

Detección topológica de comunidades estelares en el Puente de Magallanes mediante Machine Learning y datos de Gaia DR3

Topological detection of stellar communities in the Magellanic Bridge using Machine Learning and Gaia DR3 data

Doctorando:

Martin CASATTI

<https://orcid.org/0000-0002-1586-6355>

Universidad Tecnológica Nacional, Facultad Regional Córdoba, Argentina

mcasatti@frc.utn.edu.ar

Director:

Marcelo MARCISZACK

Codirector:

Carlos FEINSTEIN

Resumen

La creciente disponibilidad de datos astrométricos y fotométricos masivos, como los provistos por la misión Gaia (DR3), exige el desarrollo de enfoques analíticos escalables. En este trabajo presentamos un *pipeline* automatizado para la detección y validación de comunidades estelares en la región del Puente de Magallanes. Reemplazamos el enfoque tradicional de búsqueda de sobredensidades puramente espaciales (2D) por una búsqueda topológica. Para ello, modelamos las agrupaciones construyendo un grafo no dirigido sobre un espacio de fases ampliado de cinco dimensiones, constituido por las coordenadas espaciales (α , δ), el paralaje (ϖ) y los movimientos propios (μ_{α^*} , μ_{δ}). Sobre esta red k -NN aplicamos el algoritmo de Leiden para identificar comunidades cinemáticamente coherentes. Para la validación astrofísica, analizamos Diagramas Color-Magnitud (CMD) superponiendo isocronas teóricas PARSEC calibradas para poblaciones jóvenes. La metodología demuestra ser altamente robusta frente a la contaminación por estrellas de campo galácticas, estableciendo una base adaptable para futuros catálogos profundos, como los del Observatorio Vera C. Rubin.

Palabras clave: cluster; nubes de magallanes; machine learning; topología.

Abstract

The increasing availability of massive astrometric and photometric datasets, such as those provided by the Gaia mission (DR3), demands the development of scalable analytical approaches. In this work, we present an automated pipeline for the detection and validation of stellar communities in the Magellanic Bridge region. We replace the traditional approach of searching for purely spatial (2D) overdensities with a topological search. To achieve this, we model the groupings by constructing an undirected graph over an expanded five-dimensional phase space, consisting of spatial coordinates (α, δ), parallax (ϖ), and proper motions ($\mu_{\alpha*}, \mu_{\delta}$). We apply the Leiden algorithm to this k -NN network to identify kinematically coherent communities. For astrophysical validation, we analyze Color-Magnitude Diagrams (CMD) by superimposing theoretical PARSEC isochrones calibrated for young populations. The methodology proves to be highly robust against galactic field star contamination, establishing an adaptable baseline for future deep catalogs, such as those from the Vera C. Rubin Observatory.

Keywords: cluster; magellanic cloud; machine learning; topology.

Introducción

El Puente de Magallanes representa una de las estructuras de marea más interesantes de nuestro grupo local, conectando la Pequeña y Gran Nube de Magallanes. Sin embargo, la detección de agrupaciones estelares débiles en esta región presenta un desafío observacional significativo debido a su baja densidad intrínseca y la alta contaminación visual por estrellas de campo de la Vía Láctea (Bica et al., 2020).

Los métodos de búsqueda tradicionales basados en inspección visual o sobre-densidades espaciales simples 2D resultan inefficientes frente a los volúmenes de datos masivos proporcionados por la misión Gaia DR3 (Gaia Collaboration, Vallenari et al., 2023). En este trabajo, proponemos un enfoque no supervisado de *Machine Learning* que modela la astrometría estelar como un problema topológico mediante la construcción de redes complejas.

Desarrollo

La muestra estelar fue extraída directamente del archivo de la misión Gaia (DR3) mediante una consulta ADQL (*Astronomical Data Query Language*). El proceso de selección y limpieza de datos se estructuró en dos etapas: una restricción inicial en el servidor para definir la geometría y eliminar la contaminación primaria, seguida de un post-procesamiento de alta fidelidad para garantizar la calidad astrométrica.

Para aislar las fuentes candidatas en la región de interés y minimizar el volumen de datos descargados, la consulta en el servidor exigió los siguientes cortes fundamentales:

Se aplicó un conjunto de filtros espaciales, geométricos y fotométricos para obtener una muestra de estrellas representativa del sistema magallánico y minimizar la contaminación por objetos no relevantes. En primer lugar, se restringió la búsqueda a una región circular centrada en el Puente de Magallanes, limitando el análisis a la estructura de interés. Además, se impuso un corte de paralaje ($\varpi < 1,0$ mas), que elimina la mayor parte de las estrellas cercanas del disco galáctico y permite aislar las poblaciones ubicadas a la distancia de las Nubes de Magallanes.

