Facultad
de
Ciencias

Detección, segmentación y clasificación de objetos extragalácticos en los datos del *SKA Science Data Challenge 1*

(Extragalactic source detection, segmentation and clasification on SKA Science Data Challenge 1 data)

Trabajo de Fin de Grado
para acceder al
GRADO EN FÍSICA

Autora: Alicia Muñoz San Miguel
Director: Diego Herranz Muñoz
Febrero - 2025

Resumen

El presente trabajo tiene como objetivo emular una participación en el Science Data Challenge 1, un reto de análisis de datos astrofísicos simulados que fue propuesto en 2018 por la organización SKAO para familiarizar a la comunidad científica con el manejo datos como los del futuro radiotelescopio SKAMid. El challenge pretende servir para fomentar el desarrollo de nuevas técnicas de análisis. Para abordarlo, primero se estudia el challenge dentro de su contexto cosmológico y astrofísico, y se analizan los datos simulados sin perder de vista su significado físico. Se lleva a cabo la detección y caracterización de las fuentes presentes en las imágenes astronómicas simuladas mediante el programa `SExtractor`*, y se propone un método de clasificación de objetos basado en su distribución espectral de flujo. Además, se analizan los resultados obtenidos usando para ello el método de puntuación empleado por SKAO para evaluar a los participantes del challenge original, que fue hecho público con la finalización del mismo. Se obtienen resultados comparables con los del resto de participantes. Más allá de sus objetivos científicos, el texto se concibe con una función didáctica y divulgativa, y pretende ser una especie de punto de partida que abarque de forma completa temas relacionados con el análisis de datos en radioastronomía para los que existe poca bibliografía a nivel intermedio.*

Palabras clave: radioastronomía, SDC1, SKAO, `SExtractor`.

Abstract

This project aims to recreate a submittable participation in the Science Data Challenge 1, a simulated astrophysical data analysis challenge developed by the SKAO organisation in 2018 as a way to help get the scientific community used to working with data similar to that from future radiotelescope SKAMid. The challenge aims to aid developing new analysis techniques. In order to be tackled, said challenge first needs to be studied within a cosmological and astrophysical context, and simulated data needs to be analysed keeping its physical meaning in mind. Source detection and characterisation on simulated astronomical images is carried out via `SExtractor`*, and a spectral flux distribution based source classification method is proposed. Results are analysed using the scoring method developed by SKAO to evaluate the original challenge participants, which was made available by the organisation when the challenge ended. The obtained results behave similarly to those submitted by the science groups for the original SDC1. Apart from its scientific goals, this text is intended to serve a didactic and divulgation purpose and aims to be a reference point encompassing a variety of subjects related to radio astronomical data analysis, for which not a lot of intermediate level bibliography exists.*

Keywords: radioastronomy, SDC1, SKAO, `SExtractor`.

Glosario y Acrónimos

A continuación se incluyen el significado de acrónimos presentes en el texto y definiciones breves de conceptos astrofísicos necesarias para la comprensión del mismo.

SKAO	*Square Kilometer Array Observatory*
SDC1	*Science Data Challenge 1*
PSF	*Point Spread Function* o función de dispersión de punto, término equivalente a *Synthesized Beam*. Función que describe la distorsión de tamaño que sufren los objetos en una imagen tomada mediante interferometría a consecuencia del instrumental.
SFG	*Star Forming Galaxy*, o galaxia formadora de estrellas.
FS-AGN	*Flat Spectrum Active Galactic Nuclei* o núcleos galácticos activos de espectro plano.
SS-AGN	*Steep Spectrum Active Galactic Nuclei* o núcleos galácticos activos de espectro inclinado.
SMBH	*Super Massive Black Hole* o agujero negro supermasivo.
WCS	*World Coordinate System* o sistema mundial de coordenadas, son los coeficientes que describen la transformación geométrica entre dos sistemas de coordenadas; por ejemplo, entre la posición de un objeto proyectada en la bóveda celeste y en coordenadas de píxel sobre una imagen.
`SExtractor`	Programa de *software* que sirve para la detección y caracterización de fuentes en imágenes astronómicas.
`PSFEx`	Programa que funciona como complemento a `SExtractor`, permite modelar la PSF en una imagen para la posterior corrección de sus efectos.
formato *.fits*	O FITS (*Flexible Image Transport System*), formato de imagen usado ampliamente en astronomía, presenta los datos con un cabecero que incluye información sobre su naturaleza; por ejemplo, el cabecero de un archivo FITS contiene su WCS. El formato también se puede emplear para manejar datos tabulados.
formato *.psf*	Formato de archivo característico del modelo de PSF que devuelve el programa `PSFEx`, se trata de una tabla FITS binaria con su encabezado.
Jansky [Jy]	Unidad de flujo de uso extendido en radioastronomía que no pertenece al Sistema Internacional, se define como: 1 Jy $= 10^{-26}\mathrm{W/Hz}\cdot\mathrm{m}^2$.
redshift	O corrimiento al rojo (z), se refiere a la variación que existe entre la frecuencia a la que emite radiación un objeto en movimiento con respecto a la frecuencia que se observa a una distancia determinada del mismo, sea: $$(1+z) = \frac{\nu_e}{\nu_o} \quad (1)$$ Desde un punto de vista cosmológico, la existencia del *redshift* constituye una prueba de la expansión acelerada del universo. Observando a diferentes frecuencias con su correspondiente *redshift* se puede estudiar el universo a distintas edades.

Índice

1. Introducción

Este trabajo tiene como objetivo la elaboración de un catálogo de fuentes a partir de una imágenes astronómicas con las características necesarias para participar el *Science Data Challenge 1*. Con este objetivo, se han de analizar las imágenes de datos simuladas propuestas para detectar, caracterizar y clasificar las fuentes presentes en dichas imágenes. El trabajo también responde a un objetivo didáctico: pretende impulsar el desarrollo de habilidad multidisciplinares y servir para familiarizarse con el modo de trabajo en ambientes científicos de cooperación internacional, además de ayudar a comprender la relevancia de SKAO en el contexto científico actual.

1.1. El SDC1

Square Kilometer Array Observatory (SKAO) es una organización intergubernamental que inició en 2022 la construcción de dos grandes instalaciones de radiotelescopios en Sudáfrica y Australia, de nombres SKAMid y SKALow[1]. Cuando estén operativos, dichos telescopios observarán el cielo en rangos de frecuencia de entre 0.35 y 15.4 GHz, y 50 y 350 MHz respectivamente[2]: como su nombre indica, a frecuencias media y baja.

Ambos telescopios serán radiointerferómetros astronómicos, y se prevé que la complejidad de los datos que se obtengan de su funcionamiento exija el desarrollo de nuevas técnicas de análisis.

En 2018, SKAO presentó el *Science Data Challenge 1* o SDC1, el primero de una serie de *challenges* que la organización ideó con el objetivo servir como guía a la comunidad científica para el momento en que los telescopios SKA estén operativos. El SDC1 es una propuesta de análisis de datos simulados de características similares a los que en un futuro obtendrá el SKAMid.

1.1.1. SKAO y su contexto científico

Las inquietudes astrofísicas y cosmológicas actuales pasan por estudiar el proceso de formación de diferentes estructuras galácticas observables hoy en día. Comprender la evolución y mecanismos que se dan en el interior de los cuerpos celestes implica ahondar en la historia del universo. Para ello, es necesario observar una época sobre la que se tiene muy poca información: el período de tiempo que transcurrió entre la era de formación del Fondo Cósmico de Microondas, unos cuatrocientos mil años después del *Big Bang*, y la formación de las galaxias modernas, datada unos cinco mil millones de años después. El estudio de las regiones o épocas del universo en que todavía no habían comenzado los procesos de formación estelares (el llamado 'universo invisible') en el rango óptico o infrarrojo puede proporcionar información sobre sus edades más tempranas. Proyectos de radioastronomía como SKAO pretenden indagar sobre estos temas estudiando el elemento más abundante del universo: el hidrógeno [3].

Existe una línea de emisión espectral del hidrógeno neutro a una longitud de onda de 21 cm que se corresponde con una transición prohibida entre dos niveles hiperfinos de su

estado fundamental. En la época previa a la formación estelar, el universo estaba poblado por nubes de hidrógeno neutro en cuyo interior ocurrían procesos de excitación mediante mecanismos de colisión o absorción. La traza de la radiación emitida o absorbida como resultado de estos procesos es una fuente de información directa sobre la naturaleza del universo en esta época; y la línea de emisión a 21 cm está poblada por los rastros de dichos mecanismos. De esta forma, observar esta línea espectral es una buena forma de obtener información sobre la evolución del universo.

El estudio de la línea de 21 cm para dar respuesta a cuestiones cosmológicas presenta una gran ventaja. Por su longitud de onda, interactúa muy poco con la materia (a excepción de con los plasmas ionizados). Por ejemplo, no es absorbida por nubes de polvo y llega hasta la Tierra sin sufrir grandes alteraciones. Sin embargo, también presenta ciertos inconvenientes que suponen, sobre todo, un reto técnico. Se trata de una transición prohibida y, por lo general, muy poco intensa y con una traza débil y difícil de observar: para hacerlo se necesita un instrumental muy sensible. Otra desventaja deriva de que se encuentre en el régimen de las ondas de radio: la resolución del instrumental está inversamente relacionada con la longitud de onda, por lo que observar a frecuencias bajas implica hacerlo con antenas cada vez más grandes. Para evitar esto, en radioastronomía se recurre a la interferometría con el objetivo de observar con mayor precisión disponiendo varias antenas separadas entre sí, en lugar de enfrentarse al reto logístico que supone fabricar antenas grandes. La sensibilidad necesaria para la observación de la línea de 21 cm de hidrógeno neutro requiere del uso de un telescopio con antenas muy separadas que cubran un área recolectora muy grande, del orden de 1 km cuadrado de extensión. De esta necesidad surgen el concepto principal y el nombre de la organización SKAO.

El proyecto surgió a finales de la década de los 80, con el objetivo de dar respuesta a las inquietudes cosmológicas de la comunidad científica, inquietudes que se trasladan hasta el día de hoy. De la interacción entre científicos en el ámbito internacional y el desarrollo tecnológico y publicación científica consiguientes a lo largo de las tres décadas posteriores terminó por consolidarse SKAO. La organización se creó oficialmente en noviembre de 2011, y junto con el *European Southern Observatory* (ESO), es la única organización intergubernamental dedicada a la astronomía. Entre otras, las inquietudes astrofísicas de SKAO pasan por el estudio de la formación de planetas similares a la Tierra, la detección de distorsiones gravitatorias del espacio-tiempo, el estudio del origen de los campos magnéticos cósmicos y la formación y crecimiento de agujeros negros [1].

Sin embargo, desde un punto de vista cosmológico, la razón de ser de SKAO es el estudio de las diferentes eras de la historia del universo a través de la observación de la línea de 21 cm.

