el emparejamiento de las posiciones de los astros en el cielo. Para ello, se mide la distancia angular entre las coordenadas (ascensión recta, declinación) de dos objetos dados y se establece un radio, dentro del cual se considera que se trata del mismo objeto. En este caso, se han cruzado los objetos con espectros atípicos en SDSS con los objetos del catálogo llamado Simbad, con un radio de un segundo de arco. Simbad es un catálogo astronómico general que incluye clasificaciones obtenidas para un gran número de objetos mediante distintos métodos, véase [69]. De los 10000 objetos con espectros atípicos, se han encontrado más de 2000 objetos clasificados en Simbad, con coordenadas dentro del radio de emparejamiento. En las 2000 etiquetas recuperadas se encuentran las clases: AGN, Seyfert1, Seyfert2, Bl-Lac, galaxy, QSO, fuentes de radio (en general), white dwarf, brown dwarf, y low-mass. Por simplicidad, las clases AGN, Seyfert1, Seyfert2 y Bl-Lac se han agrupado bajo la etiqueta AGN, que representa cualquier tipo de objeto extragaláctico activo que no sea un quásar. El resumen de tipos de etiquetas entre las recuperadas de Simbad se muestra en la tabla 3.8, mientras que la figura 3.25 muestra la distribución de las etiquetas en el mapa SOM. Como puede observarse, las distintas clases de objetos caen en zonas significativamente separadas del mapa. Cabe destacar que tanto las enanas blancas, como los quásares y las enanas marrones ocupan las posiciones similares a las que ocupan en la identificación por medio



(a) Precentil 100 (b) Percentil 80

(c) Percentil 60 (d) Percentil 40

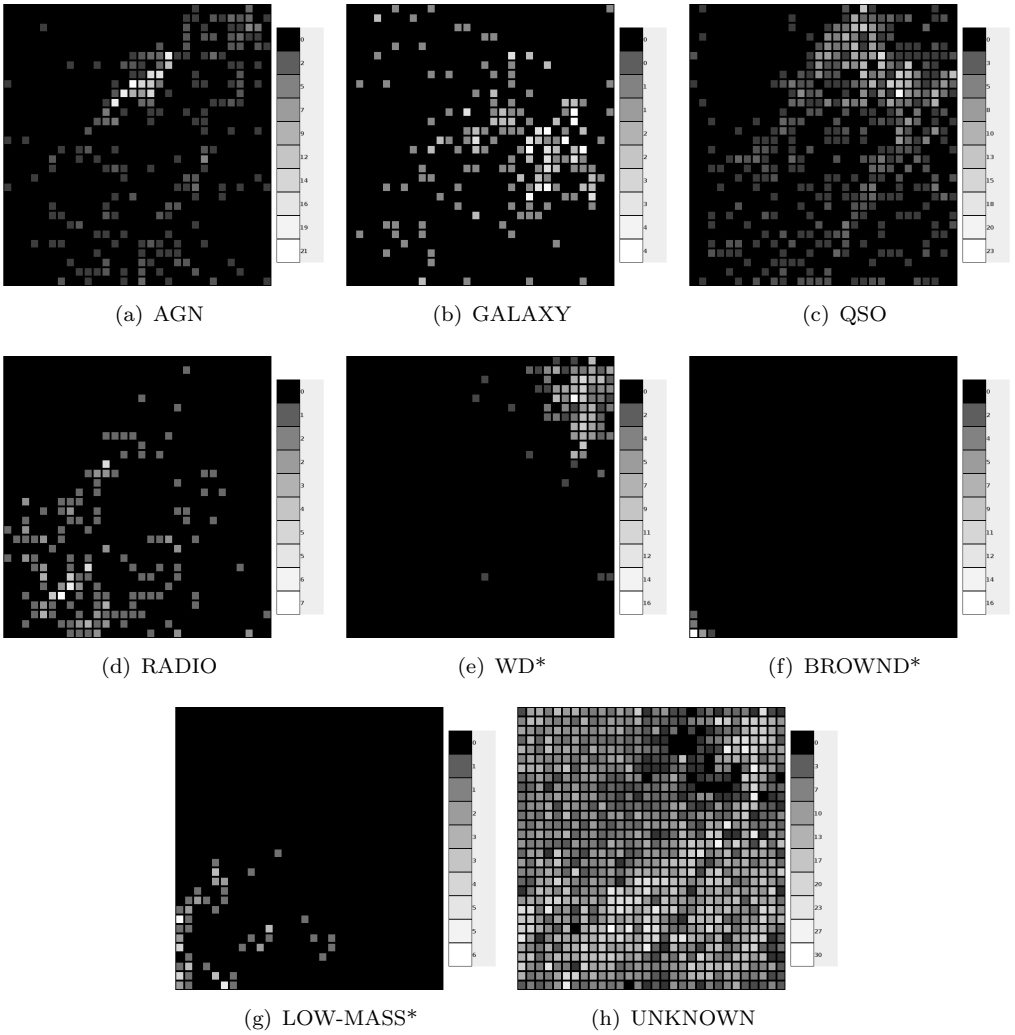
FIGURA 3.24: Identificaciones obtenidas mediante el cruce con espectros BP/RP simulados para Gaia. Se aplican varios percentiles para filtrar las clasificaciones no fiables.

de espectros simulados, véase la figura 3.24, lo cual incrementa nuestra confianza en el método. La figura 3.26(a) proporciona una vista más compacta de la distribución de etiquetas en el mapa. Dicha figura se ha obtenido de forma similar a la figura 3.11(a), sólo que en este caso no se ha tenido en cuenta la clase UNKNOWN para obtener la clase más frecuente, de forma que un grupo sólo recibe la clase UNKNOWN si no contiene ningún objeto etiquetado en Simbad. Además, algunos grupos reciben la etiqueta UNDEFINED, que se asigna cuando en un grupo existe una frecuencia similar entre dos o más clases (la clase ganadora debe ser un 10% más frecuente que la segunda). Por otro lado, la figura 3.26(b) muestra la pureza de los grupos en función de las etiquetas contenidas por los mismos. Se confirma de nuevo la hipótesis de que la región superior izquierda está compuesta por observaciones erróneas, ya que casi en ningún grupo de dicha región de encuentran objetos identificados en Simbad.

A través de las etiquetas recuperadas de Simbad, también es posible confeccionar una matriz de confusión, como la mostrada en la tabla 3.9. Puede observarse que, a pesar de la incertidumbre inducida por el método de cruce por emparejamiento de coordenadas y

TABLA 3.8: Número de objetos por cada tipo de etiqueta entre las recuperadas de Simbad. La etiqueta UNKNOWN se refiere a objetos para los cuales no se ha encontrado correspondencia.

Tipo de etiqueta	**Número de objetos**
AGN	438
GALAXY	221
QSO	988
RADIO	183
WD*	297
BROWND*	27
LOW-MASS*	73
UNKNOWN	7898

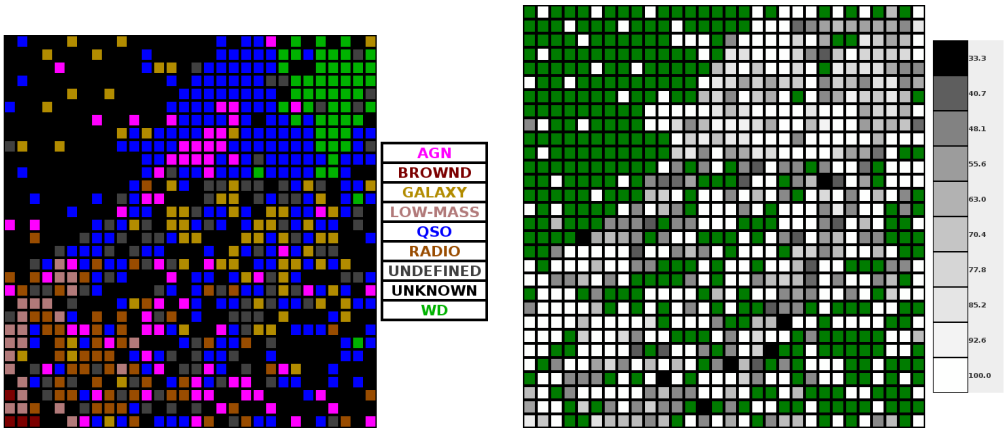


(a) AGN (b) GALAXY (c) QSO

(d) RADIO (e) WD* (f) BROWND*

(g) LOW-MASS* (h) UNKNOWN

FIGURA 3.25: Número de impactos por cada una de las clases entre las etiquetas recuperadas de Simbad.



(a) Clase más frecuente

(b) Pureza. Un color verde indica un grupo sin etiquetas

FIGURA 3.26: Identificación de los grupos de la red SOM calculada mediante las etiquetas recuperadas de Simbad.

TABLA 3.9: Matriz de confusión obtenida mediante la distribución de etiquetas recuperadas de Simbad sobre la red SOM entrenada con espectros atípicos de SDSS.

	AGN	Galaxy	QSO	Radio	UNKNOWN	WD*	brownD*	low-mass*	UNDEFINED
AGN	60,05	1,14	10,5	2,97	0	9,82	0	0,46	15,07
Galaxy	4,52	47,96	18,55	2,26	0	1,36	0	0	25,34
QSO	4,55	1,01	79,05	1,32	0	3,95	0,1	0,2	9,82
Radio	4,37	1,09	7,65	58,47	0	0	0	0	28,42
UNKNOWN	6,75	8,88	17,61	7,42	32,53	7,9	0,22	6,05	12,65
WD*	2,69	0,34	4,71	0	0	85,86	0	0	6,4
brownD*	0	0	0	0	0	0	81,48	0	18,52
low-mass*	0	0	1,37	2,74	0	0	4,11	76,71	15,07

por el proceso de clasificación por el que se obtuvieron las etiquetas, la clasificación no supervisada resulta considerablemente efectiva. Además, de la tabla pueden extraerse las posibilidades de descubrimiento de objetos novedosos entre los que aún están por clasificar, observando la fila de objetos con clase UNKNOWN. Puede observarse que, de los 8000 objetos sin etiqueta en Simbad, 624 son candidatos a convertirse en nuevas enanas blancas, 1674 son candidatos a ser nuevos quásares y así sucesivamente.