Por otra parte, se estableció un límite de magnitud aparente ($G < 20,5$) para garantizar una calidad astrométrica adecuada, evitando que las grandes incertidumbres de las fuentes más débiles afecten la construcción del grafo topológico. Finalmente, solo se consideraron estrellas con índices de color válidos, requisito indispensable para la posterior validación astrofísica mediante Diagramas Color-Magnitud (CMD), asegurando la integridad del flujo de análisis.

Sobre el catálogo inicial se aplicó un proceso de depuración para eliminar fuentes espurias, sistemas binarios no resueltos y mediciones de baja calidad. Para ello se conservaron únicamente las estrellas con una medición de paralaje estadísticamente significativa, exigiendo una relación señal-ruido ($\varpi/\sigma_{\varpi} > 5$), un bajo ruido astrométrico ($\epsilon_i < 1,0$ mas), alta precisión fotométrica ($F_G/\sigma_{F_G} > 50$) y pequeñas incertidumbres en los movimientos propios ($\sigma_{\mu_{\alpha}} < 0,5$ mas yr⁻¹ y $\sigma_{\mu_{\delta}} < 0,5$ mas yr⁻¹). Estos criterios garantizan la confiabilidad de las mediciones de distancia, fotometría y cinemática empleadas en el análisis.

Finalmente, para evitar sesgos dimensionales durante la construcción del modelo de similitud, las coordenadas espaciales y los movimientos propios se normalizaron mediante escalado estándar. Este procedimiento asegura que las diferencias de escala entre las variables, expresadas en grados y milisegundos de arco por año, no introduzcan ponderaciones artificiales, permitiendo que todas las características contribuyan de forma equilibrada a la métrica utilizada para identificar agrupamientos estelares.

El corazón del algoritmo consiste en proyectar el espacio de fases de 5 dimensiones (RA, Dec, ϖ , μ_{α} , μ_{δ}) en un grafo no dirigido. La topología de la red se definió conectando cada estrella con sus k vecinos más cercanos (k -NN) basándose en la distancia Euclidiana en el espacio escalado.

Para la detección de agrupaciones, se aplicó el algoritmo de Leiden (Traag et al., 2019), el cual supera las limitaciones de resolución de métodos anteriores optimizando iterativamente la partición del grafo para maximizar su modularidad.

El núcleo analítico de este trabajo consiste en transicionar de una búsqueda de sobredensidades puramente espaciales a un enfoque topológico y cinemático. Para lograrlo, las observables astrométricas de Gaia se mapearon hacia una red compleja no dirigida, aislando las verdaderas asociaciones físicas subyacentes mediante técnicas de partición de grafos.

Cada fuente estelar es modelada como un nodo aislado dentro de un espacio de fases de cinco dimensiones, definido por sus coordenadas ecuatoriales, paralaje y vector de movimiento propio: $\mathbf{v}_i = (\alpha, \delta, \varpi, \mu_{\alpha}, \mu_{\delta})$.

Dada la marcada heterogeneidad de las magnitudes físicas involucradas (grados para la posición espacial frente a milisegundos de arco para la cinemática), se aplicó una estandarización previa (*Standard Scaler*). Cada variable se centró en su media y se escaló a varianza unitaria, garantizando así un espacio de características isotrópico donde ningún parámetro domina la métrica por una mera diferencia de escala.

Sobre este espacio normalizado, se construyó un grafo de k -vecinos más cercanos (k -NN). Se establece una arista entre los nodos i y j si j se encuentra entre los k vecinos espaciales y cinemáticos más próximos de i , minimizando la distancia Euclidiana. La densidad de esta matriz de adyacencia refleja la probabilidad de que un grupo de estrellas comparta un origen dinámico común.

Algoritmo de Leiden y Optimización de Modularidad

Para la extracción de los cúmulos estelares a partir de la intrincada red de conexiones, se implementó el algoritmo de detección de comunidades de Leiden. Este método representa una evolución crítica sobre el algoritmo tradicional de Louvain,

diseñado específicamente para evitar el problema de las comunidades internamente desconectadas o débilmente enlazadas.

El algoritmo opera optimizando iterativamente una función de calidad conocida como modularidad (Q). La modularidad mide la densidad de las aristas dentro de una comunidad en comparación con la fracción esperada de conexiones si las aristas se distribuyeran de manera puramente aleatoria. Físicamente, maximizar Q equivale a identificar subgrafos altamente cohesivos; es decir, grupos de estrellas que viajan juntas por el espacio tridimensional con una alta coherencia interna cinemática.

Al garantizar matemáticamente que las comunidades resultantes estén bien conectadas, el uso de Leiden asegura que los cúmulos extraídos sean estructuras topológicamente continuas, reduciendo significativamente la tasa de falsos positivos generados por alineaciones visuales aleatorias del campo estelar galáctico.