La longitud de onda de 21 cm se corresponde con una frecuencia de 1.4204 GHz. A esa frecuencia se observa la línea de hidrógeno para el universo actual, es decir, sin *redshift*. Tal y como establece la expresión 1, mirando a frecuencias diferentes se puede observar la línea de 21 cm a diferentes *redshifts* para estudiarla en diferentes épocas de la historia del universo, sean:

- Las Edades Oscuras: el periodo de tiempo en el universo temprano en que todavía no se habían formado las primeras fuentes de luz. Las trazas de radiación datadas de esta época se ven afectadas por un *redshift* ultra alto, por encima de $z \sim 30$. Las mediciones de radiación observada para este *redshift* describen de forma directa la distribución de materia oscura por el universo antes de la formación de las primeras galaxias y estrellas. Para observar a este *redshift* es necesario descender a frecuencias muy bajas, del orden de unos 46 MHz.

- El Alba Cósmica: la era de formación de las primeras estrellas, observable a un *redshift* alto, entre $z \sim 20$ y $z \sim 12$. La distribución de la radiación de la línea de 21 cm deja de describir simplemente las perturbaciones cosmológicas al aparecer objetos luminosos, pero a cambio permite estudiar la estructura y mecanismos de los primeros objetos celestes. El Alba Cósmica termina con la Era de Reionización, la época en que la radiación ultravioleta resultante de la formación estelar continuada proporcionó energía suficiente para ionizar las nubes de hidrógeno neutro presentes en el medio intergaláctico, una época en que las trazas de radiación de la línea de 21 cm son muy débiles. Para estudiar esta era se trabaja en frecuencias del orden de los 100 MHz.

- La Post-Reionización: Durante esta época, la distribución de la línea de 21 cm sigue estando dominado por la densidad de materia y teniendo una expansión decelerada. El estudio de la línea para la posterior era de expansión acelerada, que empieza a ser observable a un *redshift* más bajo (entre $z \sim 0{,}4$ y $z \sim 0{,}7$) y situada en el tiempo 13700 millones de años después del *Big Bang*, proporciona información sobre la distribución de la masa en el universo actual, todo dentro del rango de las ondas de radio [4].

Actualmente el proyecto del telescopio *Square Kilometer Array* (SKA) se encuentra en la fase de desarrollo SKA1. Esta fase contempla la construcción y puesta en funcionamiento de los radiotelescopios SKALow y SKAMid en Australia y Sudáfrica. Cada telescopio funciona con un conjunto de antenas y discos orientables. Como se ha establecido previemante, SKALow observará en el rango de frecuencias entre 50 y 350 MHz, y SKAMid entre 0.35 y 15.4 GHz. De esta forma, permitirán observar la línea de 21 cm del hidrógeno neutro a *redshifts* más altos y más bajos respectivamente. El MeerKAT, rediotelescopio del *South African RadioAstronomy Observatory* (SARAO) inaugurado en 2018, está emplazado en la región de Karoo, la misma zona de construcción del telescopio SKAMid, al que ha servido de precursor. El Meerkat ha supuesto una pieza muy relevante en el campo de la radioastronomía por sí solo, elaborando por ejemplo mapas de emisiones de galaxias activas. Sus 64 discos se integrarán en la zona central del complejo de antenas del SKAMid [5].

El estudio de la línea de 21 cm, y de la radiación del Fondo Cósmico de Microondas en general, presenta problemas adicionales. Una vez desarrollado el instrumental necesario para la observación de la línea de emisión a diferentes *redshifts*, el tratamiento de la imagen resultante sigue presentando retos. En primer lugar, las observaciones obtenidas del cielo no limitan su contenido a la radiación que proviene de la línea de 21 cm: a cada frecuencia se recogen también las emisiones de diferentes cuerpos celestes. Las estre-

llas por ejemplo, que emiten como un cuerpo negro, presentan un pico a una frecuencia determinada pero emiten en todo el continuo del espectro de frecuencias. Las imágenes detalladas a frecuencias bajas de diferentes objetos presentan una gran fuente de información astrofísica, pero la traza de dichos objetos ha de ser filtrada si se quiere estudiar los datos desde un punto de vista cosmológico. Además, el tamaño de los datos obtenidos por los telescopios SKAMid y SKALow presentará un reto sin precedente, con imágenes que deberán ser comprimidas en tiempo real durante la observación debido al gran volumen y complejidad de la información.

El manejo de unos datos de tales características necesita del desarrollo de nuevas técnicas de análisis, junto con la preparación y entrenamiento de la comunidad científica. Para ayudar con la puesta a punto de astrónomos y cosmólogos de todo el mundo, SKAO desarrolló los *data challenges* como una muestra de los futuros datos reales de SKAMid y SKALow, con retos de análisis y caracterización que han ido escalando en dificultad para cada *challenge*. El primero de ellos, el SDC1, propone el análisis de un grupo de datos simulados que se corresponden con las observaciones del futuro SKAMid, y están en parte basadas en las observaciones reales del telescopio MeerKAT [6].

1.1.2. Los *Data Challenges*

Cuando termine del construirse, el *Square Kilometer Array* (SKA) será el radiotelescopio más grande del mundo. El análisis de los datos que se obtendrán de sus observaciones presentará un gran reto debido al volumen y la complejidad de los mismos. Es por este motivo que la organización SKAO desarrolló los *data challenges*: con el objetivo de preparar a la comunidad científica, ayudando con el desarrollo de nuevas técnicas para el análisis de los datos que se obtendrán de las observaciones de SKA.

Hasta la fecha actual SKAO ha propuesto tres *challenges* con diferentes niveles de dificultad. Mientras que el primer *data challenge* invita a la detección y caracterización de los objetos compactos (galaxias y núcleos activos) presentes en unas imágenes a distintas frecuencias, el segundo aumenta la dificultad ofreciendo como material un cubo de datos simulados a lo largo de un continuo de frecuencias. El tercer *challenge*, actualmente activo, incluye además como objetivo la caracterización y eliminación de objetos no compactos de las observaciones para su uso en cosmología (nubes de emisión sicrotrón, por ejemplo) con el objetivo de separar el flujo emitido por la línea de 21 cm del resto de elementos de la imagen; además del estudio de diferentes parámetros cosmológicos [7]. Los *data challenges* se plantean como una competición entre científicos, con objetivos claros a cumplir y un sistema de puntuación que clasifica a los participantes en función de su grado de éxito en el análisis. La evaluación y puntuación de los resultados tiene como objetivo motivar la participación, además de discriminar entre los métodos más adecuados y útiles para trabajar con las observaciones de los telescopios de SKA en un futuro. Además de los datos simulados, los equipos cuentan con un *training set* o muestra de la caracterización correcta que se pide para poder afinar los métodos.

El primero de los *challenges* fue publicado en 2018 bajo el nombre de *Science Data*

Challenge 1 (SDC1) [8]. Mediante conjuntos de datos simulados con las mismas características que los futuros datos reales, la propuesta SDC1 invitaba a distintos equipos científicos a participar en la elaboración de un catálogo de fuentes a partir de imágenes simuladas del cielo correspondientes al rango de observación de SKAMid, el radiotelescopio de SKA que estudiará el cielo en frecuencias intermedias. Con la finalización del *challenge* en abril de 2019, se hicieron públicos los resultados obtenidos por los equipos participantes, así como el método seguido por SKAO para calificarlos [9], y los catálogos simulados de fuentes correctos completos (la versión completa de los *training sets*, de nombre *true catalogues*). La publicación de los resultados de los distintos participantes y del método de evaluación permite analizar los datos y puntuar los resultados obtenidos de manera independiente, simulando una participación en el *challenge*, que es el escenario sobre el que se plantea el presente trabajo .

El método implementado está inspirado a grandes rasgos en el procedimiento seguido por de uno de los participantes del SDC1: los científicos del IPM (*Institute for Research in Fundamental Sciences*), que llevaron a cabo la elaboración del catálogo de fuentes mediante el programa *Source Extractor* (`SExtractor`)[9], un software para la detección y caracterización de fuentes en imágenes astronómicas basado en la segmentación. La segmentación es un procedimiento de análisis que divide los datos en trozos que contienen los objetos celestes presentes en una imagen astronómica. Es un modo de proceder especialmente útil para imágenes con una alta densidad de fuentes, tales como las de la propuesta SDC1. Huelga decir que `SExtractor` no es el único programa posible para afrontar el *challenge*: la razón de ser de estos retos es el desarrollo de técnicas de análisis diversas para su evaluación y futura implementación. El resto de participantes del SDC1 emplean procedimientos variados, y utilizan tanto programas de dominio público (como `SExtractor`), como métodos propios. Entre las técnicas empleadas por los equipos se cuentan códigos desarrollados a partir de las librerías `scipy` y `astropy` de `Python`; los programas de extracción de fuentes `ProFound`, `ConvoSource` o `ClaRAN`; además de algoritmos concebidos especialmente para el SDC1, entre otros.

De entre todas las opciones posibles, se ha decidido atacar el *challenge* mediante `SExtractor` porque se trata de una herramienta muy versátil. La configuración es muy personalizable, pero el proceso de detección y caracterización de las fuentes está muy automatizado, y es rápido y preciso. Permite trabajar en diferentes modos y es un programa cuyo uso está muy consolidado, dado que fue desarrollado en la década de los 90.

Una vez se han detectado, caracterizado y clasificado las fuentes, se analiza el catálogo obtenido siguiendo el mismo proceso empleado para evaluar a los participantes del *challenge* basándose en el método de puntuación desarrollado por SKAO para la evaluación del SDC1 [10].

El SDC1 propone el análisis de imágenes simuladas del cielo a tres frecuencias diferentes y con tiempos de integración variados. Las tres bandas de observación se corresponde con frecuencias de 560 MHz, 1.4 GHz y 9.2 GHz. El material proporcionado por SKAO consiste en 9 archivos en formato *.fits* que emulan las futuras observaciones del telescopio

SKAMid. Además de estas imágenes, se cuenta también con los *Primary Beam* y *Synthesized Beam* simulados correspondientes para hacer posible su corrección [8]. Se discutirá su relevancia y significado en secciones posteriores.

Las imágenes han sido simuladas para que contengan dos tipos de objetos celestes: *Active Galactic Nuclei* (AGN) o núcleos galácticos activos, y *Star Forming Galaxies* (SFG) o galaxias formadoras de estrellas. Dentro de las AGN se incluyen dos subtipos de fuentes: *Steep Spectrum Active Galactic Nuclei* (SS-AGN) y *Flat Spectrum Active Galactic Nuclei* (FS-AGN). Se refieren a galaxias que emiten radiación con un espectro cuya forma tiene una pendiente inclinada o se comporta de manera plana, respectivamente. En la siguiente sección se tratarán la morfología y características de los diferentes objetos en profundidad. Además, los cuerpos que aparecen en el catálogo pueden haber sido simulados como fuentes compactas o extendidas dependiendo del tamaño en píxeles de sus ejes [6].