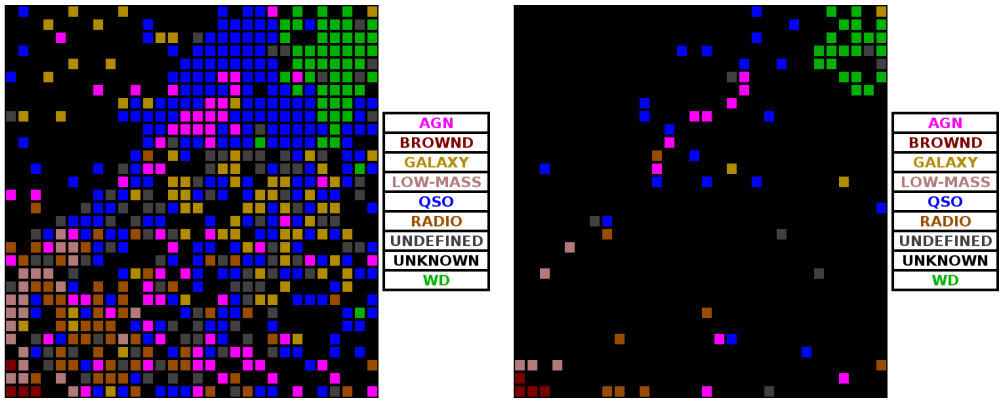
Todo el análisis descrito en esta sección se ha repetido con los espectros en su formato original de SDSS. Como conclusión, se ha determinado que la calidad de los mapas obtenidos no mejoran con respecto a los que se entrenan con espectros en formato BP/RP. Esto se debe, por un lado, a la robustez de las redes SOM al ruido y, por otro lado, a la baja señal a ruido de los espectros atípicos de SDSS, cuya magnitud se mueve alrededor del valor 19, de forma que la baja resolución del espectro BP/RP se ve compensada con un mejor SNR.

TABLA 3.10: Matriz de confusión obtenida mediante el agrupamiento conjunto de espectros atípicos de SDSS.

	AGN	Galaxy	QSO	Radio	UNKNOWN	WD	brownD	low-mass	UNDEFINED
AGN	66,1	0	1,69	3,39	22,03	1,69	0	0	5,08
Galaxy	0	0	0	0	66,67	0	0	0	33,33
QSO	7,27	0	54,55	0	27,27	0	0	0	10,91
Radio	0	0	0	46,15	38,46	0	0	0	15,38
UNKNOWN	1,75	0	0,48	0,16	94,13	0,79	0,32	0	2,38
WD	0	0	0	0	77,03	14,86	0	0	8,11
brownD	0	0	0	0	36	0	64	0	0
low-mass	0	0	0	0	84,62	0	15,38	0	0
Num. objetos	59	6	55	13	630	74	25	13	0

3.5.2 Agrupamiento conjunto de espectros de SDSS

Como ha sido descrito en la sección 3.4.7, los métodos de agrupamiento conjunto permiten la selección de grupos homogéneos de objetos. A continuación, se aplica la técnica de combinación por filtrado, tomando como partición base la obtenida con espectros originales (en formato BP/RP), mientras que las particiones obtenidas con la representación de Fourier y la aproximación wavelet de nivel 3 se toman como particiones de filtrado. Se toma un porcentaje mínimo de tan sólo el 20%, ya que de otra forma se filtrarían un gran número de objetos. Esto se debe a que el número de objetos por grupo en este caso es considerablemente menor que en el caso de las particiones obtenidas con simulaciones de Gaia. Los resultados obtenidos mediante la técnica de combinación pueden observarse en la figura 3.27 y en la tabla 3.10. Como puede observarse, la combinación resulta en una partición más pura, en el sentido de que existe una mezcla menor entre las clases de Simbad. Sin embargo, tan sólo 875 objetos han pasado el filtrado. Los grupos combinados pueden utilizarse para un análisis más fiable del conjunto de datos, o bien como una aproximación inicial a la exploración del conjunto.



(a) Clase predominante, SOM original (b) Clase predominante, SOM filtrada

FIGURA 3.27: Distribución de los grupos obtenidos mediante agrupamiento conjunto por filtrado, para espectros procedentes de SDSS y $t = 20$.

3.6 Métodos para el análisis de objetos atípicos en Gaia

Con respecto al análisis realizado con espectros atípicos de SDSS, el análisis de los espectros BP/RP que obtendremos con Gaia supondrá afrontar nuevos retos:

1. Será necesario corregir el enrojecimiento al que se verán sometidos los espectros, principalmente en la zona del plano galáctico, cuya observación no contempla SDSS, pero sí Gaia. Desafortunadamente, no es posible estimar el enrojecimiento mediante espectros BP/RP sin conocer la naturaleza de los objetos, ya que estos no han sido clasificados cuando son procesados por OA. Para resolver dicho problema, se estimará el enrojecimiento de los espectros mediante un mapa de extinción en tres dimensiones. Sin embargo, es posible que dicho mapa no proporcione soluciones fiables cuando no se pueda estimar la distancia a los objetos de forma precisa.

2. La cantidad de cómputo necesario para entrenar los mapas auto-organizativos será mucho mayor, ya que OA procesará presumiblemente millones de espectros atípicos.

3. OA procesará dos conjuntos diferentes con objetos atípicos: uno con objetos descartados por el detector de objetos atípicos de DSC y el otro con objetos clasificados con baja probabilidad.

4. El cruce mediante el emparejamiento de posiciones en el cielo resultará complicado en algunos casos, ya que Gaia tiene una resolución angular mucho mayor que la de otros catálogos existentes.

A pesar de las dificultades que entrañará el procesado de datos, Gaia permitirá obtener datos no disponibles en catálogos actuales, como paralajes y movimientos propios para un gran número de estrellas, además de variabilidad fotométrica. Dicha información podría ser utilizada por OA para asistir el proceso de identificación de espectros atípicos.

3.7 Implementación

Los experimentos realizados en este capítulo han sido realizados mediante un software implementado y testado en el lenguaje de programación Java. Dicho software se compone de varias librerías, las cuales incluyen pruebas de unidad para asegurar su correcto funcionamiento. La principal librería es la llamada *ClusteringToolkit*, en la cual se implementan los algoritmos de agrupamiento presentados en este capítulo. Además, se

hace uso de librerías extra como *libwavelet* en el empleo de técnicas de procesado de señal. El manejo de espectros BP/RP y de la preparación de los experimentos se realiza mediante la librería *OutlierAnalysisTests*.

3.7.1 Implementación en APSIS

El paquete OA ha sido integrado con el resto de paquetes de APSIS. Será ejecutado en el ciclo siguiente a DSC y OCA, de forma que procesará los millones de objetos que no han podido ser clasificados con fiabilidad. La implementación de OA tiene un diseño modular, de forma que sus funcionalidades están desacopladas. OA está formado por cinco módulos, cuyo flujo de datos se muestra en la figura 3.28 y se describe a continuación:

1. En primer lugar, cada una de las fuentes observadas por Gaia es procesada por el módulo de selección, el cual decide si la fuente debe ser procesada o no, a través de las salidas obtenidas para dichas fuentes por los paquetes DSC y OCA. De esta forma, se generan dos conjuntos de datos, el conjunto UNKNOWN y el conjunto UNDEFINED.

2. Una vez se han seleccionado los conjuntos de objetos a procesar, éstos son preprocesados, con el objetivo de prepararlos de cara al proceso de agrupamiento. Primero, se corrige el enrojecimiento de los espectros. Para ello, se obtiene una estimación de los parámetros de extinción mediante el mapa de Drimmel (ver [70]) para, después, corregir el enrojecimiento mediante la curva de extinción de Cardelli (ver [71]). Una vez corregido el enrojecimiento, se unen los espectros BP y RP. Finalmente, se realiza la normalización por flujo integrado.

3. El proceso de agrupamiento se realiza por separado para cada conjunto (UNDEFINED,UNKNOWN). Dicho proceso conlleva diversas dificultades, ya que el hardware disponible en el centro de procesado de datos (CNES) está compuesto por una serie de nodos con un máximo de 2GB de RAM, con lo que el conjunto de datos que procesará OA no puede ser almacenado en memoria dinámica. Además, el entrenamiento del mapa auto-organizativo en operación real llevaría demasiado tiempo, dado que cada ciclo de procesado de APSIS tiene una extensión de 90 días, con lo que es necesario establecer un modelo de computación distribuida. Por lo tanto, se ha diseñado un esquema en el que se subdividen los conjuntos de datos en varios subconjuntos, los cuales son tratables por diferentes instancias del módulo *SOMUpdater*. Cada instancia de *SOMUpdater* se encarga, para cada objeto en su subconjunto de observaciones, de obtener la neurona ganadora mediante la

ecuación 3.6 y de computar las actualizaciones correspondientes de dicha neurona ganadora y su vecindad, mediante la ecuación 3.8. Las actualizaciones de pesos para cada observación se acumulan, de forma que cada *SOMUpdater* obtiene las actualizaciones de pesos correspondientes a su subconjunto.

4. El módulo *SOMMerger* se encarga de unir las asignaciones y las actualizaciones de pesos de obtenidas por cada *SOMUpdater*, para luego actualizar los pesos del mapa. Este módulo también chequea la convergencia del proceso de entrenamiento. En caso de no converger, se vuelve al paso anterior.

5. Finalmente, se han computado dos mapas auto-organizativos independientes. Este módulo se encarga de identificar las neuronas de ambos mapas, mediante los métodos expuestos en la sección 3.5.1. Además, el módulo calcula las distribuciones de los distintos parámetros astrométricos (posiciones, paralajes, etc.) para cada neurona de los mapas, en función de los objetos que han sido asignados a la misma.