Resultados y Validación Astrofísica

Una vez identificadas las comunidades espacial y cinemáticamente coherentes mediante el algoritmo topológico, se procedió a validar su naturaleza física como posibles cúmulos estelares. Para ello, se analizó la distribución de sus estrellas en Diagramas Color-Magnitud (CMD), utilizando la fotometría de Gaia y comparándola con modelos evolutivos teóricos.

La interpretación de estos diagramas se realizó mediante isocronas PARSEC v1.2S, compatibles con las bandas fotométricas de Gaia (G , G_{BP} y G_{RP}), obtenidas a través de la herramienta CMD 3.7 del Observatorio Astronómico de Padova. Los modelos se parametrizaron con una metalicidad subsolar ($Z = 0,004$), representativa del material estelar del Puente de Magallanes, y con edades del orden de los 100 millones de años ($\log(\text{Age}) \approx 8,0$), proporcionando el marco teórico para contrastar las propiedades fotométricas de las poblaciones detectadas.

Para cada comunidad identificada por el algoritmo de Leiden se construyó un Diagrama Color-Magnitud (CMD), representando la magnitud aparente G en función del índice de color $BP - RP$. La validación consistió en superponer una isocrona teórica PARSEC, ajustada mediante un módulo de distancia de $\mu \approx 18,9$ y una corrección por enrojecimiento interestelar $E(BP - RP)$, para llevar el modelo a las condiciones observacionales del sistema magallánico.

La concordancia entre la distribución de las estrellas observadas y la isocrona teórica permitió confirmar la naturaleza física de las comunidades detectadas. Aquellas que reproducen la Secuencia Principal y el punto de giro (*Turn-off*) característicos de una población estelar coetánea fueron clasificadas como cúmulos estelares jóvenes, mientras que las agrupaciones sin esta correspondencia se descartaron como alineaciones fortuitas de estrellas del campo galáctico.

Validación Cruzada Espacial (*Cross-Match*)

Para cuantificar la eficiencia, robustez y confiabilidad del algoritmo de detección topológica propuesto, se llevó a cabo un proceso de validación cruzada espacial frente a un "verdadero terreno" (*ground truth*) establecido en la literatura.

La validación de las detecciones se realizó utilizando como referencia el catálogo de cúmulos y asociaciones estelares de Bica et al. (2020), obtenido a través de la base de datos Vizier, por constituir el inventario más completo de agrupaciones confirmadas en la Gran Nube de Magallanes (LMC), la Pequeña Nube de Magallanes (SMC) y el Puente de Magallanes. Sobre este catálogo se aplicó un procedimiento de *cross-match*, implementado con la librería *astropy*, que calculó la distancia angular mínima entre las comunidades detectadas por el algoritmo y las posiciones registradas en el sistema de coordenadas ICRS, permitiendo cuantificar la correspondencia entre las predicciones del modelo y los cúmulos previamente conocidos, por

medio de los siguientes pasos:

1. **Cálculo de centroides:** Para cada comunidad detectada (con al menos N_{min} miembros) se calculó el centroide espacial mediante el promedio de sus coordenadas ecuatoriales ($\bar{\alpha}, \delta$).
2. **Comparación espacial:** Se determinó la distancia angular mínima ($\Delta\theta$) entre cada centroide detectado y los cúmulos del catálogo de Bica et al. (2020).
3. **Clasificación:** Se consideró una coincidencia cuando $\Delta\theta \leq r_{max} = 2,0$ minutos de arco, clasificando la detección como Verdadero Positivo (TP); en caso contrario, se la clasificó como Falso Positivo (FP), correspondiente a una agrupación no catalogada o a una fluctuación del campo estelar.

Conclusiones y Trabajo Futuro

En este trabajo, hemos presentado y validado una metodología no supervisada para la detección automatizada de agrupaciones estelares débiles en la compleja región del Puente de Magallanes. Alejándonos de las tradicionales búsquedas de sobredensidades puramente espaciales (2D), demostramos que modelar el espacio de fases astrométrico (5D) provisto por Gaia DR3 como una red topológica k -NN permite aislar estructuras cinemáticamente coherentes que de otro modo quedarían ocultas bajo la severa contaminación del campo galáctico en primer plano.

La aplicación del algoritmo de partición de grafos de Leiden probó ser altamente eficiente para maximizar la modularidad de estas redes, garantizando la extracción de comunidades estelares cohesivas. La validación física de estas detecciones, realizada mediante la superposición automatizada de isocronas teóricas PARSEC ($Z = 0,004$, $\log(\text{Age}) \approx 8,0$) sobre Diagramas Color-Magnitud (CMD), confirmó que el *pipeline* es capaz de recuperar poblaciones estelares jóvenes, consistentes con los episodios de formación estelar característicos de esta estructura de marea.