La simulación también contempla las distorsiones que causaría el instrumental en una observación real, es decir, incluye los efectos de los *Primary* y *Synthesized Beams*, consecuencia del uso de antenas en el telescopio y de la configuración en forma de interferómetro astronómico de SKAMid. Se detallará el alcance de las simulación en este ámbito en las secciones correspondientes pero, en esencia, su corrección para la posterior elaboración del catálogo constituye una parte importante del proceso, tanto en el contexto del SDC1 como lo sería para un conjunto de datos reales.

Pese a lo detallado de las imágenes, los datos simulados tienen limitaciones. Las observaciones del SDC1 no incluyen errores de calibración ni la mayoría de errores sistemáticos: por ejemplo, no se simulan las fluctuaciones en la ganancia inducidas por el efecto de la atmósfera (es decir, en estos datos no existe *seeing*) [6]. La simulación tampoco incluye errores de *pointing*. En un futuro, los efectos de estas distorsiones se deberán corregir para tratar con los datos reales.

En resumen: El *challenge* consiste en la detección, caracterización y clasificación de los objetos presentes en las imágenes de datos simulados. Como paso previo a la detección de fuentes, se han de compensar los efectos del instrumental en las imágenes corrigiendo las distorsiones causadas por el *Synthesized Beam* y el *Primary Beam*. De las imágenes 'corregidas' del cielo a distintas frecuencias se han de detectar y extraer las fuentes simuladas para elaborar un catálogo con unas características concretas. El catálogo incluye categorías que tienen que ver con la posición, tamaño y fotometría de los objetos celestes. Posteriormente, se ha de clasificar las detecciones en tipos de objetos diferentes, y por último comparar el catálogo obtenido con el *true catalogue* de fuentes simuladas que se ha utilizado para elaborar las imágenes. Durante el desarrollo del *challenge*, este último paso fue llevado a cabo por SKAO, pero con la publicación posterior del método de calificación empleado y el *true catalogue*, es posible emularlo, que es lo que se va a hacer.

Pese a que el SDC1 y el presente trabajo tratan con datos simulados, es importante entender el *challenge* en su contexto físico y no perder de vista la filosofía del proceso. El objetivo último de trabajar con estos datos es desarrollar estrategias para el análisis de

datos reales en un futuro, por lo que en la medida de lo posible las observaciones se han de tratar como si fueran reales. En las siguientes secciones se van a discutir los elementos del *challenge* como si se tratase de un proceso de análisis de datos reales: se van a estudiar las características reales de los objetos que conforman las imágenes y la relevancia de su estudio dentro del desarrollo de los objetivos con los que se fundó SKAO; el funcionamiento real de SKAMid; y el tratamiento de las imágenes y su corrección como si fuese un proceso realista. Enfrentarse a la simulación desde este ángulo también responde al objetivo didáctico de este trabajo, que pretende servir para ampliar conocimientos relacionados con el tema, y ayudar con el desarrollo de habilidades que son necesarias para familiarizarse con el modo de trabajo en un proyecto científico de las características del SDC1, como pueden ser la programación, el manejo de grupos de datos de gran volumen o la resolución de problemas.

Antes de explicar el proceso que se ha de seguir para desarrollar los objetivos pedidos para el SDC1, es importante estudiar bien el material con que se trabaja, tanto las imágenes como el telescopio.

1.2. Objetos simulados en el SDC1: SFGs y AGNs

Como paso previo a trabajar con las imágenes, se ha de entender qué tipo de objetos van a aparecer en las observaciones y la relevancia de su estudio desde un punto de vista astrofísico, para posteriormente poder clasificarlas. En esta sección se va a discutir la naturaleza de los distintos objetos que aparecen en el catálogo, además de su equivalente simulado en los datos del *challenge*.

1.2.1. Características principales

Como se ha introducido en la sección anterior, en las imágenes del SDC1 se incluyen *Star Forming Galaxies*, y *Active Galactic Nuclei* de dos tipos: FS-AGNs y SS-AGNs.

La evolución galáctica es un ámbito de investigación en constante desarrollo en la astronomía moderna. En concreto, el papel que juegan los agujeros negros supermasivos (SMBH), presenta muchas incógnitas en lo que respecta a su impacto en la morfología de su galaxia, aunque se sabe que existe una correlación entre su masa y la luminosidad de la galaxia [11].

El ritmo al que se forman estrellas en el interior de una galaxia puede proporcionar información sobre su edad, y permite clasificar la población de SFGs en grupos según su etapa evolutiva. En una SFG activa, la formación de estrellas se da con el colapso gravitacional de nubes de gas muy frío. Durante esa fase de su vida, las estrellas jóvenes que se forman en el interior de la galaxia emiten principalmente radiación ultravioleta, confiriendo a la galaxia un color azul. Según va disminuyendo la cantidad de gas frío disponible, su población estelar envejece y va 'migrando' a una zona de emisión que otorga a la galaxia un color rojizo. La zona de transición entre ambas fases se conoce como 'valle verde', denominada así por el desplazamiento en frecuencia que sufre el máximo dominante de su

espectro de emisión al pasar de una zona a otra, y el consiguiente cambio de color. Este proceso de transición se denomina *quenching*. Con el conocimiento actual, se considera un *redshift* de $z \sim 1$ como indicador de la transición de una etapa a otra [12], pero la evolución de las poblaciones estelares en el interior de galaxias es uno de los ámbitos de investigación en desarrollo que se beneficiarán de la existencia de proyectos como SKAO.

Una fracción de las galaxias que contienen un SMBH en su centro tiene también un núcleo galáctico activo (AGN). Para que este se forme han de darse unas condiciones muy concretas en la zona central de la galaxia: se necesita la presencia de gas y una morfología específica, además de otras características concretas. Las AGN obtienen energía de la acreción de materia que sucede alrededor del SMBH. Dentro de una misma galaxia pueden coexistir un núcleo galáctico activo y mecanismos de formación de estrellas, de hecho, se cree que la radiación emitida por el núcleo activo puede afectar a los procesos de enfriamiento de gas necesarios para la formación estelar, acelerando así el proceso de transición de una SFG activa a la zona de *quenching*[12].

1.2.2. Mecanismos de emisión de radiación y morfología

Uno de los objetivos del SDC1 es clasificar las fuentes detectadas en el catálogo en AGNs y SFGs. Aunque los mecanismos de emisión de radiación que se van a discutir pueden coexistir en una misma galaxia, el objetivo es distinguir cuál de ellos predomina en el espectro luminoso medido por SKAMid, y clasificarla de la forma correspondiente.

En el interior de una galaxia se pueden encontrar diferentes objetos que emiten en un espectro continuo, cada uno con un pico a una frecuencia determinada. Como consecuencia de esto, las galaxias emiten también en un continuo de frecuencias, con los diferentes rangos poblados por las trazas de radiación que resultan de diversos mecanismos. Los picos de emisión en este espectro indican qué procesos dominan la radiación emitida por la galaxia, y permiten estudiarla y clasificarla.

Las regiones del ultravioleta y el espectro visible, por ejemplo, están pobladas por la radiación que emiten las estrellas. Dicha radiación, además, calienta las nubes de polvo en el interior de la galaxia y llevan a una emisión de tipo cuerpo negro con un pico en el registro infrarrojo. El movimiento de electrones en el interior de campos magnéticos da lugar a la radiación sincrotrón, que destaca en el régimen de las ondas de radio.

Para ser apreciable, la radiación sincrotrón necesita que en el interior de la galaxia se den mecanismos de emisión de electrones relativistas, y que además exista un campo magnético en que se puedan mover. En el caso de las SFGs, por ejemplo, estrellas muy masivas (con una masa del orden de $M > 8M_{\odot}$) explotan en supernovas de tipos II y Ib, acelerando electrones de rayos cósmicos, que rotan siguiendo los campos magnéticos galácticos y emiten radiación sincrotrón [13]. No se trata de campos magnéticos muy fuertes, por lo que el pico de emisión correspondiente a la radiación sincrotrón no es el más alto en el espectro de las SFGs. En cambio, para una AGN, la interacción entre el núcleo activo y el disco de acreción del SMBH en el centro de la galaxia da como resultado

la emisión de *jets* de electrones relativistas, además de campos magnéticos muy intensos. Por tanto, el espectro de emisión de una galaxia con un núcleo activo está caracterizado por un pico en la región de frecuencias asociada a la radiación sincrotrón. La radiación emitida por este mecanismo responde a una distribución espectral de tipo ley de potencias, en la que se tiene que el flujo dependerá de la frecuencia de la forma:

$$S_\nu \propto \nu^{-\alpha}, \tag{2}$$

donde α es el índice espectral de la radiación. Se puede emplear α para caracterizar el espectro de la AGN como una FS-AGN o una SS-AGN.

Aunque en el espectro de una SFG también se aprecie un pico que corresponde con la emisión de radiación sinctrotrón, no es el mecanismo que predomina en el interior de la galaxia. En el caso de las SFGs, caracterizadas por los procesos de formación estelar que se dan en su interior, la radiación emitida por las estrellas calienta las nubes de polvo dando como resultado una radiación térmica de gran potencia que emite con una distribución de cuerpo negro con un pico en el infrarrojo.

Por tanto, las emisiones de radiación sincrotrón y de radiación térmica de las nubes de polvo se pueden dar en ambos tipos de galaxias, pero predominan la primera para las AGNs, y la segunda para las SFGs. En el contexto de las observaciones de SKAMid, esto se traduce en que la presencia de cada objeto será más abundante en las bandas en que se sitúe su pico de frecuencias.

En cuanto a su morfología, en el SDC1, las SFGs se han simulado como un perfil de Sérsic, proyectado dentro de un elipsoide con el cociente entre el tamaño de sus semiejes mayor y menor como parámetro, y en una posición determinada [6]. Pese a ser, en este caso, objetos totalmente simulados, el uso del perfil de Sérsic hace su comportamiento muy similar al que tendrían los datos reales. Como criterio general, debido a que el mecanismo de emisión principal de las SFGs es la radiación térmica en forma de cuerpo negro de las nubes de polvo, su presencia será más notoria en las bandas de observación a frecuencias más altas. La SFGs son el objeto más abundante en las imágenes del SDC1, tal y como se puede ver en el *true catalogue* proporcionado por SKAO. Además de para la clasificación en diferentes tipos de los objetos, la morfología generada a partir de un perfil de Sérsic se habrá de tener en cuenta a la hora de elegir cómo se obtienen las categorías del catálogo final que están relacionadas con el tamaño y forma de los objetos simulados.

Por otra parte, los AGN constituyen uno de los tipos de fuentes más brillantes del universo y, como se ha mencionado anteriormente, tienen el pico de emisión en el rango de frecuencias correspondiente a la radiación sincrotrón. En el *challenge* se incluyen AGNs de dos tipos en función de la forma que tenga su espectro de emisión a lo largo del rango de frecuencias: con un espectro inclinado (SS-AGN) y con un espectro plano (FS-AGN).

