Aprendizaje automático para identificar poblaciones estelares en galaxias cercanas

G. Baume[1,2], M.J. Rodriguez[2], C. Feinstein[1,2] & E. Gularte[1,3]

[1] *Facultad de Ciencias Astronómicas y Geofísicas, UNLP, Argentina*
[2] *Instituto de Astrofísica de La Plata, CONICET–UNLP, Argentina*
[3] *Geodesia Espacial y Aeronomía, FCAG–UNLP, Argentina*

Contacto / gbaume@fcaglp.unlp.edu.ar

Resumen / Se ha realizado un estudio de diferentes poblaciones estelares en galaxias cercanas. Éste se ha basado en datos fotométricos multibanda obtenidos con el *Hubble Space Telescope*. En el análisis se han aplicado técnicas de aprendizaje automático no supervisado a fin de reconocer tanto las poblaciones estelares, como los grupos de estrellas en la población más joven. En ambos casos se han utilizado diferentes algoritmos de agrupamiento y se ha evaluado la eficiencia de los mismos. La metodología aplicada ha permitido llevar a cabo la tarea evitando el uso de criterios preconcebidos. Adicionalmente, se ha logrado caracterizar la distribución espacial de cada una de las poblaciones estelares considerando sus similitudes con una estructura de tipo fractal. De esta forma, ha sido posible identificar a las poblaciones mas jóvenes con una estructura jerárquica y a las poblaciones mas evolucionadas con distribuciones homogéneas, salvo fluctuaciones a muy gran escala.

Abstract / A study of different stellar populations in nearby galaxies has been carried out. This has been based on multi-band photometric data obtained with the Hubble Space Telescope. In the analysis, unsupervised machine learning techniques have been applied in order to recognize both the stellar populations and the groups of stars in the youngest population. In both cases, different clustering algorithms have been used and their efficiency has been evaluated. The applied methodology has allowed to carry out the task without the need for preconceived criteria. Additionally, it has been possible to characterize the spatial distribution of each of the stellar populations considering their similarities with a fractal-type structure. In this way, it has been possible to identify the youngest populations with a hierarchical structure and the more evolved populations with homogeneous distributions, except for fluctuations on a very large scale.

Keywords / Methods: data analysis – Galaxies: photometry — Galaxies: star clusters – Galaxies: stellar content

1. Introducción

El volumen de datos producidos por diferentes relevamientos celestes requiere de metodologías y herramientas automatizadas que permitan su análisis sistemático y homogéneo. En este sentido, los diferentes métodos del aprendizaje automático (*machine learning*) proveen un poderoso mecanismo para cumplir con esos objetivos. En particular, los métodos no supervisados permiten ser aplicados a los datos y obtener resultados sin tener conceptos preconcebidos sobre ellos. Uno de los métodos no supervisados más empleados son los de agrupamiento (*clustering*). Esos métodos tienen como objetivo, la identificación de grupos de objetos dentro de los datos sobre la base de características (*features*) similares para los objetos de un dado grupo y disimiles para los objetos de grupos diferentes.

Una de las bases de datos más relevantes, son las diferentes compilaciones de los datos fotométricos de galaxias cercanas del *Hubble Space Telescope* (HST). Estas fuentes de datos tienen la particularidad de brindar información multi-banda de extremadamente alta resolución espacial, permitiendo separar las componentes estelares de las galaxias cercanas (Dalcanton et al. 2009; Lee et al. 2014).

Por otro lado, las galaxias se hallan constituidas por diversas poblaciones estelares. Estas poblaciones permiten investigar la historia de la formación estelar (SFH, *Star Formation History*). Los diagramas fotométricos de cada galaxia son una herramienta relevante para identificar las diferentes poblaciones estelares aunque su identificación y separación es una tarea compleja.

En el presente trabajo, se han aplicado diferentes métodos de agrupamiento sobre los datos fotométricos de las galaxias cercanas NGC 1313 y NGC 2403 (ver Tabla 1). El objetivo ha sido separar las diferentes poblaciones estelares en cada galaxia e identificar los diferentes grupos estelares en las respectivas poblaciones jóvenes.