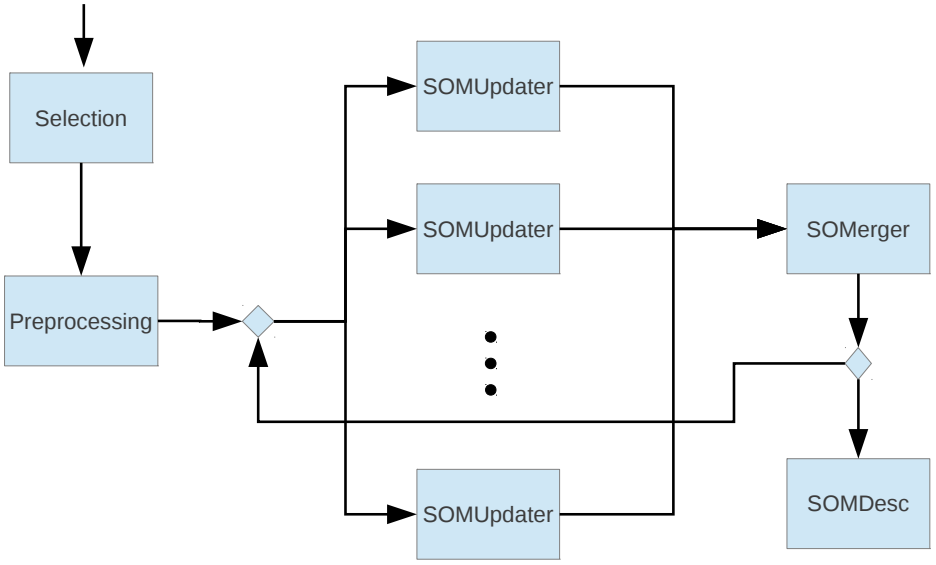


FIGURA 3.28: Diagrama de flujo de datos en el procesado de datos del paquete OA.

Uno de los factores a tener en cuenta es la inicialización de los dos mapas para el procesado en APSIS. En este caso, el mapa inicial utilizado para el primer ciclo de procesado será el obtenido a través del entrenamiento con la simulación descrita en la sección 3.4. Al inicializar los mapas de OA con un mapa del cual conocemos su topología, podemos observar cómo dicha topología ha cambiado y cuán diferentes son

los prototipos obtenidos con datos de la misión. Adicionalmente, la inicialización de cada ciclo de procesado se realizará con el mapa del ciclo anterior, lo cual nos permitirá monitorizar cambios significativos en los datos observados por Gaia, de forma que será posible identificar efemérides, como son las novas y las supernovas.

3.7.2 Implementación en MapReduce

De forma adicional a la implementación en APSIS, se ha realizado una implementación distribuida del algoritmo de entrenamiento de los mapas auto-organizativos sobre *MapReduce*. *MapReduce* es un paradigma de programación escalable que ha ganado gran popularidad en los últimos debido tanto a su capacidad para procesar grandes cantidades de datos como a su simplicidad. Actualmente, la implementación del paradigma más utilizada es la proporcionada por la fundación Apache, llamada *Hadoop*. *Hadoop* incluye un esquema de programación en java para algoritmos *MapReduce* y un sistema de ficheros distribuido llamado HDFS, véase [72]. El algoritmo de entrenamiento de la red SOM, implementado con *Hadoop*, está compuesto por tres componentes:

- El *driver*: Este componente se encarga de inicializar y ejecutar el algoritmo de aprendizaje. Para cada iteración, el *driver* llama a un trabajo *MapReduce* que realiza las fases de competición y cooperación de forma distribuida. Entonces, el *driver* reúne los resultados y completa la adaptación de las neuronas. Finalmente, el *driver* se encarga de chequear la convergencia del entrenamiento.
- El *mapper*: Los *mappers* se encargan de realizar cálculos sobre una única observación. Por lo tanto, se toma ventaja de los mismos para computar la neurona ganadora para cada observación en el conjunto de entrada. Como resultado, el *mapper* devuelve una pareja clave-valor, siendo la clave el número de neurona ganadora y el valor la correspondiente observación.
- El *reducer*: Este componente se usa para obtener las actualizaciones de la red SOM correspondientes a las observaciones que comparten la misma neurona ganadora. Cada *reducer* recibe una lista de observaciones, retornando la contribución correspondiente a la actualización de las neuronas de la red.

Un esquema del algoritmo de entrenamiento de la red SOM, implementado en *MapReduce*, es mostrado en la figura 3.29. Este esquema ha permitido el cómputo de una red SOM de tamaño 30*30 para representar todo el conjunto de espectros proveniente del catálogo Hamburg/ESO (HES), compuesto por aproximadamente $5 * 10^6$ espectros. Dicho cómputo fue realizado en tan sólo 3 días, en el Centro de Supercomputación de

Galicia (CESGA), mediante un clúster Hadoop virtual con 100 nodos, preparado a tal efecto. El análisis de dicho mapa auto-organizativo será realizado en el futuro próximo por expertos en dicho catálogo.

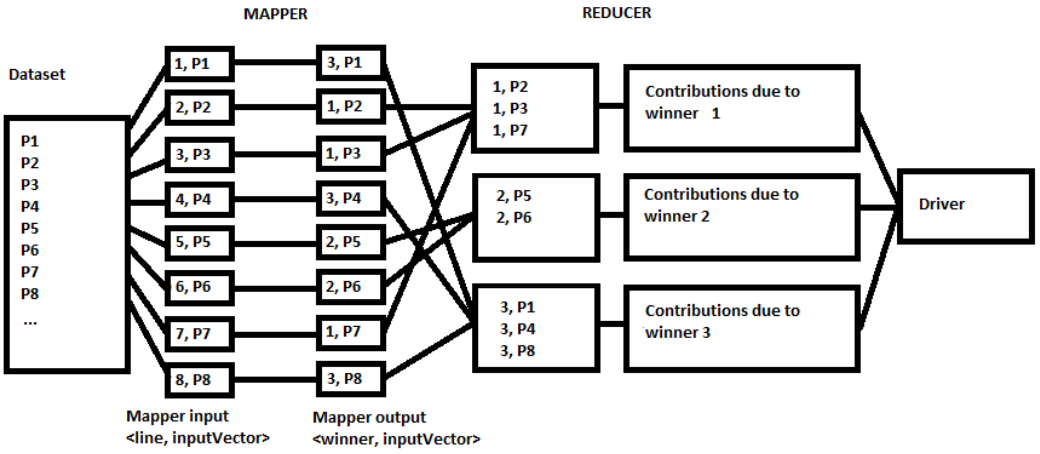


FIGURA 3.29: Esquema de la implementación del algoritmo de entrenamiento de una red SOM mediante el paradigma MapReduce.

3.7.3 Técnicas especiales de aceleración del entrenamiento de la red SOM

A pesar de la implementación en modo distribuido del algoritmo de entrenamiento de la red SOM, el proceso requerirá una gran cantidad de recursos computacionales. Afortunadamente, es posible tomar ventaja de la vecindad de las redes SOM para recortar el tiempo de cómputo, sin necesidad de recurrir a técnicas de selección de características, basándose en las técnicas desarrolladas en [73]. En primer lugar, es posible restringir la búsqueda de la neurona ganadora ($win(s)$) para una observación dada, ya que, cuanto mayor sea la ordenación del mapa, mayor será la probabilidad de que la ganadora sea la misma que en la iteración previa del aprendizaje ($win(s-1)$), o una neurona cercana en el mapa. Este comportamiento se muestra en la figura 3.30, donde puede observarse que, en la gran mayoría de las ocasiones, la distancia media entre $win(t)$ y $win(t-1)$ disminuye con las iteraciones, en consonancia con la mejor ordenación del mapa. Por otro lado, es posible acelerar el proceso de aprendizaje mediante el aumento progresivo del tamaño del mapa. Para ello, en primer lugar se entrena un mapa de pequeño tamaño, para luego interpolarlo y entrenar un mapa considerablemente mayor, el cual puede realizarse rápidamente utilizando la restricción en la búsqueda del ganador. Además de acelerar el entrenamiento, este proceso permite también crear una red SOM multiresolución, algo deseable de cara al posterior análisis, ya que se puede seleccionar el tamaño de mapa deseado para realizar la exploración de los datos.

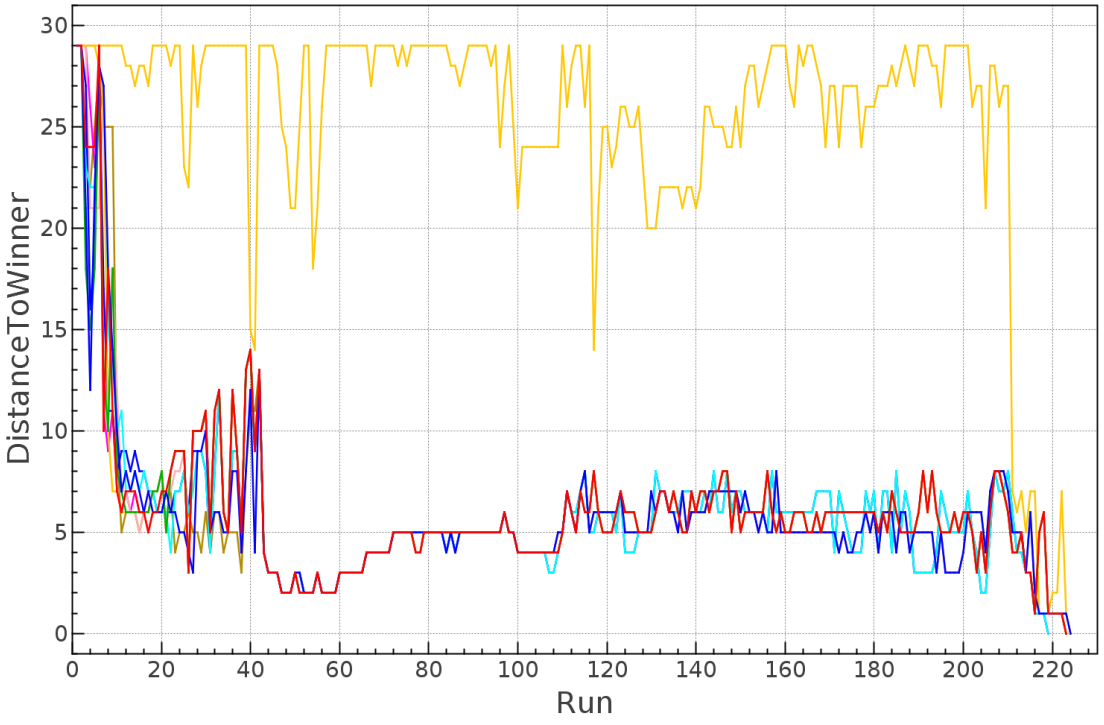


FIGURA 3.30: Distancia media por iteración entre la neurona ganadora actual $win(s)$ y la ganadora de la iteración previa $win(s-1)$.