Las SS-AGN, o *Steep Spectrum Active Glactic Nuclei* se caracterizan por que su espectro de emisión decae rápidamente en intensidad con el aumento de la frecuencia (presentan un espectro con una gran pendiente). Típicamente se pueden clasificar en dos morfologías

de doble lóbulo: de tipo FRI (o *core-bright*) y FRII (*edge-bright*). En las SS-AGN de tipo FRI, la zona alejada del núcleo o bulbo es menos brillante que el mismo, y en las de morfología FRII hay zonas brillantes más alejadas del centro de la galaxia [14]. Los objetos SS-AGN generados en las imágenes del SDC1 también presentan estos dos tipos de morfología. En este caso las galaxias no están completamente simuladas, sino que han sido incorporadas a partir de una librería de imágenes reales de SS-AGNs, pero reescaladas y rotadas para no ser iguales a objetos reales [6]. Al venir de imágenes de verdad, el comportamiento espectral de las SS-AGN será muy realista. Su presencia estará concentrada en las bandas de frecuencia más baja, como corresponde a su espectro, que refleja que el flujo emitido decae rápidamente al aumentar la frecuencia.

Las FS-AGNs (o *Flat Spectrum Active Galactic Nuclei*) sin embargo emiten un espectro más plano, que se mantiene uniforme a lo largo de un rango amplio de frecuencias. Tienen una forma más compacta que las SS-AGNs, con un único lóbulo anexado al bulbo. En las imágenes simuladas, las FS-AGNs aparecen como una pareja de componentes: una fracción de flujo correspondiente al *core* como un objeto muy pequeño y compacto, y una componente más grande que representa el lóbulo [6]. Las FS-SGN tienen una presencia más consistente a lo largo de las tres bandas de frecuencia, dado que su espectro de emisión es plano con respecto a la frecuencia.

Los datos simulados también distinguen entre fuentes compactas y extendidas, el límite de tamaño entre ellas es de tres píxeles. Aunque ese tamaño separa los objetos en dos categorías no tiene significado físico, se han escogido tres píxeles como límite por convenio. Las fuentes extendidas están añadidas como sellos en la imagen, y las fuentes compactas están integradas con un filtro gaussiano de convolución en la imagen final [6].

Pese a que todos los objetos cuentan con una distribución espectral de flujo continua, la simulación no incluye variaciones de flujo dentro de cada banda [6]. Como los canales de observación tienen un ancho, el flujo debería cambiar dentro de los mismos en función de la variación de la frecuencia y de la respuesta espectral del instrumento. El valor de flujo que se proporciona para cada objeto en una banda es, por tanto, la integral de su distribución espectral de energía dentro de la banda.

1.3. Sistemática: funcionamiento de SKAMid

Una vez estudiado el contenido de las imágenes del SDC1, queda detenerse en las características físicas del telescopio que las obtendrá. El SKAMid será un radiointerferómetro de síntesis de apertura constituido por 197 telescopios orientables que conformarán en total un área de recolección de 33000 m^2. Se integrarán en su disposición los discos del ya existente radiotelescopio MeerKAT[15].

En esta sección se van a discutir los principios físicos en los que está basado su funcionamiento, y posteriormente se van a estudiar sus características más técnicas, así como su funcionamiento desde un punto de vista más práctico.

1.3.1. Interferómetro astronómico y teoría de la coherencia

SKAMid es un radiointerferómetro astronómico. Está basado en el mismo principio físico que el interferómetro estelar de Michelson y funciona de forma similar. Se cimienta en la interferometría y la teoría de la coherencia, y su funcionamiento es posible gracias a la naturaleza ondulatoria de la luz. Opera bajo la premisa de que la distribución de intensidad de la luz emitida por una fuente y su posición en el cielo pueden deducirse del estudio de la coherencia de la luz emitida a través de la visibilidad del patrón de interferencia en franjas que genera en el plano de observación de un interferómetro.

Las diferentes fuentes y la luz que emiten se pueden clasificar en función de su coherencia. Además de ser coherente o incoherente, una fuente puede presentar coherencia parcial: en un objeto extendido pueden incluso coexistir distintos grados de coherencia en diferentes regiones. Los objetos celestes actúan como una fuente de luz no puntual que emite radiación. La radiación emitida por todos los puntos que conforman la superficie del cuerpo en diferentes direcciones y en relación de fase unas con otras es clásicamente considerada incoherente, o de una coherencia muy baja. Se dice que una fuente es coherente cuando la relación de fase entre las ondas que emite es constante. Esta coherencia tiene una componente espacial y otra temporal o longitudinal: la primera tiene que ver con la forma y tamaño de la fuente y la segunda con su pureza espectral. Una fuente monocromática y puntual emite ondas totalmente coherentes entre sí.

El intervalo de tiempo durante el cual la fase de una onda se puede predecir con precisión para cualquier punto se denomina tiempo de coherencia. Es la medida del tiempo durante cual existe uno de los trenes de onda promedio que componen la onda y modulan su frecuencia. Para una onda monocromática, este intervalo es infinito. Para una onda cuasimonocromática, dura el tiempo que se extiendan sus trenes de onda. El tiempo de coherencia es la medida de la coherencia temporal[16].

La coherencia espacial tiene que ver con el frente de onda del campo emitido, que es una consecuencia de la geometría de la fuente. Cuando dos puntos en un mismo frente de onda mantienen una relación de fase constante, la fuente tiene coherencia espacial. Esto puede darse en ondas con tiempos de coherencia tanto finitos como infinitos, es decir, una fuente puede tener coherencia espacial sin ser temporalmente coherente. Para una fuente puntual que emite la misma radiación en todas las direcciones es fácil ver que cada frente de onda mantiene una relación de fase constante en todos sus puntos independientemente de si el campo es monocromático o no [16].

El estudio del patrón de interferencia de una fuente consigo misma proporcionará, tal y como se ha establecido, información sobre la coherencia de dicha fuente. Para el estudio de la luminosidad del patrón de interferencia proyectado sobre una superficie a una cierta distancia de la fuente, se define la irradiancia I como la energía radiante promedio por unidad de área y tiempo (la 'cantidad' de luz que ilumina la superficie)[16]:

$$I = \langle \vec{E}^2 \rangle_T \tag{3}$$

Si se tienen dos ondas de campo eléctrico $\tilde{E}_1(t)$ y $\tilde{E}_2(t)$ que interfieren en un punto P generando a su vez un campo eléctrico tal que:

$$\tilde{E}_P(t) = \tilde{K}_1\tilde{E}_1(t - t_1) + \tilde{K}_2\tilde{E}_2(t - t_2), \tag{4}$$

(donde $\tilde{K}_1$ y $\tilde{K}_2$ son los propagadores), entonces la irradiancia del patrón resultante de la interferencia de ambas ondas en P resulta:

$$I = \langle\tilde{E}_P(t)\tilde{E}_P^*(t)\rangle_T = I_{S1} + I_{S2} + I_{S1S2}, \tag{5}$$

donde los términos individuales:

$$\begin{aligned} I_{S1} &= \langle\tilde{E}_1(t)\tilde{E}_1^*(t)\rangle_T \\ I_{S2} &= \langle\tilde{E}_1(t)\tilde{E}_2^*(t)\rangle_T \\ I_{S1S2} &= \tilde{K}_1\tilde{K}_2^*\langle\tilde{E}_1(t+\tau)\tilde{E}_2^*(t)\rangle_T + \tilde{K}_1^*\tilde{K}_2\langle\tilde{E}_1^*(t+\tau)\tilde{E}_2(t)\rangle_T \end{aligned} \tag{6}$$

I_{S1} e I_{S2} se refieren al aporte a la irradiancia que hacen por separado cada uno de los campos eléctricos. El término de la irradiancia correspondiente a la interferencia entre ambos campos, I_{S1S2}, es el factor cuyo estudio da información relacionada con la coherencia de la fuente.

A continuación se describe el aspecto del patrón de interferencia correspondiente a varios tipos de fuentes clasificadas en función de su coherencia, y cómo el grado de la misma se relaciona con el término de interferencia de la irradiancia planteado en la ecuación 6:

- Fuente coherente: Se proyecta un patrón de interferencia con franjas bien definidas. La relación de fase entre las ondas que salen de S_1 y S_2 es constante, por lo que el término I_{S1S2} de la ecuación 6 y, en consecuencia, la irradiancia total (ecuación 5) dependerán de la diferencia de fase entre ambas ondas. I_{S1S2} podrá tomar tanto valores positivos como negativos, pero nunca será nulo y oscilará a lo largo del tiempo, dando como resultado un patrón de interferencia en franjas.

- Fuente parcialmente coherente: Se observa un patrón de interferencia que es la superposición del patrón generado por cada onda individual. En un interferómetro, está modulado por el patrón de difracción de Fraunhofer. En función del ángulo que subtienda la fuente y de su forma, el patrón estará más o menos definido. Si los diferentes patrones quedasen superpuestos máximo con máximo, el patrón de interferencia resultante total estaría bien definido, aunque los mínimos no llegarían a ser completamente nulos. Si por el contrario se superpusieran con un desfase de π, de tal forma que mínimos y máximos coincidiesen, entonces el patrón se desdibujaría en una mancha de luz de irradiancia uniforme.

- Fuente incoherente: No se produce patrón observable.

Para una fuente extendida con coherencia parcial, se puede establecer la condición límite para un patrón de interferencia bien definido con el objetivo de estudiar el límite de la coherencia. Se puede definir un área contenida dentro del patrón proyectado dentro

de la cual el patrón de interferencia en franjas esté bien definido. Dicha superficie se denomina área de coherencia, A_c, y su tamaño aumenta cuando lo hace la distancia entre la fuente y la superficie en que se está observando el patrón[16]. Es decir, el área que delimita la zona en la que el patrón de interferencia está bien definido y en la que luz se comporta como si fuera coherente es más grande cuanto más alejados estén la fuente y el plano de observación. Así, la luz emitida por una fuente que en un principio podría no ser considerada coherente, se comporta de una forma cada vez más coherente según se aleja de su fuente de emisión.

La calidad del patrón de interferencia se puede medir de manera cuantitativa a partir de la visibilidad:

$$\mathscr{V} = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{7}$$

La visibilidad del patrón del interferencia y el término de interferencia I_{S1S2} de la expresión de la irradiancia (ecuación 5) se pueden relacionar a través de la llamada función de coherencia mutua:

$$\tilde{\Gamma}_{12}(\tau) = \langle \tilde{E}_1(t+\tau)\tilde{E}_2^*(t)\rangle_T \tag{8}$$

La función de coherencia mutua es, matemáticamente, la correlación cruzada entre dos puntos del campo eléctrico emitido por la fuente, 1 y 2. En su versión normalizada se denomina grado de coherencia complejo, $\tilde{\gamma}_{12}(\tau)$, y su módulo toma valores entre 0 y 1 e indica el grado de coherencia de la luz emitida por dicha fuente. Para una fuente completamente coherente $\tilde{\gamma}_{12}(\tau) = 1$, en una fuente incoherente $\tilde{\gamma}_{12}(\tau) = 0$, y para las fuentes con diferentes grados de coherencia parcial $\tilde{\gamma}_{12}(\tau)$ toma valores acotados entre esos dos límites.