2. Datos

Se han utilizado datos obtenidos con el *Wide Field Channel* de la *Advanced Camera for Surveys* (WFC/ACS) y con el *Ultraviolet and Visible Light Channel* de la *Wide Field Camera 3* (UVIS/WFC3). Las imágenes utilizadas ya se encuentran pre-reducidas y los datos fotométricos ya se hallan calibrados. Ellos corresponden al Legacy Extragalactic UV Survey (LEGUS; Lee et al. 2014) y al The ACS Nearby Ga-

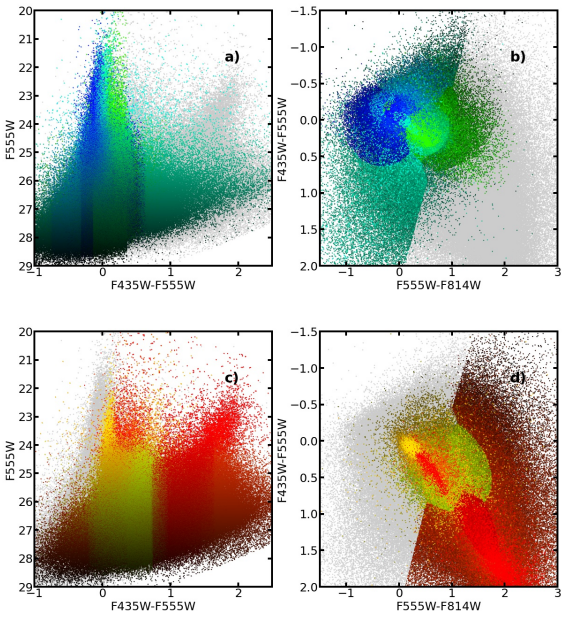

Figura 1: Diagramas fotométricos (CMDs y TCDs) de la galaxia NGC 1313, destacando la población azul (paneles a y b) y la población roja (paneles c y d). Los colores indican las diferentes componentes gaussianas de cada población.

Tabla 1: Características de las galaxias estudiadas.

Parámetro	NGC 1313	NGC 2403	Ref.
Morfologia	SB(s)d	SAB(s)cd	(1)
Distancia	4.2 Mpc	3.1 Mpc	(1)
A_V	0.29	0.11	(1)
Tamaño$^{(*)}$	9.1'× 7.1'	21.9'× 12.3'	(1)
Ang. posición	23.4°	126.3°	(2)
Inclinacion	34.8°	61.3°	(2)

Notas: (*) Ejes principales (V); (1) NED; (2) LEDA.

LAXY SURVEY (ANGST; Dalcanton et al. 2009). Los datos utilizados cubren un porcentaje importante de los campos de las galaxias NGC 1313 y NGC 2403 (ver imágenes en LEGUS y en ANGST). En el caso de NGC 1313 se disponen datos en los filtros $F275W$, $F336W$, $F435W$, $F555W$ y $F814W$, mientras que en el caso de NGC 2403 se disponen datos en los filtros $F435W$, $F475W$, $F606W$ y $F814W$.

3. Metodología

Los métodos utlizados en diferentes fases del proceso fueron *Gaussian mixture method* (GMM), *K-Means*, *Agglomerative clustering* (AgC), *Density based spatial clustering of applications with noise* (DBSCAN), *Hierarchical DBSCAN* (HDBSCAN) y *Path linkage criterion* (PLC). Los detalles de los métodos se pueden encontrar en Battinelli et al. (2000), en Pedregosa et al. (2011) y en McInnes et al. (2017).