Apéndice A

Aplicaciones en otros dominios

En este apéndice se describen dos casos en los que se han aplicado, directa o indirectamente, técnicas que provienen del trabajo desarrollado en esta tesis. El primero de ellos también tiene una conexión con la ciencia espacial, ya que se trata la segmentación de imágenes de un radar equipado en un satélite llamado Envisat, de la ESA, con el objetivo de segmentar las mismas, obteniendo zonas candidatas a ser vertidos de fuel. En el segundo caso se aplican redes SOM para la creación de un sistema de detección de intrusos innovador, basado en anomalías. Estos dos casos demuestran la versatilidad de las técnicas y métodos de análisis de datos, las cuales encuentran un número creciente de aplicaciones.

A.1 Detección de vertidos de fuel mediante imágenes SAR

Junto con la deforestación y el cambio climático, la contaminación marina es uno de los principales problemas medioambientales de nuestra época. Sin embargo, en la mayoría de los países, la población, los gobiernos y las organizaciones internacionales menosprecian el grave daño ambiental que los océanos están sufriendo y las catastróficas consecuencias que esto puede acarrear en el futuro. Hoy podemos observar la contaminación del océano en forma de manchas masivas de plástico (la mancha del Pacífico, por ejemplo, tiene actualmente dos veces el tamaño de los Estados Unidos). Es también el resultado de accidentes de petroleros, como el hundimiento en 2002 del Prestige [74], en las costas de Galicia (España), que causó un tremendo daño ambiental. Por último, muchos barcos vacían la sentina de sus tanques en el mar para reducir los costos, causando de esta manera aún más contaminación que a través de accidentes ocasionales.

La vigilancia del océano se ha realizado tradicionalmente mediante aeronaves y fuerzas de la guardia costera. Sin embargo, a la hora de cubrir grandes superficies del océano,

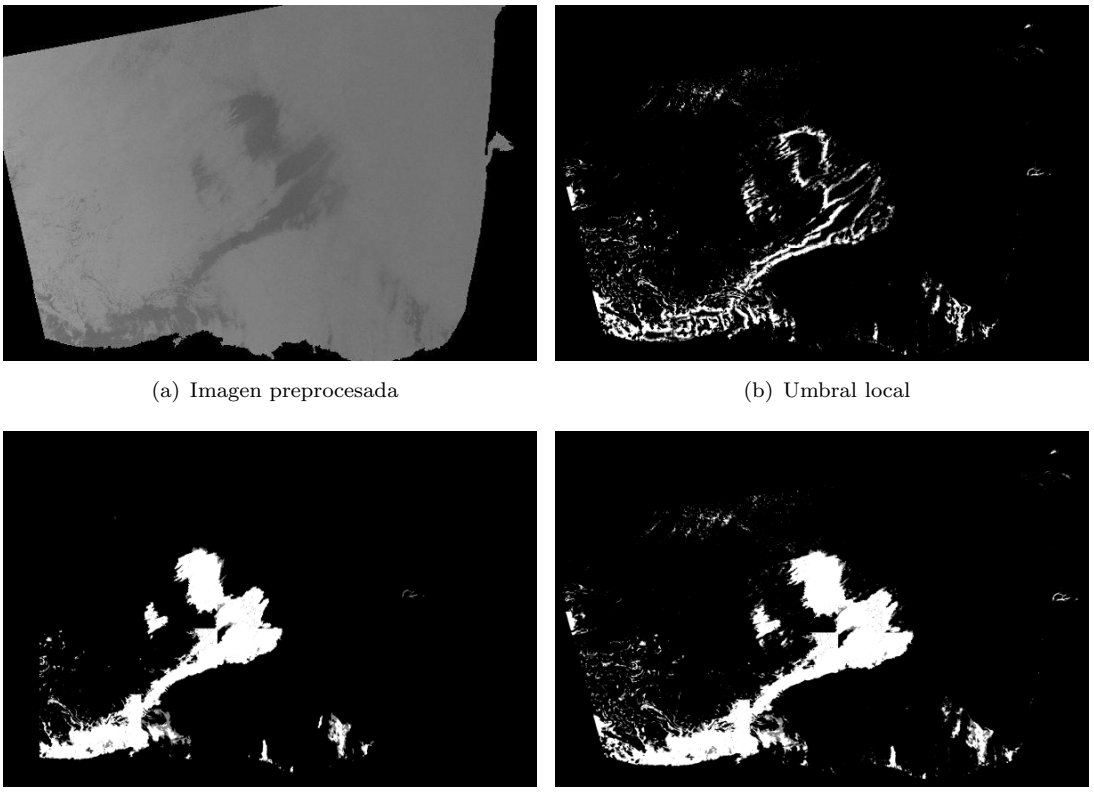
las naves espaciales proporcionan una mejor solución. Con respecto al tipo de sensores utilizados en la vigilancia del océano, las microondas son preferibles a los sensores ópticos, ya que pueden proporcionar datos bajo todas las condiciones meteorológicas y de luminosidad. Este tipo de sensores son llamados radares de apertura sintética (SAR, por sus siglas en inglés), que capturan la dispersión de las microondas en una determinada superficie y son óptimas para la captura de la rugosidad de la superficie del mar, independientemente de las condiciones meteorológicas. Esta dispersión se puede utilizar para distinguir entre una superficie normal del mar, que mostrará valores altos de dispersión, y entidades "anómalas", como los buques, las formaciones de algas o las descargas de contaminantes, que tienen valores de dispersión inferiores.

Nuestro equipo de investigación ha desarrollado una herramienta para monitorizar las costas mediante la búsqueda de vertidos, que recibe el nombre de "Sentinazos". La herramienta se ejecuta en la Nube, proporcionando un marco integrado, donde los usuarios pueden compartir información y métodos. Se permite preprocesar las imágenes para facilitar su análisis posterior. Para realizar dicho análisis, los usuarios pueden subir sus propios algoritmos, de forma que los mismos pasan a ser escalables. Además, se proporcionan una variedad de métodos para la validación de los algoritmos. Por último, todas las imágenes y objetos están geográficamente referenciados, lo que permite el análisis de las relaciones espaciales entre los mismos.

El preprocesado de imágenes SAR se divide en varias fases. Primero se corrigen los cambios de intensidad producidos por la variación en el ángulo de incidencia del radar, para luego realizar un filtrado del ruido de tipo Speckle que afecta a las imágenes. Finalmente, se realiza una proyección cartográfica que permita geolocalizar los píxeles de la imagen y se eliminan las zonas costeras, ya que las mismas no son de interés.

El objetivo principal de la detección de vertidos de fuel por medio de imágenes SAR consiste en aislar esas manchas oscuras que son candidatas a ser vertidos y entonces determinar si son vertidos reales o falsos positivos. Los falsos positivos pueden ser producidos por fenómenos naturales, como las algas, zonas con viento suave, zonas de lluvia, etc.

Con el objetivo de aislar las manchas oscuras en las imágenes, se han desarrollado varias técnicas de segmentación basadas tanto en algoritmos de agrupamiento como en transformaciones de señal y en algoritmos basados en umbrales sobre la intensidad de los píxeles. Los algoritmos de agrupamiento utilizados son el clásico k-means y sus extensiones fuzzy k-means (FCM), kernelized fuzzy k-means (KFCM) y spatial kernelized fuzzy k-means (SKFCM). Este último toma ventaja de las relaciones espaciales entre los píxeles para mejorar la robustez al ruido, véase [75]. A través de la aplicación de estos algoritmos, se ha determinado que los mejores resultados se obtienen al



(a) Imagen preprocesada (b) Umbral local

(c) SKFCM (d) SKFCM+Umbral local

FIGURA A.1: Segmentación de una imagen SAR mediante varias técnicas incluidas en Sentinazos.

combinar el SKFCM con un algoritmo de umbrales locales, ya que los umbrales locales ofrecen buenas segmentaciones de manchas pequeñas, mientras que el SKFCM ofrece los mejores resultados a la hora de segmentar manchas grandes. La figura A.1 ilustra el funcionamiento de este algoritmo de segmentación combinado.

Para proporcionar toda la funcionalidad involucrada en Sentinazos, se ha diseñado una arquitectura basada en herramientas de software libre, mostrada en la figura A.2. Dicha arquitectura está formada por una base de datos relacional con extensiones geográficas (PostGIS), un servidor de funcionalidades geográficas (Geoserver), un sistema de computación distribuida para cálculo intensivo (Hadoop) y un servidor de aplicaciones Web (Tomcat). Más información sobre Sentinazos puede encontrarse en [76].