Para una situación en la que los términos I_1 e I_2 de la expresión 5 sean iguales, se tiene que:

$$\mathscr{V} = |\tilde{\gamma}_{12}(\tau)| \tag{9}$$

De forma que, en las condiciones descritas, la visibilidad de las franjas guarda una relación directa con la parte real de $\tilde{\gamma}_{12}(\tau)$, y por tanto de $\tilde{\Gamma}_{12}(\tau)$. En un dispositivo que proyecte el patrón de interferencia de una fuente sobre una pantalla, como un interferómetro, la visibilidad se puede medir directamente de las intensidades máxima y mínima del patrón de interferencia (ecuación 7) para obtener la función de coherencia mutua y el grado de coherencia complejo.

El teorema de van Cittert-Zernike establece que, para una fuente incoherente, si se trata dicha fuente como una apertura de su misma forma y tamaño, S, y la distancia entre la fuente y el plano de observación en el que se proyecta el patrón de interferencia es mucho mayor que S; entonces la función de coherencia mutua medida entre dos puntos cercanos en la pantalla 1 y 2 es la transformada de Fourier de la distribución de intensidad de la fuente. De forma que, en estas condiciones, medir la visibilidad del patrón de difracción de Fraunhofer proyectado permite determinar la función de coherencia mutua, que además de indicar el grado de coherencia parcial de la fuente permite obtener su imagen

a partir de una transformada de Fourier [17].

Este es el principio en el que se basa el funcionamiento del interferómetro estelar de Michaelson. El interferómetro se vale de una combinación de espejos para adaptar el camino óptico descrito por la luz emitida por el objeto de tal forma que dos rendijas sean iluminadas con haces de luz que proceden del mismo objeto y mantengan una relación de fase constante. El patrón de interferencia se proyecta en una pantalla y su visibilidad se puede medir para determinar la función de coherencia mutua y el espectro luminoso del cuerpo celeste, tal y como garantizan las condiciones del teorema de van Cittert-Zernike.

Un radiointerferómetro funciona de manera similar. En vez de combinar dos haces de luz, se mide electrónicamente la correlación cruzada entre dos puntos dentro de un mismo haz y, de nuevo bajo las condiciones previamente definidas para el teorema de van Cittert-Zernike, se obtiene la distribución de intensidad de la fuente a través de una transformada de Fourier [17]. Como se trabaja en el registro de las ondas de radio, la señal se ha de medir con una antena y, tal y como se ha establecido en secciones anteriores, el tamaño de la misma limita la resolución de las observaciones, sobre todo para longitudes de onda así de grandes. La técnica de síntesis de apertura permite abordar este problema de una forma eficiente.

A partir de ahora, se define el plano u, v como un plano tangente a la bóveda celeste en el punto en que se encuentra el objeto observado.

1.3.2. Síntesis de apertura y características del telescopio SKAMid y su funcionamiento

La respuesta del radiotelescopio a una fuente de luz puntual es el *Synthesized Beam*, también llamada *Point Spread Function* (PSF) [18]. En un caso ideal, una fuente de luz puntual suscitaría un equivalente también puntual en una imagen obtenida del telescopio, pero debido a su naturaleza como interferómetro astronómico, la fuente detectada aparece más grande que su tamaño real. La distorsión se cuantifica mediante la PSF. Aunque se utiliza la respuesta del instrumental a una fuente de luz puntual para modelar y obtener la PSF, la distorsión provocada afecta a todos los objetos que aparecen en la imagen tomada. La imagen obtenida de un objeto a partir de un interferómetro astronómico puede entenderse como la convolución entre la PSF y la distribución de intensidad de luz del objeto [17].

Se denomina *Primary Beam* a la transformada de Fourier del área efectiva o apertura de cada una de las antenas que conforman el radiotelescopio [19] (el 'trozo de cielo' que subtiende la apertura de una antena visto en el espacio de Fourier). Todas las imágenes interferométricas van multiplicadas por el *Primary Beam* de la antena con que han sido tomadas. Dada su importancia, el estudio y corrección de la PSF y el *Primary Beam* serán tratados en profundidad en una sección propia.

Un radiointerferómetro astronómico como SKAMid está constituido por un mínimo

de dos antenas que funcionan de forma acoplada. SKAMid en concreto contará con 197. Cada antena es un elemento independiente y puede tener características individuales, no es necesario que sean todas iguales: de hecho, en SKAMid no lo son, ya que además de sus antenas propias cuenta también con las del MeerKAT, que están montadas sobre discos de radios diferentes.

Todas las antenas del interferómetro reciben radiación. Sin embargo, dependiendo del punto de rotación en que se encuentre la Tierra en ese momento, las ondas de radio llegarán a cada antena en distintos grados de desfase. Para el caso en el que el plano que contiene las antenas de los telescopios coincida con la línea de visión, las ondas estarán en fase y habrá un máximo de interferencia. Como se ha establecido en el aparatado anterior, los mínimos del patrón de interferencia de dichas ondas nunca serán nulos, tal y como corresponde al patrón generado por una fuente no puntual bajo las condiciones que postula el teorema de van Cittert-Zernike. De esta forma, la visibilidad equivaldrá al módulo de la parte real de la coherencia compleja, y su transformada de Fourier será equivalente a la distribución de intensidad de la fuente.

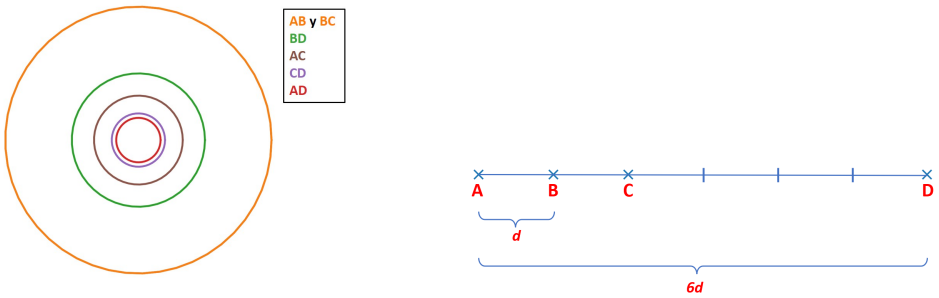


Figura 1: *Ejemplo de disposición de antenas en un interferómetro astronómico de síntesis de apertura. En función de la distancia entre las dos antenas varía el radio del círculo proyectado en la bóveda celeste. Para dos antenas del mismo diámetro que observan en la misma longitud de onda, el radio del círculo será dicha longitud de onda entre la distancia que las separa. Es decir, para las antenas dispuestas en la imagen los radios serán:* $r_{AB} = r_{BC} = \lambda/d$; $r_{BD} = \lambda/5d$; $r_{AC} = \lambda/2d$; $r_{CD} = \lambda/4d$; $r_{AD} = \lambda/6d$ *Se ve que todas las antenas con el mismo espaciado cubren el mismo radio, aunque su posición sea diferente, de forma que no hace falta colocar una antena en todas las posiciones para cubrir el Primary Beam completo. La mayor resolución se corresponde con las antenas que están separadas la distancia máxima (en este caso, A y D) [20].*

SKAMid funcionará mediante la técnica de síntesis de apertura, que se refiere al modo en que las antenas del radiointerferómetro interactúan entre ellas para garantizar el funcionamiento del aparato. La síntesis de apertura es una alternativa al uso de una sola antena de grandes dimensiones para la detección de ondas de radio. Permite ampliar la resolución de las imágenes y el área recolectora del telescopio sin los problemas que supone la fabricación y manejo de una antena de las dimensiones necesarias.

Los interferómetros de síntesis de apertura recurren a la disposición estratégica de varias antenas para obtener imágenes más detalladas de las fuentes que permitan estudiar su estructura. Los distintos discos se colocan a distancias variadas entre ellos de tal forma que, con la rotación terrestre, la posición de cada antena describa un círculo o elipse de radio diferente al proyectarse sobre la bóveda celeste cada 12 horas. Si se cubren todos los espaciados posibles entre las dos antenas con máxima separación, se puede obtener una imagen muy detallada de los objetos contenidos dentro del trozo de cielo que subtiende la apertura de una antena individual (es decir, el trozo de cielo que contiene el *Primary Beam*). En la figura 1 se puede ver un esquema que ilustra el funcionamiento de la técnica de síntesis de apertura.

Al igual que el aumento del radio del disco de una antena aumenta la resolución de la imagen tomada, cuanto mayor sea la separación máxima entre las antenas del interferómetro mayor será la resolución a la que permita observar el telescopio, de ahí la relevancia del proyecto SKA en el contexto científico actual. Las antenas con mayor espaciado de SKA-Mid están alejadas 150km, lo que otorga al radiotelescopio una resolución sin precedente [1].

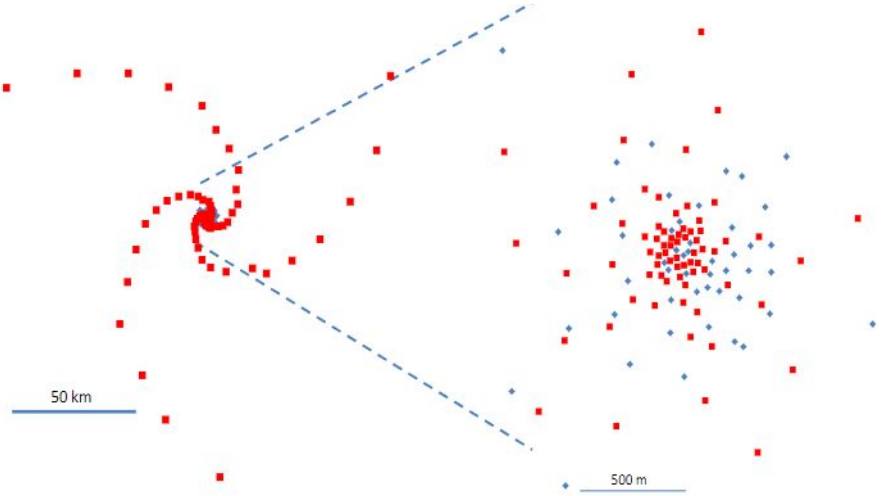


Figura 2: *En la figura de la izquierda se puede ver la configuración de los discos del SKAMid. A la derecha, la zona central, con los discos del MeerKAT (de radio diferente) en color azul[21].*

Existen varias configuraciones habituales a la hora de colocar las antenas: en una línea que se extiende de este a oeste, en forma de Y o T... Los discos del SKAMid en conjunto con los del MeerKAT se dispondrán en forma de espiral, tal y como se puede ver en la figura 2. En total, el complejo incluye 133 antenas nuevas con un disco de 15 m de diámetro, y las 64 antenas del MeerKAT de $13,5$ m de diámetro. La parte externa es una espiral de tres brazos, disposición que cubre muy bien el plano u, v, y los discos están espaciados de forma logarítmica con respecto al centro. La parte central está distribuida de forma diferente, con una gran densidad de antenas para cubrir zonas menos brillantes[21].