El proceso se basó en el empleo de los diagramas color-magnitud (CMDs) y color-color (TCDs) de los objetos de las galaxias bajo estudio para distinguir sus componentes. Los pasos que se siguieron fueron:

- Se pre-procesaron los datos fotométricos, estimando las magnitudes faltantes con el método *Iterative Inputer*. (Pedregosa et al., 2011).
- Se separaron los objetos brillantes ($V < 24$) de los objetos débiles ($V > 24$), donde la magnitud V corresponde a la banda $F555W$ o a la banda $F606W$ dependiendo de los datos disponibles en la zona de cada galaxia.
- Se identificaron las diferentes componentes utilizando GMM sobre los TCDs de cada galaxia. O sea, se utilizaron como características a los índices de color disponibles para cada galaxia. La cantidad de componentes se estableció mediante el uso del *Bayesian information criterion* (Schwarz, 1978).
- Las diferentes componentes gaussianas fueron agrupadas entre sí con el método *K-Means*, aplicado sobre sus correspondientes centros en el espacio de índices de color y fijando una búsqueda de dos grupos. Se distinguieron entonces dos poblaciones principales denominadas población azul y población roja.
- Se estimó la dimensión fractal (D, Mandelbrot 1982) utilizando el método perímetro-área sobre los mapas de densidad estelar de los objetos brillantes de la población azul y de la población roja (ver detalles en Rodríguez et al. 2019)
- Se aplicaron diferentes métodos de agrupamiento sobre la distribución espacial de los objetos brillantes de la población azul y se evaluaron sus parámetros más relevantes. En este proceso se utilizaron como características las coordenadas deproyectadas (ξ_P, η_P). Así se tuvo en cuenta la inclinación de cada galaxia.
- Los diferentes métodos fueron validados utilizando el índice Silhouette (Rousseeuw, 1987). Este índice es una medida de qué tan similares son los objetos de un dado grupo en comparación con otros grupos. El índice varía en el rango [-1, 1] y cuanto mayor es, mejor es el método evaluado.

4. Resultados

En la Fig. 1 se presentan los CMDs y TCDs de una de las galaxias estudiadas con la separación de sus poblaciones utilizando el GMM. Por otro lado, la aplicación de los diferentes métodos de agrupamiento para la identificación de los agrupaciones estelares de las poblaciones jóvenes (azules) de cada galaxia ha permitido obtener los parámetros presentados en la Tabla 2. En la Fig. 2 se presentan los resultados obtenidos con AgC y HDBSCAN.

En relación con la estructura espacial de las poblaciones estelares, los valores obtenidos de la dimensión fractal para los objetos brillantes de las poblaciones azules de cada galaxia se presentan en la Tabla 2. No fue posible establecer valores representativos para las poblaciones rojas.

Respecto a las agrupaciones estelares identificadas con los diferentes métodos de agrupamiento, se encontró que todos los métodos utilizados tienen índices Silhouette con valores aceptables. No obstante, los parámetros indicados en la Tabla 2 permitieron establecer que todos

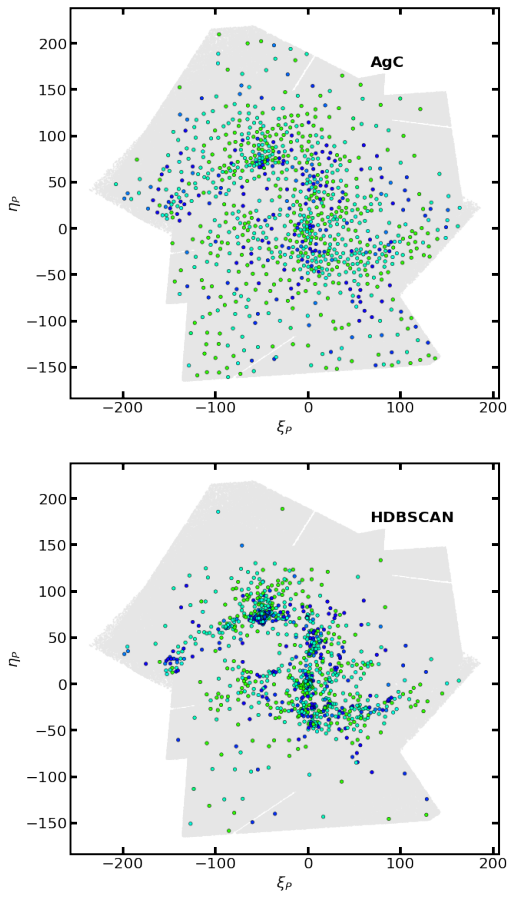

Figura 2: Cartas buscadoras de la zona de la galaxia NGC 1313 en coordenadas deproyectadas (ξ_P, η_P). En ellas se indican los centros de las agrupaciones estelares identificadas por dos de los métodos de agrupamiento. Los colores indican las componentes gaussianas a la que pertenece cada agrupación estelar.

los métodos, excepto AgC, identifican agrupaciones estelares vinculadas con las sobre-densidades estelares. En particular, las agrupaciones identificadas con DBSCAN son pequeñas y todas ellas poseen tamaños similares. Por otro lado, HDBSCAN y PLC proveen agrupaciones de tamaños variados.