Adicionalmente, la validación cruzada espacial contra catálogos empíricos de referencia (e.g., Bica et al., 2020) cuantificó la robustez del método, evidenciando una alta tasa de recuperación de cúmulos verdaderos (TP) y aislando candidatos (FP) que ameritan observaciones fotométricas de seguimiento profundo.

Trabajos futuros

Si bien el algoritmo de Leiden ha demostrado un desempeño sobresaliente, el análisis topológico de datos astronómicos masivos abre la puerta a la importación de técnicas avanzadas desarrolladas originalmente para el Análisis de Redes Sociales (*Social Network Analysis*, SNA), tales como *Infomap* (Rosvall y Bergstrom, 2008) y Propagación de Etiquetas (Raghavan et al., 2007). Como trabajo futuro, proponemos evaluar el rendimiento computacional y la pureza de extracción de los siguientes enfoques de detección de comunidades en grafos:

- **Infomap:** Basado en teoría de la información, identifica comunidades minimizando la longitud de descripción de una caminata aleatoria, lo que permitiría detectar grupos estelares mediante la retención del flujo cinemático.
- **Propagación de Etiquetas (*Label Propagation*):** Algoritmo iterativo de complejidad casi lineal, adecuado para analizar catálogos de cientos de millones de estrellas con bajo costo computacional.

- **Modelos de Bloques Estocásticos (SBM):** Permiten inferir la estructura latente de la red e identificar tanto cúmulos compactos como corrientes estelares con patrones de conexión similares.

Asimismo, la versatilidad de esta arquitectura topológica permite su extrapolación inmediata a otras regiones de interés astrofísico. En particular, su extensión a galaxias cercanas presenta antecedentes y casos de uso relevantes en espirales del Grupo Local (Johnson et al., 2015), sistemas masivos con poblaciones compactas (Hughes et al., 2021) y galaxias enanas del Volumen Local (Weisz et al., 2014). Entre los objetivos prioritarios a corto plazo destacamos:

- **La Corriente Magallánica (Magellanic Stream):** Extender el volumen de búsqueda hacia las estructuras de gas neutro que se extienden tras las Nubes de Magallanes, buscando poblaciones estelares ultra-débiles asociadas a interacciones de marea extremas.
- **Cúmulos Abiertos Disolviéndose (Tidal Tails):** Aplicar los algoritmos de redes en el disco de la Vía Láctea para rastrear las colas de marea estelares que se desprenden de cúmulos en proceso de disolución dinámica.
- **Corrientes Estelares en el Halo Galáctico:** Identificar asociaciones cinemáticas en el halo, remanentes de galaxias enanas acretadas a lo largo de la historia evolutiva de nuestra galaxia.

La naturaleza agnóstica de este *pipeline* frente a la dimensionalidad de los datos lo posiciona como una herramienta metodológica fundamental. El modelo basado en grafos se encuentra estructuralmente preparado para ingerir y procesar los inmensos volúmenes de datos fotométricos profundos que proveerá el inminente *Legacy Survey of Space and Time* (LSST) del Observatorio Vera C. Rubin.

Referencias

- Bica, E., P. Westera, C. Schröder y A. Duarte (2020). «A multi-band view of the Magellanic Clouds». En: *Monthly Notices of the Royal Astronomical Society* 498, págs. 205-222.
- Gaia Collaboration, A. Vallenari et al. (2023). «Gaia Data Release 3. Summary of the content and survey properties». En: *Astronomy & Astrophysics* 674, A1.
- Hughes, A. K., D. J. Sand, A. C. Seth et al. (2021). «NGC 5128 Globular Cluster System». En: *The Astrophysical Journal* 914, pág. 16.
- Johnson, L. C., A. C. Seth, J. J. Dalcanton et al. (2015). «PHAT Stellar Cluster Survey. II. Andromeda Project Cluster Catalog». En: *The Astrophysical Journal* 802, pág. 127.
- Raghavan, U. N., R. Albert y S. Kumara (2007). «Near linear time algorithm to detect community structures in large-scale networks». En: *Physical Review E* 76, pág. 036106.
- Rosvall, M. y C. T. Bergstrom (2008). «Maps of random walks on complex networks reveal community structure». En: *Proceedings of the National Academy of Sciences* 105, págs. 1118-1123.
- Traag, V. A., L. Waltman y N. J. van Eck (2019). «From Louvain to Leiden: guaranteeing well-connected communities». En: *Scientific Reports* 9, pág. 5233.
- Weisz, D. R., A. E. Dolphin, E. D. Skillman et al. (2014). «The Star Formation Histories of Local Group Dwarf Galaxies». En: *The Astrophysical Journal* 789, pág. 147.