1.3.3. Tratamiento de la imagen que se obtiene con un interferómetro: qué son la PSF y el *Primary Beam* y por qué es importante corregirlos

Una vez entendidos el contenido de las imágenes del SDC1 y el funcionamiento de SKAMid, queda estudiar cómo afecta la naturaleza del telescopio a sus observaciones, y cómo se ha de tener esto en cuenta como paso previo al análisis de las imágenes propuesto por el *challenge*.

Del funcionamiento del radiointerferómetro derivan distorsiones que se han de corregir en la imagen final. Estas distorsiones vienen provocadas por la respuesta del interferómetro astronómico a la luz y por el uso de antenas, como es necesario en el rango de las ondas de radio. Como se ha introducido brevemente, el efecto de estas distorsiones está cuantificado y descrito por el *Syntehsized Beam* y el *Primary Beam*.

El *Synthesized Beam* o *Point Spread Function* (PSF) es la respuesta del telescopio a una fuente de luz puntual. La imagen final obtenida por el instrumento es una convolución entre la PSF y la distribución de intensidad de luz de los objetos observados. Todas las imágenes tomadas por un interferómetro astronómico están 'ensuciadas' por la PSF, que distorsiona y difumina la forma de las fuentes detectadas, por lo que su efecto ha de corregirse para medir correctamente los tamaños y posiciones de los objetos.

El *Primary Beam* indica la ganancia de la antena en función de la dirección, y así proporciona información sobre el patrón de difracción que aparece cuando las ondas de radio atraviesan la apertura del telescopio, además de indicar la sensibilidad angular de cada antena como magnitud direccional [22]. Todas las imágenes tomadas por un telescopio de ondas de radio van multiplicadas por el *Primary Beam* de respuesta de sus antenas, y por tanto este se ha de modelar y corregir. Su efecto es más notorio en los bordes de la imagen. Así, en una imagen en la que no se ha corregido el *Primary Beam* se pueden perder fuentes cerca de los bordes, además de distorsionarse la información fotométrica de todos los objetos.

Una imagen captada por un radiointerferómetro como SKAMid presenta, por tanto, distorsiones debidas a las características del interferómetro al completo (la PSF), y a los efectos de cada antena individual (el *Primary Beam*). Para corregir sus efectos se ha de 'limpiar' la imagen haciendo una deconvolución de la imagen con la PSF para que los objetos recuperen su forma real; y se ha de dividir la imagen entre el *Primary Beam* de respuesta de la antena que la ha tomado. Ese es en esencia el proceso, pero presenta complejidades derivadas sobre todo de la morfología del *Primary Beam* y de la dificultad que conlleva su modelado.

Ambas correcciones se pueden aplicar en distinto orden y de formas diferentes, con resultados ligeramente diferentes. La deconvolución de la PSF se puede llevar a cabo mediante algoritmos bien establecidos, como por ejemplo el algoritmo CLEAN.

La respuesta del *Primary Beam* se puede corregir antes, después o durante el proceso

de deconvolución. Para imágenes con un tamaño angular tal que contengan una vez el *Primary Beam* (es decir, con un solo *pointing*), lo más apropiado es corregir el mismo después de hacer la deconvolución entre la imagen y la PSF. Si se tiene sin embargo una imagen que forma parte de un mosaico en el que participan las imágenes tomadas por varias antenas (es decir, el tamaño de la imagen supera el campo de visión de la antena), es más útil hacer la corrección del *Primary Beam* como paso previo a la deconvolución; aunque se ha de tener en cuenta que las regiones del cielo en las que la ganancia del *Primary Beam* sea baja la imagen final tendrá más ruido [23].

La descripción del *Primary Beam* para su corrección es, como se ha mencionado antes, algo compleja. Para un radiointerferómetro astronómico como es SKAMid, en cada imagen tomada por el telescopio se habrá de corregir el *Primary Beam* de su antena correspondiente. El *Primary Beam* es distinto incluso en antenas con discos del mismo tamaño: pequeñas imperfecciones o diferencias entre ellas se reflejan mucho en su forma. En SKAMid, además, se trabajará con antenas que pertenecen a diferentes telescopios (las suyas propias y algunas del MeerKAT) que tienen diámetros distintos, así que la corrección personalizada para cada antena será muy importante. La morfología del *Primary Beam* no es lo único a tener en cuenta, el proceso de corrección también requiere conocer su polarización. Como se ha detallado en la sección anterior, la obtención de una imagen del cielo con un dispositivo como SKAMid pasa por colocar las antenas del interferómetro a distancias tales que cubran la porción del cielo deseada al rotar la Tierra a lo largo del día, provocando que la bóveda celeste gire en relación a las antenas del telescopio. Esta rotación coloca las antenas en orientaciones diferentes según pasan las horas y, como la ganancia de cada antena varía de forma direccional, la polarización del *Primary Beam* no es igual en todas las direcciones.

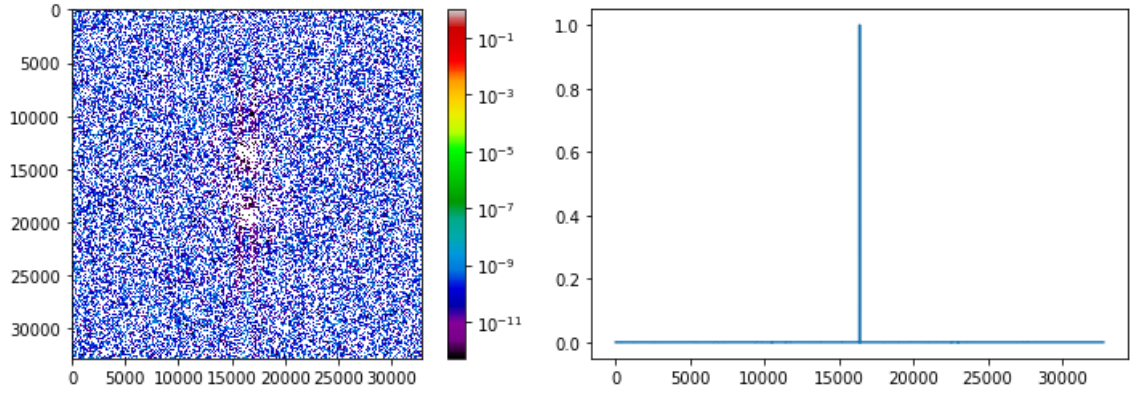
Para corregir el *Primary Beam* empleando un modelo que considere su polarización en todas las orientaciones relativas a la bóveda celeste, se han de calcular sus parámetros de Stokes. Si se tiene la polarización bien descrita para todas las orientaciones, se puede corregir el efecto del *Primary Beam* a la vez que se hace deconvolución de la PSF, proporcionando los parámetros de Stokes al algoritmo que limpia la imagen [24]. Esta es la forma más exhaustiva de las tres de corregir el *Primary Beam*. Las otras dos, que dividen la imagen entre el beam antes o después de la deconvolución, lo hacen solo para una de sus orientaciones, por lo que se obtiene una versión aproximada de la imagen, corregida con un *Primary Beam* simplificado y estático.

En las próximas secciones se detallan la morfología y características de ambos *beams* y sus equivalentes simulados, y se establece el método de corrección más adecuado en función del alcance de la precisión de la simulación en este ámbito.

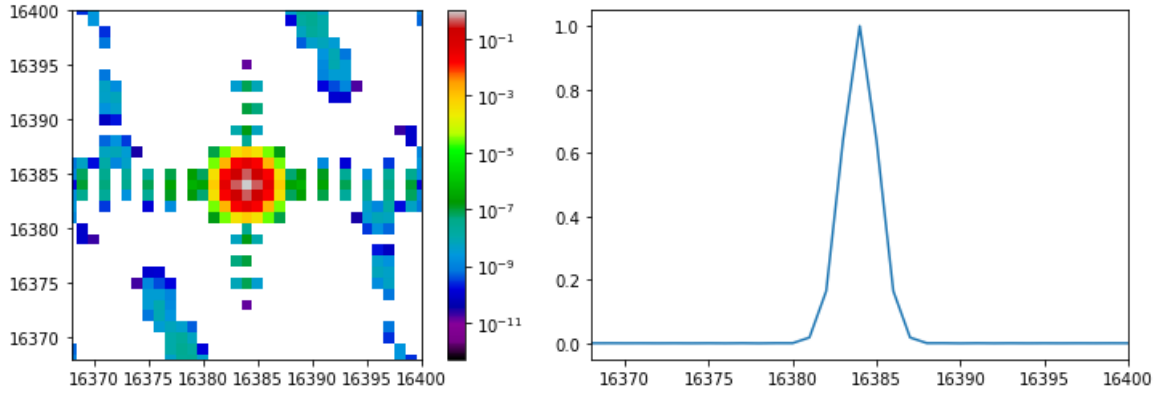
1.3.4. *Point Spread Function* y su corrección

Los efectos de la PSF se aprecian en todas las imágenes tomadas por un interferómetro. Es la respuesta del telescopio a una fuente puntual, y se estudia como una función que va convolucionada con la imagen final. Su corrección es necesaria para estimar de forma

correcta los tamaños y posiciones de los objetos detectados. Dicha corrección cobrará más importancia cuando el tamaño del objeto sea de un orden de magnitud similar al ancho de la función que describe la PSF. Si se tiene una PSF muy estrecha en relación al tamaño promedio de los objetos que aparecen en la imagen, los parámetros medidos no se verán tan afectados por esta distorsión.



(a) Imagen y perfil de la PSF completa.



(b) Imagen y perfil de la región central de la PSF.

Figura 3: *Diferentes vistas del Synthesized Beam simulado para la banda 1 (frecuencia de 560MHz).En la figura (a) se pueden ver la PSF completa y su perfil central. Como el pico principal es muy estrecho y la imagen contiene mucho ruido simulado, no se aprecia la morfología. Es por eso que se ha incluido en la figura (b) una vista de la zona central de la PSF, que muestra la forma gaussiana del pico.*

Si el interferómetro astronómico tuviese una respuesta perfecta a la luz emitida por las fuentes que detecta, la PSF tendría forma de delta de Dirac, y así las fuentes puntuales generarían una respuesta puntual en el instrumental y aparecerían como tal en la imagen final. Una fuente puntual como una estrella quedaría marcada como un objeto de 1 píxel de tamaño en la imagen (en este caso, el elemento más pequeño posible). La distorsión del aparato sin embargo ensucia los contornos del objeto detectado, y ese 'ensanchamiento' viene cuantificado por el ancho de la PSF. La morfología de la PSF se describe comúnmente como una función de tipo gaussiano muy estrecha, aunque dicha forma puede variar de un instrumento a otro. Presenta unas pequeñas oscilaciones a los

lados del pico consecuencia del ruido de fondo en las observaciones, pero su modelado no es tan relevante a la hora de corregir la forma de los objetos.