5. Conclusiones

En base al estudio realizado, se encuentra que la metodología aplicada a los datos fotométricos de alta resolución espacial producidos por el HST han permitido separar diferentes poblaciones estelares en las galaxias NGC 2403 y NGC 1313. En particular, se han podido separar dos poblaciones en cada galaxia denominadas población azul (joven) y población roja (evolucionada). Además, se han identificados los grupos estelares vinculados con la población azul.

Por otro lado, los objetos brillantes de la población azul parece tener una estructura fractal con características consistentes con las halladas en otras galaxias espirales (Rodríguez et al., 2019), mientras que la población roja es homogénea salvo fluctaciones a muy gran escala.

Finalmente, en relación con los diferentes métodos utilizados para identificar las agrupaciones estelares en la población azul, se encuentra que los métodos HDBSCAN y PLC son los más apropiados para esta finalidad.

Tabla 2: Parámetros obtenidos con los diferentes métodos de agrupamiento aplicados sobre la distribución espacial de los objetos de la población azul de las dos galaxias estudiadas.

NGC 2403 ($D = 1.66 \pm 0.02$)

Parámetro	AgC	DBSCAN	HDBSCAN	PLC
N_{CL}	603	585	472	446
R_{med}	4.3”	0.3”	2.2”	1.4”
σ_R	1.4”	0.1”	1.4”	0.7”
Silhouette	0.39	0.55	0.53	0.36

NGC 1313 ($D = 1.46 \pm 0.02$)

Parámetro	AgC	DBSCAN	HDBSCAN	PLC
N_{CL}	916	854	925	697
R_{med}	2.4”	0.2”	1.2”	1.1”
σ_R	1.0”	0.1”	1.1”	0.7”
Silhouette	0.43	0.63	0.57	0.38

Notas: D = Dimensión fractal; N_{CL} = Nro. de cúmulos; R_{med} = Radio medio; σ_R = Dispersión de radios.

6. Perspectivas a futuro

Los resultados obtenidos son preliminares y en el futuro se realizará un refinamiento de la metodología utilizada considerando otras variantes en la selección de objetos en los CMDs/TCDs y en el modelado de las distribuciones en todos los diagramas. Además se buscará una vinculación entre las componentes encontradas y las fases evolutivas de los objetos. Se intentará entonces aplicar el procedimiento a una muestra amplia de galaxias cercanas utilizando el mismo tipo de datos.

Agradecimientos: Trabajo parcialmente financiado por el PIPs 112-201701-00055 de CONICET y los Programas de Incentivos 11/G158 y 11/G168 de la UNLP. El estudio se ha basado en observaciones realizadas con el *Hubble Space Telescope* (NASA - ESA), y obtenidas del Hubble Legacy Archive (HLA), que es una colaboración entre el STScI (NASA), la ST-ECF (ESA) y el CADC (NRC-CSA). Los autores agradecen al editor y al árbitro de este artículo por sus sugerencias y comentarios.

Referencias

Battinelli P., et al., 2000, A&A, 357, 437
Dalcanton J.J., et al., 2009, ApJS, 183, 67
Lee J.C., et al., 2014, *American Astronomical Society Meeting Abstracts*, vol. 223, 217.01
Mandelbrot B.B., 1982, *The fractal geometry of nature*, vol. 1, WH freeman New York
McInnes L., Healy J., Astels S., 2017, The Journal of Open Source Software, 2
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Rodríguez M.J., Baume G., Feinstein C., 2019, A&A, 626, A35
Rousseeuw P., 1987, Journal of Computational and Applied Mathematics, 20, 53
Schwarz G., 1978, Ann. Statist., 6, 461