A.2 Sistema de detección de intrusiones y monitorización de tráfico de red

Los sistemas de detección de intrusiones (IDS, por sus siglas en inglés) son un componente fundamental en el modelo de seguridad de una organización. Su uso, en

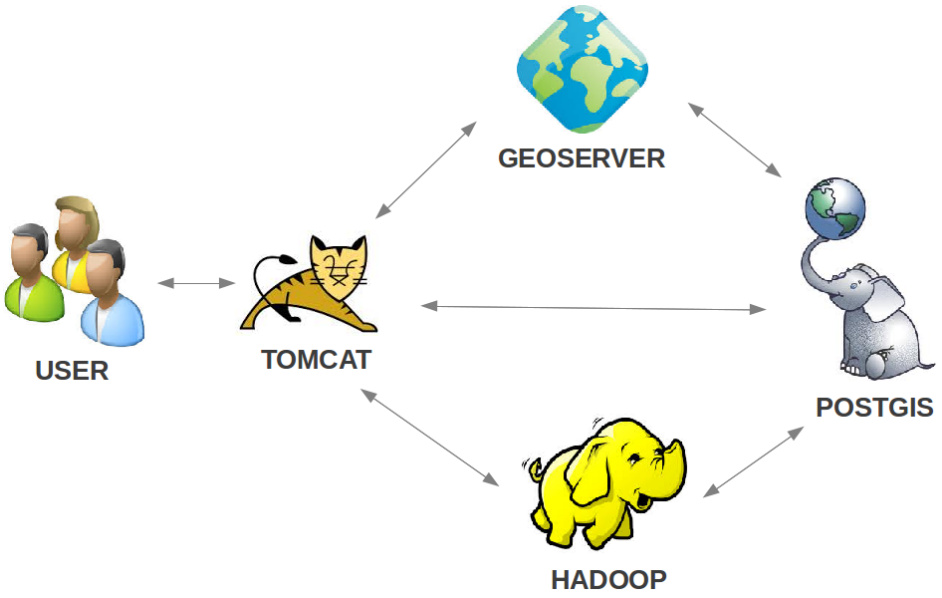
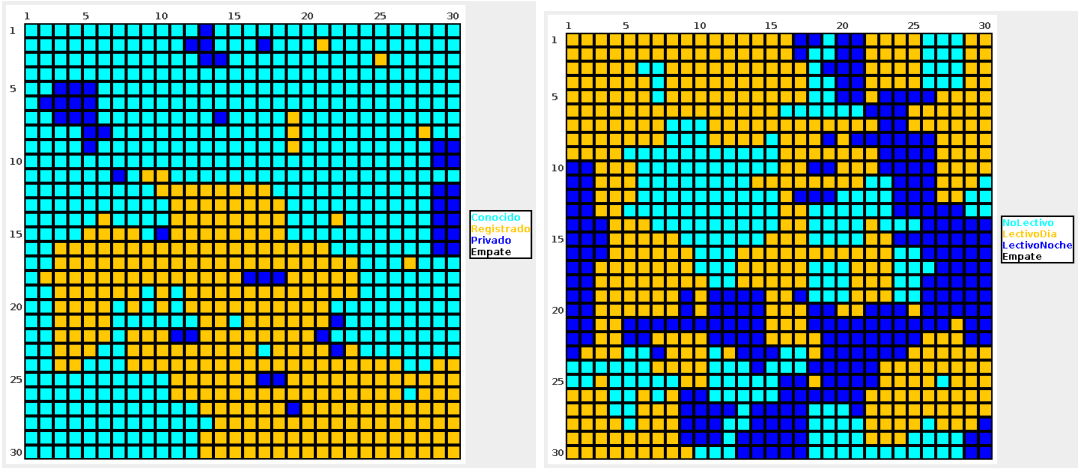


FIGURA A.2: Arquitectura general de Sentinazos.

conjunción con un firewall, forma una doble línea de defensa frente a los numerosos y cada vez más sofisticados ataques a los que nos encontramos expuestos al utilizar las redes informáticas. Sin embargo, no son la panacea. Al estar basados en firmas no disponen de mecanismos para detectar ataques no conocidos, ni modificaciones de ataques conocidos por lo que requieren un ajuste y una supervisión continua por parte de personal cualificado. Para solucionar estos problemas, se ha desarrollado un nuevo modelo de IDS, basado en SOMs. El hecho de que las redes SOM aprendan de forma no supervisada y su capacidad de preservar la topología del espacio de entrada, la hacen idónea para la creación de un IDS basado en la detección de casos atípicos.

Para la creación del IDS basado en SOMs, en primer lugar ha sido necesario un proceso de selección de características mediante el que se han extraído las propiedades más importantes de entre la información proporcionada por cada línea de log del firewall de la red de docencia de la Facultad de Informática de la Coruña. De esta forma, se obtiene un conjunto de características mixto, con características numéricas, como son una serie de factores de repetición de paquetes de red con valores en el intervalo [0,1], y con características categóricas como son el protocolo de red (TCP, UDP) o el tipo de fecha (Lectivo, No Lectivo, Lectivo Noche). De esta forma, el algoritmo de entrenamiento habitual para una red SOM, pensado para características numéricas, deja de ser válido, por lo que se ha tenido que acudir a técnicas especiales como el algoritmo Numeric-Categorical SOM (NCSOM) [77]. Debido a deficiencias encontradas en el algoritmo NCSOM, se ha desarrollado una variante nueva de las redes SOM, denominada Frequency neuron Mixed SOM (FMSOM).



(a) Tipo de puerto (b) Tipo de fecha

FIGURA A.3: Ejemplos de planos de componentes obtenidos por el IDS basado en SOMs.

Además de servir como IDS, la herramienta desarrollada también funciona a modo de sistema de monitorización del tráfico de una red corporativa, mediante visualizaciones como las descritas en la sección 3.4.4.1, véase la figura A.3. En el futuro, está planeada la integración del IDS desarrollado en empresas del sector de las telecomunicaciones, así como la publicación de los correspondientes resultados científicos en revistas internacionales.

Conclusiones

En esta tesis doctoral, se han desarrollado varios métodos de extracción de conocimiento en bases de datos astronómicas, de cara a su aplicación en el análisis de datos de la próxima misión Gaia. Los métodos revisados pertenecen primariamente a la categoría de redes neuronales (ANNs), las cuales se han aplicado a la resolución de dos problemas diferentes: la estimación de parámetros astrofísicos y el análisis de objetos atípicos.

En primer lugar, se ha estudiado el problema de la estimación de parámetros astrofísicos estelares, realizada por el paquete GSP-Spec, considerando ANNs como técnica base. Se ha seguido una metodología formal con el objetivo de obtener un conjunto de parámetros óptimos para el entrenamiento de las ANNs, basada en algoritmos de optimización genéticos (PSO). Además, se han aplicado técnicas de procesado de señal (wavelets) para mejorar las estimaciones de las ANNs, con el resultado de un conjunto de APs más robustos, tanto en el caso de estrellas con tipos espectrales F, G y K débiles, como en el caso de estrellas de tipo A de cualquier magnitud, tal y como se muestra en la figuras 2.9 y 2.10. Por último, se han aplicado algoritmos genéticos clásicos para la selección de características en el espectro RVS. Dicha selección se ha mostrado efectiva, ya que conserva los resultados de estimación, si bien el método requiere de un gran esfuerzo computacional.

Las ANNs clásicas, utilizadas para resolver problemas de regresión, tienen la desventaja de ofrecer resultados poco transparentes. El explotador de los resultados no tiene una medida de incertidumbre sobre los mismos y no conoce la bondad del ajuste realizado. Una posible solución a dicho problema es invertir la ANN y predecir el espectro RVS a partir de un conjunto de APs. De esta forma, es posible obtener una medida de bondad del ajuste y una estimación de la incertidumbre en los APs. Se ha desarrollado y descrito una nueva arquitectura, que hemos denominado GANN, que ha demostrado ser efectiva en la estimación de APs, sobretodo en el caso de estrellas débiles, donde supera a la versión común. Este resultado ha mejorado nuestras expectativas, ya que se esperaban resultados ligeramente inferiores, o bien similares, a los obtenidos por las ANNs tradicionales, véanse la figuras 2.13 y 2.14. Como desventaja, este método conlleva una mayor cantidad de cómputo. El uso de distintas distribuciones a priori sobre los APs, que pueden incluir información externa como el diagrama HR o astrometría, promete ofrecer un conjunto de APs más robusto, y será explorado cuando se disponga de los datos de la misión.

El otro problema que se ha estudiado en esta tesis es el análisis de objetos atípicos de clasificación, realizado por el paquete OA con el objetivo de ayudar en la calibración

de los paquetes de clasificación de Gaia y de permitir el descubrimiento de objetos novedosos. Para ello, se ha asumido que los métodos a utilizar a tal efecto deben ser no supervisados, ya que, por definición, se trata de un conjunto de objetos que no sigue ningún tipo de modelo o regla establecida. Además, se ha considerado que, entre los datos provenientes de los instrumentos de Gaia, la espectrofotometría BP/RP proporciona la información más completa de cara a la clasificación de objetos astrofísicos.

Se ha realizado un estudio del rendimiento que ofrecen diversas técnicas automáticas de clasificación no supervisada, cuando éstas se aplican a espectros BP/RP simulados de la misión Gaia. No obstante, poco después de comenzar tal estudio, ha quedado patente el impacto que tiene la representación de los datos en las técnicas no supervisadas. Por lo tanto, es necesario preprocesar correctamente los datos, si se quiere extraer el máximo conocimiento de los mismos, prestando atención a la robustez que la representación de datos tiene ante la presencia de ruido. Después de un estudio formal de rendimiento, se han escogido los mapas auto-organizativos (SOM) como técnica base para el análisis realizado por OA, ya que éstos ofrecen un buen rendimiento de agrupamiento, véase tabla 3.2, junto con la posibilidad de visualizar el conjunto de datos en un mapa bidimensional, véase la figura 3.11. Dicho poder de visualización se ha comprobado con simulaciones de Gaia, ya que distintas clases de objetos se agrupan en diferentes regiones del mapa. Por último, se han desarrollado dos técnicas de agrupamiento conjunto que permiten obtener grupos más homogéneos mediante la combinación de particiones computadas con diferentes representaciones del espectro BP/RP. Sin embargo, en el proceso se pierde un gran número de objetos, como se muestra en las tablas 3.5 y 3.6. Además, la complejidad computacional del agrupamiento conjunto es significativamente mayor. En el futuro, y en función del avance en los recursos computacionales, serán estudiadas otras técnicas de agrupamiento y visualización de entre las muchas existentes.