El SDC1 proporciona una imagen que contiene la PSF para cada uno de los tres canales de observación de SKAMid. En la figura 3 se puede ver la PSF simulada para la banda de frecuencia más baja. Las PSFs están en parte generadas a partir de visibilidades e imágenes de ruido medidas con los discos del radiotelescopio Meerkat. Para construir las funciones, se ha sometido a un mapa de visibilidades tomado por los discos a un ajuste gaussiano, con el objetivo de generar un *beam* con una forma adecuada. Para las tres bandas de frecuencia en las que observa SKAMid, el objetivo era generar tres PSFs con un FWHM de 1.5 arcsec, 0.60 arcsec y 0.0913 arcsec para las bandas de 560MHz, 1.4Ghz y 9.2GHz respectivamente. A partir de las imágenes de ruido se han simulado las oscilaciones a los lados del pico o *sidelobes*, que toman una amplitud absoluta máxima de $4 \cdot 10^{-4}$ para el beam normalizado[6]. Como resultado, las imágenes simuladas en el SDC1 para los *Synthesized Beams* emulan bien la morfología gaussiana deseada. En la figura 4 se puede apreciar la forma de los *sidelobes* consecuencia del ruido junto al pico principal de la PSF.

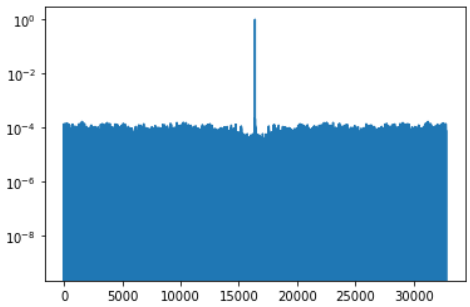


Figura 4: *Perfil de la PSF simulada para la banda 1 en escala logarítmica. La escala logarítmica permite apreciar los sidelobes consecuencia del ruido, que no superan una amplitud de* $4 \cdot 10^{-4}$ *en sus oscilaciones para el beam normalizado.*

Dada la poco significativa amplitud de los *sidelobes* de ruido frente al pico central de la PSF, de querer modelarse para su corrección una función gaussiana sería una buena aproximación. No obstante, el proceso de corrección no pasa por su modelado manual. Como se ha mencionado en la sección anterior, para corregir los efectos de distorsión del telescopio se han de deconvolucionar la imagen y la PSF. Existen algoritmos como CLEAN que llevan a cabo este proceso de forma automatizada. No obstante, la deconvolución es un procedimiento delicado, y estos algoritmos en realidad hacen operaciones equivalentes en el espacio de Fourier sobre la imagen para evitar hacer una deconvolución. Esto es debido a que el proceso de deconvolución es muy susceptible a los niveles de ruido de la imagen, por lo que puede hacer que estos se disparen y la enturbien en vez de corregirla.

Debido a la naturaleza de los datos del SDC1 el uso de una herramienta como CLEAN es excesivo y poco adecuado. No obstante, debido al tamaño del FWHM de las PSFs simuladas en relación con el tamaño de los objetos que aparecen en las imágenes, la corrección de la PSF es un paso necesario para la correcta elaboración del catálogo de fuentes pedido por SKAO. El proceso llevado a cabo se detallará en la sección correspondiente, implica el uso del programa `PSFEx`, un complemento de `SExtractor`.

1.3.5. *Primary Beam* y su corrección

Las distorsiones asociadas al *Primary Beam* en SKAMid son consecuencia del uso de antenas. El *beam* representa la ganancia direccional de cada antena. Su corrección es importante para la detección de fuentes cerca de los bordes del campo de visión, y para la obtención de datos fotométricos precisos de todos los objetos. El *Primary Beam* se ha de corregir de forma personalizada para cada antena, pues varía con sus características físicas. Esto es especialmente importante para los datos obtenidos con SKAMid, que vienen de discos con diámetros diferentes.

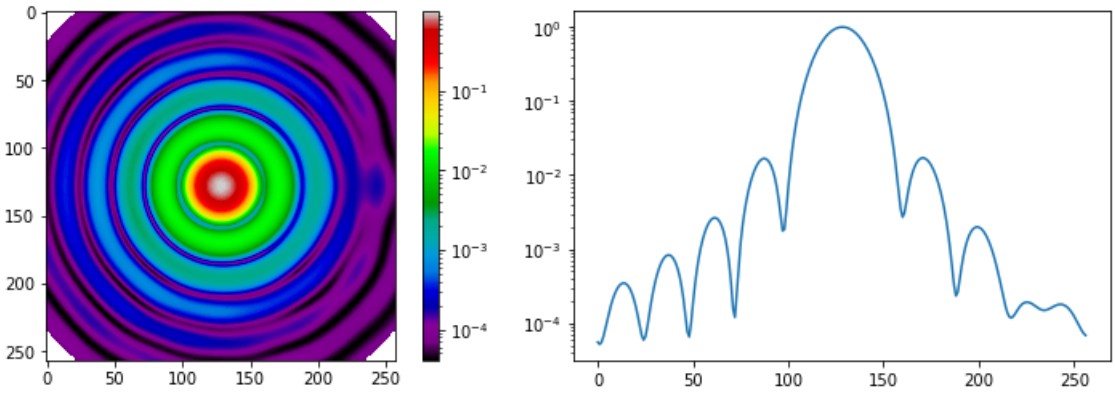
La morfología del *Primary Beam* no tiene un ajuste tan evidente como la de la PSF. Presenta también un pico central con forma gaussiana, más ancho que en el caso de la PSF, pero a diferencia de esta tiene unos lóbulos a los lados con una amplitud que no es despreciable. Para el SDC1 se ha simulado el *Primary Beam* de respuesta correspondiendo a una observación a 1.4 GHz, es decir, la banda intermedia de observación de SKAMid. Los *Primary Beams* para las otras dos frecuencias se han reescalado de forma lineal para contener el campo de visión en función del tamaño de píxel a cada frecuencia. El *beam* está simulado a partir de la respuesta media de dos polarizaciones lineales perpendiculares para una orientación fija, ambas elípticas antes de calcular la respuesta promedio. El *beam* simulado se corresponde con un disco de diámetro de 15 m. Posteriormente se ha multiplicado el modelo simulado del cielo por la respuesta de este *Primary Beam* promedio para emular sus efectos[6]. La figura 5 ilustra la morfología del *Primary Beam* simulado para para la banda 1.

Al multiplicar el cielo por una imagen del beam homogéneo y estático se genera una versión muy simplificada de una situación real. Para corregir las distorsiones que genera el *Primary Beam* en el SDC1, no es necesario tener en cuenta los efectos de las diferentes orientaciones en la polarización, o las variaciones que sufre para cada antena individual, pero para una situación real esto sería de suma importancia. Para esta simulación con alcance limitado, es suficiente dividir las imágenes generadas del cielo entre este *Primary Beam* estático con polarización constante.

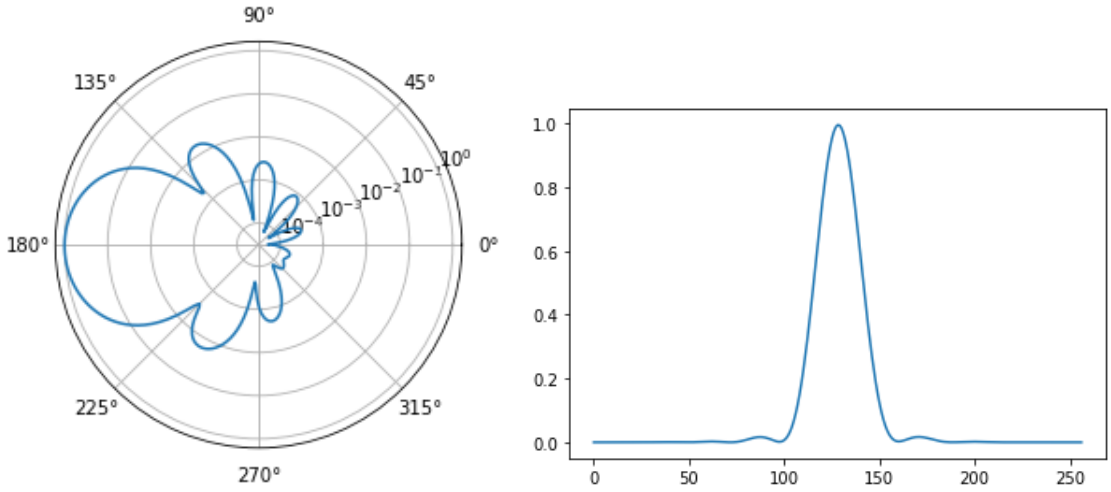
De entre los diferentes procedimientos de corrección enumerados en el apartado anterior, hacerlo durante el proceso de deconvolución de la PSF no es posible dado que sólo se conoce la polarización del *beam* para una orientación. Se ha de corregir antes o después del tratado de la PSF, para la orientación estática del *Primary Beam* con la que se cuenta. Como las imágenes de *challenge* subtienden un sólo *pointing* y no forman parte de un mosaico, lo más apropiado será corregir sus efectos como paso posterior a la corrección de la PSF. Se detallará cómo se han implementado ambas medidas en conjunto en la sección siguiente.

De todas maneras, dado que en la simulación la imagen del cielo se ha multiplicado por el *Primary Beam*, para atenuar sus efectos basta con dividir la imagen entre el mismo. Aunque el procedimento es muy simple, presenta problemas, dado que las imágenes que proporciona para cada frecuencia el SDC1 del cielo y *Primary Beam* simulado correspon-

diente no son del mismo tamaño, por lo que no se puede dividir una imagen entre otra directamente. Para conseguir una imagen del *Primary Beam* adecuada para cada frecuencia se han explorado varias opciones, concluyéndose que la mejor solución es modelarlo para poder reescalarlo al tamaño necesario para cada una de las frecuencias.



(a) Imagen y perfil del *Primary Beam* en escala logarítmica.



(b) Perfil del *Primary Beam* en coordenadas polares y cartesianas.

Figura 5: *Diferentes vistas del Primary Beam de la banda 1 (560MHz).*
En la figura (a) se pueden ver el Primary Beam simulado y su sección central, ambos en escala logarítmica para poder apreciar bien la amplitud de los sidelobes con respecto al pico central. En la figura (b) se puede ver la misma sección en escala lineal, en coordenadas polares y cartesianas para ver la forma de la ganancia de la antena.

Dada la morfología del *Primary Beam*, esta tarea no es trivial. Una función gaussiana describe bien la forma del pico central pero no incluye los efectos de los lóbulos laterales. Pese a que su forma recuerda a una función de Airy, esta no es una buena descripción tampoco. Para el *Primary Beam* de una antena 'perfecta' (es decir, circularmente simétrica y sin mezclas entre los parámetros de Stokes), el mejor ajuste lo proporciona una función de tipo cosenoidal [25]. La forma del *Primary Beam* con sus sidelobes viene bien descrita por la función de potencia:

$$P(\theta) = \left[\frac{cos(\pi\theta D/\lambda)}{1-4[\theta D/\lambda]^2}\right]^2 \tag{10}$$

donde D se refiere a la apertura de la antena y λ a la longitud de onda en la que observa [22]. A través de la expresión $FWHM \approx 1{,}2\frac{\lambda}{D}$, se puede reescribir 10 como:

$$P(\theta) = \left[\frac{cos(\pi\theta 1{,}2/FWHM)}{1-4[\theta 1{,}2/FWHM]^2}\right]^2 \tag{11}$$

de forma que se pueda generar una función que describa el *Primary Beam* con su FWHM como parámetro, que es el dato que se proporciona en la documentación del SDC1. La implementación de este método ha venido motivada por su uso exitoso para describir la forma del *Primary Beam* de los discos del MeerKAT [26]. Una vez modelado a partir de la expresión 11, se puede generar un *Primary Beam* del tamaño necesario y dividir entre él la imagen del cielo, atenuándose así su efecto.