Una vez que las redes SOM han sido correctamente configuradas mediante las simulaciones de Gaia, éstas han sido utilizadas para el análisis de objetos atípicos reales, provenientes del catálogo SDSS. Se ha descrito cómo, mediante las visualizaciones que proporciona la red SOM, es posible separar observaciones astrofísicas de artefactos observacionales. Además, se han caracterizado los grupos del mapa mediante dos técnicas de identificación diferentes. La primera de ellas se basa en la búsqueda de espectros similares a los prototipos de cada neurona en la red, mientras que la segunda consiste en recuperar clasificaciones de la base de datos Simbad para los objetos atípicos de SDSS, mediante el emparejamiento de las respectivas posiciones en la bóveda celeste. De esta forma, se ha podido determinar la presencia de varias clases de objetos astronómicos entre las observaciones sin clasificación previa en SDSS, como enanas blancas y quásares, tal y como se muestra en las figuras 3.24 y 3.26. El estudio de la mezcla de poblaciones de objetos, representadas en el mapa, permite la selección de

objetos sin clasificar, candidatos a convertirse en nuevos descubrimientos. Se espera que, a través de observaciones complementarias, muchos objetos de SDSS sin clasificar sean finalmente identificados, véase la matriz de confusión en la tabla 3.9. El mismo proceso será realizado cuando los datos de Gaia estén disponibles, lo cual conllevará nuevos retos, pero por otro lado promete reportar un número aún mayor de descubrimientos.

Conclusions

In this thesis, we have developed several methods of knowledge extraction in astronomical databases, with the aim of applying them in the analysis of data obtained during Gaia mission. The revised methods belong primarily to the category of ANNs, which have been applied to solve two different problems: astrophysical parameter estimation and outlier analysis.

We started by studying the problem of estimating stellar astrophysical parameters by the GSP-Spec package, considering ANNs as technical basis. A formal methodology was followed in order to obtain a set of optimal parameters for training the neural networks, using genetic-based optimization algorithms (PSO). Furthermore, signal processing techniques (wavelets) have been applied to improve the ANNs performance, with the result of a set of more robust AP estimates, both in the case of weak stars with spectral types F, G and K, as in the case of A stars of any magnitude, as shown in figures 2.9 and 2.10. In this way, mission requirements are met for stars with magnitude $G_{rvs} < 13$ in T_{eff} and $[Fe/H]$ estimation and $G_{rvs} < 12$ in the case of $log\, g$ estimation for F, G and K stars. For A type stars, requirements are met with $G_{rvs} < 11$ in T_{eff} and $[Fe/H]$ estimation and $G_{rvs} < 14$ in $log\, g$ estimation. Finally, we have applied classical genetic algorithms for feature selection in the RVS spectrum. Such selection has proven to be effective in retaining the AP estimation results, although the method requires a large computational effort.

Classical ANNs, used to solve regression problems, have the disadvantage of providing little transparent results. The results exploiter does not have a measure of uncertainty about them and does not know the goodness of fit. One possible solution to this problem is to reverse the ANN and predict the RVS spectrum from a set of APs. It thus becomes possible to obtain a measure of goodness-of-fit and an estimate of the uncertainty in the APs. We developed and described a new architecture, called GANN, which, besides providing uncertainty, has proven to be effective in estimating APs, especially in the case of faint stars, where it surpasses the common version. This result has exceeded our expectations since we expected slightly inferior or similar results to those obtained by traditional ANNs, see figures 2.13 and 2.14. With GANNs, mission requirements are met up to $G_{rvs} < 13$ in $log\, g$ estimation for F, G and K stars, whereas in the case of A type stars, requirements are met up to $G_{rvs} < 15$ in $log\, g$ estimation and up to $G_{rvs} < 13$ in $[Fe/H]$ estimation. As a disadvantage, the method involves a greater amount of computation. Using different priors on APs, based on the HR diagram and/or in external information (astrometry), promises a more robust set of APs, and will be explored when the data is available from the mission.

The other problem that was studied in this thesis is the analysis of classification outliers, performed by the OA package in order to assist in the calibration of Gaia classification packages and to enable the discovery of novel objects. It has been assumed that the methods used to that effect should be unsupervised because, by definition, it is an object set that do not follow any set pattern or rule. Furthermore, it has been considered that, among the data from the Gaia instruments, BP/RP spectrophotometry provides the most complete information with regard to the classification of astrophysical objects.

We studied the performance of various unsupervised automatic classification techniques when they are applied to BP/RP simulated spectra for the Gaia mission. However, shortly after initiating such a study, it became clear that the impact of data representation in the unsupervised techniques is crucial. Therefore, it is necessary to preprocess the data correctly, if it is wanted to extract the maximum knowledge from it, paying attention to the robustness of the data representation in the presence of noise. After a formal study of performance, we have chosen the self-organizing maps (SOM) as the technical basis for the OA analysis, since these offer good clustering performance, see table 3.2, together with the possibility of viewing the dataset in a two-dimensional map, see figure 3.11. This visualization power was demonstrated with Gaia simulations since different kinds of objects are clustered in different map regions. Finally, two ensemble clustering methods have been developed, which can obtain more homogeneous clusters. However, many objects are lost in the process, as is shown in tables 3.5 and 3.6. Moreover, the computational complexity of the ensemble clustering methods is significantly higher. In the future, an according to the growth of computational resources, more clustering and visualization methods will be studied.

Once the SOM networks were correctly configured by Gaia simulations, they were used for the analysis of real outliers from the SDSS catalog. It has been described how, using the visualizations provided by the SOM network, it is possible to separate 30% of objects as being astrophysical observations from 60% of objects identified as observational artifacts. Furthermore, SOM clusters were characterized using two different identification techniques. The first technique is based on finding spectra similar to the prototypes of each neuron in the network, while the second consists in retrieving classifications from the Simbad database for SDSS outliers, by matching the respective positions in the sky. Thus, we have determined the presence various classes of astronomical objects among the observations without SDSS classification, such as white dwarfs and quasars, as shown in figures 3.24 and 3.26. The study of mixed populations of objects, represented on the map, allows selection of unidentified objects, which are candidates for new types of astrophysical sources. It is hoped that, through complementary observations, many objects that are unclassified in SDSS will be finally identified, see the confusion matrix in table 3.9. The same process will be performed

when Gaia data is available, which will bring new challenges, but on the other hand promises to report a greater number of discoveries.

Bibliografía

[1] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, and et al. et al. The Seventh Data Release of the Sloan Digital Sky Survey. *apjs*, 182: 543, June 2009. doi: 10.1088/0067-0049.

[2] K. Freeman and J. Bland-Hawthorn. The New Galaxy: Signatures of Its Formation. *Annual Revision Astronomy & Astrophysics*, 40:487–537, 2002. doi: 10.1146/annurev.astro.40.060401.093840.

[3] K. C. Freeman and G. McNamara. *In Search of Dark Matter*. 2006. doi: 10.1007/0-387-27618-1.

[4] Jia Li, Surajit Ray, and Bruce G. Lindsay. A Nonparametric Statistical Approach to Clustering via Mode Identification. *J. Mach. Learn. Res.*, 8:1687–1723, December 2007. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1314498.1314555`.

[5] C. A. L. Bailer-Jones, R. Andrae, B. Arcay, T. Astraatmadja, I. Bellas-Velidis, A. Berihuete, A. Bijaoui, C. Carrión, C. Dafonte, Y. Damerdji, A. Dapergolas, P. de Laverny, L. Delchambre, P. Drazinos, R. Drimmel, Y. Frémat, D. Fustes, M. García-Torres, C. Guédé, U. Heiter, A.-M. Janotto, A. Karampelas, D.-W. Kim, J. Knude, I. Kolka, E. Kontizas, M. Kontizas, A. J. Korn, A. C. Lanzafame, Y. Lebreton, H. Lindstrøm, C. Liu, E. Livanou, A. Lobel, M. Manteiga, C. Martayan, C. Ordenovic, B. Pichon, A. Recio-Blanco, B. Rocca-Volmerange, L. M. Sarro, K. Smith, R. Sordo, C. Soubiran, J. Surdej, F. Thévenin, P. Tsalmantza, A. Vallenari, and J. Zorec. The Gaia astrophysical parameters inference system (Apsis). Pre-launch description. *ArXiv e-prints*, September 2013.

[6] T. Schmidt-Kaler. Automated spectral classification. A survey. *Bulletin d'Information du Centre de Donnees Stellaires*, 23:2, October 1982.

[7] T. Schmidt-Kaler. Quantitative two-dimensional classifications of low dispersion objective prism spectra. In A. G. D. Philip and D. S. Hayes, editors, *The HR*

Diagram - The 100th Anniversary of Henry Norris Russell, volume 80 of *IAU Symposium*, pages 29–32, 1978.

[8] M. J. Kurtz. Progress in Automation Techniques for MK Classification. In R. F. Garrison, editor, *The MK Process and Stellar Classification*, pages 136–152, 1984.

[9] T. von Hippel, L. J. Storrie-Lombardi, M. C. Storrie-Lombardi, and M. J. Irwin. Automated Classification of Stellar Spectra - Part One - Initial Results with Artificial Neural Networks. *mnras*, 269:97, July 1994.

[10] W. B. Weaver and A. V. Torres-Dodgen. Neural Network Classification of the Near-Infrared Spectra of A-Type Stars. *apj*, 446:300, June 1995. doi: 10.1086/175789.

[11] M. C. Storrie-Lombardi, O. Lahav, Jr. L. Sodre, and L. J. Storrie-Lombardi. Morphological Classification of Galaxies by Artificial Neural Networks. *mnras*, 259: 8P, November 1992.

[12] R. A. Calvo, H. A. Ceccato, and R. D. Piacentini. Neural network prediction of solar activity. *apj*, 444:916–921, May 1995. doi: 10.1086/175661.

[13] C. A. L. Bailer-Jones, M. Irwin, G. Gilmore, and T. von Hippel. Physical parametrization of stellar spectra - The neural network approach. *mnras*, 292: 157, November 1997.

[14] S. Snider, C. Allende Prieto, T. von Hippel, T. C. Beers, C. Sneden, Y. Qu, and S. Rossi. Three-dimensional Spectral Classification of Low-Metallicity Stars Using Artificial Neural Networks. *apj*, 562:528–548, November 2001. doi: 10.1086/323428.

[15] M. Manteiga, D. Ordóñez, C. Dafonte, and B. Arcay. ANNs and Wavelets: A Strategy for Gaia RVS Low S/N Stellar Spectra Parameterization. *pasp*, 122: 608–617, May 2010. doi: 10.1086/653039.

[16] C. A. L. Bailer-Jones. The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry. *mnras*, 403:96–116, March 2010. doi: 10.1111/j.1365-2966.2009.16125.x.

[17] C. Liu, C. A. L. Bailer-Jones, R. Sordo, A. Vallenari, R. Borrachero, X. Luri, and P. Sartoretti. The expected performance of stellar parametrization with Gaia spectrophotometry. *mnras*, 426:2463–2482, November 2012. doi: 10.1111/j.1365-2966.2012.21797.x.

[18] G. Kordopatis, A. Recio-Blanco, P. de Laverny, A. Bijaoui, V. Hill, G. Gilmore, R. F. G. Wyse, and C. Ordenovic. Automatic stellar spectra parameterisation in the IR Ca ii triplet region. *aap*, 535:A106, November 2011. doi: 10.1051/0004-6361.

[19] A. Recio-Blanco, A. Bijaoui, and P. de Laverny. Automated derivation of stellar atmospheric parameters and chemical abundances: the MATISSE algorithm. *mnras*, 370:141–150, July 2006. doi: 10.1111/j.1365-2966.2006.10455.x.

[20] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *Computer Journal*, 7:308–313, 1965.

[21] R. L. Kurucz. Model atmospheres for G, F, A, B, and O stars. *Astrophysical Journal Supplement Series*, 40:1–340, May 1979. doi: 10.1086/190589.

[22] Roberto Battiti. First- and second-order methods for learning: between steepest descent and Newton's method. *Neural Comput.*, 4(2):141–166, March 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.2.141. URL `http://dx.doi.org/10.1162/neco.1992.4.2.141`.

[23] Sandhya Samarasinghe. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition.* Auerbach Publications, 2006. URL `http://www.bibsonomy.org/bibtex/2bedbf1c1af49e0b5d6276626db7ef6d2/lran022`.

[24] Russell C. Eberhart, Yuhui Shi, and James Kennedy. *Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation).* Morgan Kaufmann, 1 edition, April 2001. ISBN 1558605959. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1558605959`.

[25] Ioan Cristian Trelea. The particle swarm optimization algorithm: convergence analysis and parameter selection. *Inf. Process. Lett.*, 85(6):317–325, March 2003. ISSN 0020-0190. doi: 10.1016/S0020-0190(02)00447-7. URL `http://dx.doi.org/10.1016/S0020-0190(02)00447-7`.

[26] S.W. Piche. Steepest descent algorithms for neural network controllers and filters. *Neural Networks, IEEE Transactions on*, 5(2):198–212, 1994. ISSN 1045-9227. doi: 10.1109/72.279185.

[27] J. Leonard and M.A. Kramer. Improvement of the backpropagation algorithm for training neural networks. *Computers & Chemical Engineering*, 14(3):337–341, 1990. ISSN 0098-1354. doi: 10.1016/0098-1354(90)87070-6. URL `http://www.sciencedirect.com/science/article/pii/0098135490870706`.

[28] Tao Li, Qi Li, Shenghuo Zhu, and Mitsunori Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explor. Newsl.*, 4(2):49–68, December 2002. ISSN 1931-0145. doi: 10.1145/772862.772870. URL `http://doi.acm.org/10.1145/772862.772870`.

[29] Jean-Luc Starck, Ralf Siebenmorgen, and Roland Gredel. Spectral Analysis Using the Wavelet Transform. *The Astrophysical Journal*, 482(2):1011, 1997. URL `http://stacks.iop.org/0004-637X/482/i=2/a=1011`.

[30] M. Fligge and S. K. Solanki. Noise reduction in astronomical spectra using wavelet packets. *A&AS*, 124:579–587, September 1997. doi: 10.1051/aas:1997208.

[31] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988. ISSN 1097-0312. doi: 10.1002/cpa.3160410705. URL `http://dx.doi.org/10.1002/cpa.3160410705`.

[32] Stephane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[33] Diego Fustes, Diego Ordóñez, Carlos Dafonte, Minia Manteiga, and Bernardino Arcay. Distributed Genetic Algorithm for Feature Selection in Gaia RVS Spectra: Application to ANN Parameterization. In Luis Manuel Sarro, Laurent Eyer, William O'Mullane, and Joris De Ridder, editors, *Astrostatistics and Data Mining*, volume 2 of *Springer Series in Astrostatistics*, pages 127–131. Springer New York. ISBN 978-1-4614-3322-4. doi: 10.1007/978-1-4614-3323-1_12. URL `http://dx.doi.org/10.1007/978-1-4614-3323-1_12`.

[34] David J.C. MacKay. A Practical Bayesian Framework for Backprop Networks. *Neural Computation*, 4:448–472, 1991.

[35] C K I Williams, C Quazaz, C M Bishop, and H Zhu. On The Relationship Between Bayesian Error Bars And The Input Data Density. In *In Proceedings Fourth IEE International Conference on Artificial Neural Networks*, pages 160–165, 1995.

[36] Andreas S. Weigend and David A. Nix. Predictions with Confidence Intervals (Local Error Bars), 1994.

[37] W. A. Wright. Neural network regression with input uncertainty. In *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 284–293, 1998. doi: 10.1109/NNSP.1998.710658.

[38] M. A. C. Perryman, K. S. de Boer, G. Gilmore, E. Høg, M. G. Lattanzi, L. Lindegren, X. Luri, F. Mignard, O. Pace, and P. T. de Zeeuw. GAIA: Composition, formation and evolution of the Galaxy. *Astronomy&Astrophysics*, 369:339–363, April 2001. doi: 10.1051/0004-6361:20010085.

[39] Y. Isasi, F. Figueras, X. Luri, and A. C. Robin. GUMS & GOG: Simulating the Universe for Gaia. In Jose M. Diego, Luis J. Goicoechea, J. Ignacio González-Serrano, and Javier Gorgas, editors, *Highlights of Spanish Astrophysics V*, Astrophysics and Space Science Proceedings. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11250-8. doi: 10.1007/978-3-642-11250-8_106.

[40] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:2004, 2004.

[41] C.M. Bishop. Novelty Detection and Neural Network Validation. In Stan Gielen and Bert Kappen, editors, *ICANN '93*, pages 789–794. Springer London. ISBN 978-3-540-19839-0. doi: 10.1007/978-1-4471-2063-6_225. URL `http://dx.doi.org/10.1007/978-1-4471-2063-6_225`.

[42] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, 13(7):1443–1471, July 2001. ISSN 0899-7667. doi: 10.1162/089976601750264965. URL `http://dx.doi.org/10.1162/089976601750264965`.

[43] Burr Settles. Active learning literature survey. Technical report, 2010.

[44] Rui Xu and Don Wunsch. *Clustering*. Wiley-IEEE Press, 2009. ISBN 9780470276808.

[45] J. Sánchez Almeida, J. A. L. Aguerri, C. Muñoz-Tuñón, and A. de Vicente. Automatic Unsupervised Classification of All Sloan Digital Sky Survey Data Release 7 Galaxy Spectra. *The Astrophysical Journal*, 714(1):487, 2010. URL `http://stacks.iop.org/0004-637X/714/i=1/a=487`.

[46] J. Sánchez Almeida and C. Allende Prieto. Automated Unsupervised Classification of the Sloan Digital Sky Survey Stellar Spectra using k-means Clustering. *The Astrophysical Journal*, 763:50, January 2013. doi: 10.1088/0004-637X.

[47] J. Vanderplas and A. Connolly. Reducing the Dimensionality of Data: Locally Linear Embedding of Sloan Galaxy Spectra. *The Astrophysical Journal*, 138: 1365–1379, November 2009. doi: 10.1088/0004-6256.

[48] Diego Ordóñez, Carlos Dafonte, Bernardino Arcay Varela, and Minia Manteiga. HSC: A multi-resolution clustering strategy in Self-Organizing Maps applied to

astronomical observations. *Appl. Soft Comput.*, 12(1):204–215, 2012. URL `http://dx.doi.org/10.1016/j.asoc.2011.08.052`.

[49] Diego Fustes, Carlos Dafonte, Bernardino Arcay, Minia Manteiga, Kester Smith, Antonella Vallenari, and Xavier Luri. SOM ensemble for unsupervised outlier analysis. Application to outlier identification in the Gaia astronomical survey. *Expert Syst. Appl.*, 40(5):1530–1541, April 2013. ISSN 0957-4174. doi: 10.1016/j.eswa.2012.08.069. URL `http://dx.doi.org/10.1016/j.eswa.2012.08.069`.

[50] D. Fustes, M. Manteiga, C. Dafonte, B. Arcay, A. Ulla, K. Smith, R. Borrachero, and R. Sordo. An approach to the analysis of SDSS spectroscopic outliers based on Self-Organizing Maps. *ArXiv e-prints*, September 2013.

[51] I. Brott and P. H. Hauschildt. A PHOENIX Model Atmosphere Grid for Gaia. In C. Turon, K. S. O'Flaherty, and M. A. C. Perryman, editors, *The Three-Dimensional Universe with Gaia*, volume 576 of *ESA Special Publication*, page 565, January 2005.

[52] B. Gustafsson, B. Edvardsson, K. Eriksson, U. G. Jørgensen, Å. Nordlund, and B. Plez. A grid of MARCS model atmospheres for late-type stars. I. Methods and general properties. *Astronomy & Astrophysics*, 486:951–970, August 2008. doi: 10.1051/0004-6361:200809724.

[53] R. Sordo, A. Vallenari, R. Tantalo, C. Liu, K. Smith, F. Allard, R. Blomme, J.-C. Bouret, I. Brott, P. de Laverny, B. Edvardsson, Y. Frémat, U. Heber, E. Josselin, O. Kochukhov, A. Korn, A. Lanzafame, C. Martayan, F. Martins, B. Plez, A. Schweitzer, F. Thévenin, and J. Zorec. Stellar libraries for Gaia. *Journal of Physics Conference Series*, 328(1):012006, December 2011. doi: 10.1088/1742-6596.

[54] F. Allard, P. H. Hauschildt, and A. Schweitzer. Spherically Symmetric Model Atmospheres for Low-Mass Pre-Main-Sequence Stars with Effective Temperatures between 2000 and 6800 K. *The Astrophysical Journal*, 539:366–371, August 2000. doi: 10.1086/309218.

[55] B. G. Castanheira, S. O. Kepler, G. Handler, and D. Koester. Analysis of IUE spectra of helium-rich white dwarf stars. *Astronomy & Astrophysics*, 450:331–337, April 2006. doi: 10.1051/0004-6361:20054221.

[56] R. Blomme, Y. Frémat, A. Lobel, and C. Martayan. Emission-line Stars and Early-type Stars with Gaia. *EAS Publications Series*, 45:373–376, 0 2010. ISSN 1638-1963. doi: 10.1051/eas. URL `http://www.eas-journal.org/article_S1633476045000620`.

[57] P. Tsalmantza and C.A.L. Bailer-Jones. Parametrization of Binary Stars with Gaia Observations. In Luis Manuel Sarro, Laurent Eyer, William O'Mullane, and Joris De Ridder, editors, *Astrostatistics and Data Mining*, volume 2 of *Springer Series in Astrostatistics*, pages 263–270. Springer New York, 2012. ISBN 978-1-4614-3322-4. doi: 10.1007/978-1-4614-3323-1_28. URL `http://dx.doi.org/10.1007/978-1-4614-3323-1_28`.

[58] B. Rocca-Volmerange, P. Tsalmantza, and M. Kontizas. A library of synthetic galaxy spectra for GAIA. In C. Charbonnel, F. Combes, and R. Samadi, editors, *SF2A-2008*, page 33, November 2008.

[59] P. Tsalmantza, A. Karampelas, M. Kontizas, C. A. L. Bailer-Jones, B. Rocca-Volmerange, E. Livanou, I. Bellas-Velidis, E. Kontizas, and A. Vallenari. A semi-empirical library of galaxy spectra for Gaia classification based on SDSS data and PÉGASE models. *Astronomy & Astrophysics*, 537:A42, January 2012. doi: 10.1051/0004-6361.

[60] Q. A. Parker, A. Acker, D. J. Frew, M. Hartley, A. E. J. Peyaud, F. Ochsenbein, S. Phillipps, D. Russeil, S. F. Beaulieu, M. Cohen, J. Köppen, B. Miszalski, D. H. Morgan, R. A. H. Morris, M. J. Pierce, and A. E. Vaughan. The Macquarie/AAO/Strasbourg Hα Planetary Nebula Catalogue: MASH. *MNRAS*, 373:79–94, November 2006. doi: 10.1111/j.1365-2966.2006.10950.x.

[61] C. Fabricius, C. Jordi, J. M. Carrasco, H. Voss, and M. Weiler. Gaia photometric calibration. In *Highlights of Spanish Astrophysics VII*, pages 880–885, May 2013.

[62] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theor.*, 28(2): 129–137, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489. URL `http://dx.doi.org/10.1109/TIT.1982.1056489`.

[63] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL `http://dl.acm.org/citation.cfm?id=1283383.1283494`.

[64] Teuvo Kohonen. Neurocomputing: foundations of research. chapter Self-organized formation of topologically correct feature maps, pages 509–521. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6. URL `http://dl.acm.org/citation.cfm?id=65669.104428`.

[65] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[66] S. Sarumathi, N. Shanthi, and G. Santhiya. Article: A Survey of Cluster Ensemble. *International Journal of Computer Applications*, 65(9):8–11, March 2013. Published by Foundation of Computer Science, New York, USA.

[67] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. J. Connolly, D. J. Eisenstein, J. A. Frieman, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, P. Z. Kunszt, B. C. Lee, A. Meiksin, J. A. Munn, H. J. Newberg, R. C. Nichol, T. Nicinski, J. R. Pier, G. T. Richards, M. W. Richmond, D. J. Schlegel, J. A. Smith, M. A. Strauss, M. SubbaRao, A. S. Szalay, A. R. Thakar, D. L. Tucker, D. E. Vanden Berk, B. Yanny, J. K. Adelman, Jr. J. E. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, M. Bartelmann, S. Bastian, A. Bauer, E. Berman, H. Böhringer, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, L. Carey, M. A. Carr, B. Chen, D. Christian, P. L. Colestock, J. H. Crocker, I. Csabai, P. C. Czarapata, J. Dalcanton, A. F. Davidsen, J. E. Davis, W. Dehnen, S. Dodelson, M. Doi, T. Dombeck, M. Donahue, N. Ellman, B. R. Elms, M. L. Evans, L. Eyer, X. Fan, G. R. Federwitz, S. Friedman, M. Fukugita, R. Gal, B. Gillespie, K. Glazebrook, J. Gray, E. K. Grebel, B. Greenawalt, G. Greene, J. E. Gunn, E. de Haas, Z. Haiman, M. Haldeman, P. B. Hall, M. Hamabe, B. Hansen, F. H. Harris, H. Harris, M. Harvanek, S. L. Hawley, J. J. E. Hayes, T. M. Heckman, A. Helmi, A. Henden, C. J. Hogan, D. W. Hogg, D. J. Holmgren, J. Holtzman, C.-H. Huang, C. Hull, S.-I. Ichikawa, T. Ichikawa, D. E. Johnston, G. Kauffmann, R. S. J. Kim, T. Kimball, E. Kinney, M. Klaene, S. J. Kleinman, A. Klypin, G. R. Knapp, J. Korienek, J. Krolik, R. G. Kron, J. Krzesiński, D. Q. Lamb, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, B. McLean, K. Menou, A. Merelli, H. J. Mo, D. G. Monet, O. Nakamura, V. K. Narayanan, T. Nash, Jr. E. H. Neilsen, P. R. Newman, A. Nitta, M. Odenkirchen, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. S. Peterson, D. Petravick, A. Pope, R. Pordes, M. Postman, A. Prosapio, T. R. Quinn, R. Rechenmacher, C. H. Rivetta, H.-W. Rix, C. M. Rockosi, R. Rosner, K. Ruthmansdorfer, D. Sandford, D. P. Schneider, R. Scranton, M. Sekiguchi, G. Sergey, R. Sheth, K. Shimasaku, S. Smee, S. A. Snedden, A. Stebbins, C. Stubbs, I. Szapudi, P. Szkody, G. P. Szokoly, S. Tabachnik, Z. Tsvetanov, A. Uomoto, M. S. Vogeley, W. Voges, P. Waddell, R. Walterbos, S.-i. Wang, M. Watanabe, D. H. Weinberg, R. L. White, S. D. M. White, B. Wilhite, D. Wolfe, N. Yasuda, D. G. York, I. Zehavi, and W. Zheng. Sloan Digital Sky Survey: Early Data Release. *The Astrophysical Journal*, 123:485–548, January 2002. doi: 10.1086/324741.

[68] SDSS DR7 Object Explorer. URL `http://cas.sdss.org/dr7/en/tools/explore/obj.asp`.

[69] M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier. The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astronomy & Astrophysics Supplement Series*, 143:9–22, April 2000. doi: 10.1051/aas:2000332.

[70] R. Drimmel, A. Cabrera-Lavers, and M. López-Corredoira. A three-dimensional Galactic extinction model. *Astronomy & Astrophysics*, 409:205–215, October 2003. doi: 10.1051/0004-6361:20031070.

[71] J. A. Cardelli, G. C. Clayton, and J. S. Mathis. The relationship between infrared, optical, and ultraviolet extinction. *The Astrophysical Journal*, 345:245–256, October 1989. doi: 10.1086/167900.

[72] Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596521979, 9780596521974.

[73] Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the WEBSOM method. *Inf. Sci.*, (1-3):135–156.

[74] La tragedia del prestige, evolución de la marea negra., 2009. URL `http://www.lavozdegalicia.es/albumes/index.jsp`.

[75] Dao-qiang Zhang and Song-can Chen. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artificial Intelligence in Medicine*, 32:2004, 2004.

[76] D. Fustes, D. Cantorna, C. Dafonte, B. Arcay, A. Iglesias, and M. Manteiga. A cloud-integrated web platform for marine monitoring using GIS and remote sensing. Application to oil spill detection through SAR images. *Future Generation Computer Systems*, September 2013. ISSN ISSN 0167-739X.

[77] Ning Chen and Nuno C. Marques. An extension of self-organizing maps to categorical data. In *Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence*, EPIA'05, pages 304–313, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-30737-0, 978-3-540-30737-2.