2. Metodología

Teniendo ya una idea general de los objetivos del *challenge* y comprendiendo el material de forma más detallada, la próxima sección es una descripción exhaustiva del procedimiento seguido para simular la participación en el SDC1.

2.1. Desarrollo del SDC1

SKA proporciona para el *challenge* imágenes en formato *.fits*[8]. Se incluyen 9 imágenes simuladas del cielo en tres frecuencias de observación diferentes, que emulan tres de las bandas en las que trabajará SKAMid:
- 560 MHz, en representación de la banda 1 de SKAMid (B1 a partir de ahora)
- 1.4 GHz, en representación de la banda 2 de SKAMid (B2)
- 9.2 GHz, en representación de la banda 5 de SKAMid (B5)

Además, para cada frecuencia se proporcionan imágenes correspondientes a tres profundidades de integración diferentes:
- 8h, simulando una observación de integración baja.
- 100h, una integración de profundidad media.
- 1000h, una integración de profundidad alta.

Se ha trabajado con las imágenes correspondientes a 1000h de observación.

El campo de visión para cada frecuencia de observación se ha elegido de tal forma que las imágenes simuladas contengan el *Primary Beam* de una de las antenas, es decir, un *pointing*. Esto da lugar a los tamaños angulares de mapa y píxeles indicados en la tabla 1. Cada imagen tiene 32768 píxeles de lado independientemente de la frecuencia indicada, y ocupa 4 Gb de tamaño. El campo simulado está centrado en una ascensión recta de RA = 0 y una declinación de Dec = 30 grados[6].

Banda	Tamaño mapa (grados)	Tamaño píxel (arcsec)
B1	5.5	0.60
B2	2.2	0.24
B5	0.33	0.037

Tabla 1: *Tamaños angulares para las distintas frecuencias dependiendo del campo de visión.*

El material del SDC1 también incluye imágenes simuladas de los *Primary Beam* y *Synthesized Beams* correspondientes a cada banda[8]. Las imágenes de los *Primary Beams* tienen 257 píxeles de lado y, las de los *Synthesized Beams*, 32768. Esta diferencia de tamaño es debida a que el *Primary Beam* varía de forma suave y no requiere de una resolución tan alta para que se pueda trabajar con él cómodamente.

	Categoría SDC1	Unidades	Significado físico
C1	ID	-	Número identificador de la fuente
C2	RA (*core*)	grados	Ascensión recta del bulbo o *core* de la fuente.
C3	Dec (*core*)	grados	Declinación del bulbo de la fuente.
C4	RA (centroide)	grados	Ascensión recta del centroide de la fuente.
C5	Dec (centroide)	grados	Declinación del centroide de la fuente.
C6	flux	Jy	Densidad de flujo integrada (con el *Primary Beam* corregido).
C7	core_frac	-	Cociente de la densidad de flujo integrada entre el bulbo y el total de la fuente (con el *Primary Beam* corregido).
C8	bmaj	arcsec	Dimensión del eje mayor de la elipse a la que se ha ajustado la fuente.
C9	bmin	arcsec	Dimensión del eje menor de la elipse del ajuste.
C10	PA	grados	Ángulo de posición del eje mayor de la elipse del ajuste.
C11	size	-	Método de ajuste empleado para estimar el tamaño de la fuente: 1,2,3 para LAS, *Gaussian*, *Exponential*
C12	class	-	Clase en la que se ha clasificado l fuente: 1,2,3 para SS-AGNs, FS-AGNs y SFGs

Tabla 2: *Entradas que ha de incluir el catálogo de fuentes elaborado.*

A partir de la publicación de los resultados de los equipos participantes del SDC1 en 2019, se cuenta también con los catálogos de fuentes completos correctos para cada una de las tres bandas, los *true catalogues*; y con una librería de `Python` que incluye métodos para puntuar catálogos tal y como se evaluaron las propuestas de los equipos participantes

del *challenge*: un *scorer*.

Los objetivos del SDC1 son la detección, caracterización y clasificación de fuentes. Tras la corrección de las imágenes, se han de detectar las fuentes presentes en ellas y elaborar un catálogo. Estos objetivos pretenden garantizar el desarrollo de técnicas que permitan identificar bien los objetos presentes en una observación, lo cual será necesario tanto para estudiar dichos objetos en un contexto astrofísico, como para retirarlos de la imagen y estudiar la línea de 21 cm del hidrógeno desde un punto de vista cosmológico.

El catálogo elaborado debe incluir las categorías enumeradas en la tabla 2 para cada uno de los objetos. Las categorías 2, 3, 4 y 5 tienen que ver con la posición de los centroides y *cores* de los objetos detectados. Las categorías 6 y 7 están relacionadas con su fotometría y las 8, 9 y 10 con su forma. La categoría 11 sirve para indicar la forma en que se han estimado el tamaño y forma de la fuente, y depende del criterio usado para elaborar las categorías 8, 9 y 10. Los tres modos posibles son LAS (*Largest Angular Size*, que da el tamaño como la medida más grande detectada para el objeto, una especie de eje mayor) Gaussian (ajusta el objeto a una gaussiana) y Exponential (ajusta la forma del objeto a un perfil exponencial de Sérsic). La categoría 12 es la que se corresponde con la parte de clasificación de fuentes en función de sus mecanismos de emisión en SFGs y AGNs [27].

2.2. Tratamiento de la imagen previo a su análisis: corrección de *PSF* y *Primary Beam*

En esta sección se va a abordar la corrección de los efectos de la PSF y el *Primary Beam* en las imágenes de las tres bandas como paso previo a la detección de fuentes y la elaboración de catálogos.

2.2.1. Corrección de la PSF

Las distorsiones generadas por la respuesta del telescopio a la luz son más evidentes cuando el tamaño promedio de los objetos observados es similar al ancho de la PSF del instrumental. En la tabla 3 aparecen reflejados para cada banda los FWHM de la PSF y el tamaño medio de los objetos en el *true catalogue* correspondiente. De su estudio se ve que los tamaños de ambas magnitudes son similares en los tres casos, por lo que los efectos de la PSF no son despreciables y su corrección es necesaria.

Banda	FWHM de la PSF (arcsec)	Tamaño promedio objetos (arcsec)
B1	1.5	2.25
B2	0.6	1.7
B5	0.0913	0.82

Tabla 3: *Tamaños promedio de los objetos en las tres bandas y el ancho de la PSF para cada una de ellas.*

Para la corrección de los efectos de la PSF, se va a utilizar el programa PSFEx. PSEFEx es un programa que trabaja como acompañamiento a SExtractor, que a su vez se va a usar para la posterior detección de fuentes. PSFEx emplea la imagen del cielo que se va a estudiar (previamente procesada por SExtractor) para elaborar un modelo de la PSF que posteriormente se puede utilizar para corregir sus efectos al hacer la detección de fuentes con SExtractor [28].

VIGNET (w, h)	Una ventana de dimensiones máximas (w, h) que se ajusta al tamaño de cada objeto. Los parámetros (w, h) se han elegido para contener de forma holgada el área del pico de la PSF, eligiéndose así $w = h = 29$ píxeles. Es recomendado que dicho número sea impar para que sean simétricas respecto al píxel central.
X_IMAGE	Posición X del centroide del objeto, en unidades de píxeles.
Y_IMAGE	Posición Y del centroide del objeto, en unidades de píxeles.
FLUX_RADIUS	El radio que contiene la mitad del flujo emitido por la fuente, en píxeles.
SNR_WIN	El *Signal to Noise Ratio*, o cociente entre la señal y el ruido. El sufijo _WIN indica que el parámetro se obtiene ajustando el objeto dentro de una ventana gaussiana.
FLUX_APER(1)	El flujo del objeto medido a través de una apertura circular de tamaño fijo.
FLUXERR_APER(1)	El error asociado a la magnitud anterior.
ELONGATION	La elongación del objeto, el semieje mayor al que se ha ajustado la forma de la fuente entre su semieje menor.
FLAGS	*Flags* de la extracción.
FLUX_AUTO	El flujo contenido dentro de una apertura elíptica tipo Kron.
FLUXERR_AUTO	El error de la magnitud anterior.
CLASS_STAR	Un número entre 0 y 1 que indica la puntualidad de la fuente. Cuando el número está cerca de 1 la fuente se ajusta mejor a una estrella, cuando tiende a 0 es una galaxia.

Tabla 4: *Parámetros de entrada que ha de contener el catálogo de input para PSFEx.*

El primer paso es, por tanto, procesar la imagen del cielo para cada banda con SExtractor. SExtractor es un programa que detecta las fuentes de una imagen astronómica dada, y elabora un catálogo de objetos con una serie de categorías [29]. En la sección correspondiente a la detección de fuentes se profundizará con más detalle en sus características. Para el correcto funcionamiento de PSFEx, se le ha de proporcionar un catálogo que contenga al menos las categorías enumeradas en la tabla 4 para los objetos detectados. De esta forma, PSFEx no ha de detectar las fuentes, sino que trabaja directamente en las viñetas para cada objeto extraídas por SExtractor, lo que facilita mucho el proceso de modelado de la PSF. Una vez elaborado un catálogo con las categorías nece-

sarias a través de `Sextractor`, se puede poner en funcionamiento `PSFEx`.

`PSFEx` hace una selección de los objetos presentes en el catálogo para evaluar cuáles son buenos candidatos a ser fuentes puntuales y modelar la PSF a partir de ellos, pues la PSF es la respuesta del instrumental a una fuente de luz puntual. Dependiendo de las características de la imagen observada y de la presencia de diferentes objetos, el ajuste obtenido será más o menos preciso. Idealmente, la PSF habría de modelarse a partir de estrellas o quásares, por lo que se necesita de la presencia de dichos objetos en el catálogo de entrada para el correcto ajuste de la PSF. Las imágenes del *challenge* contienen SFGs y AGNs. Ambos son tipos de galaxias, y no está documentada la presencia de ninguna estrella o quásar en el SDC1 [6]. Sin embargo, la presencia de fuentes compactas lo suficientemente pequeñas en la imagen del cielo hace que el ajuste sea posible y que `PSFEx` funcione correctamente.

La configuración de `PSFEx` empleada para extraer la PSF de las imágenes del *challenge* no se ha desviado mucho con respecto a la configuración por defecto. En la sección correspondiente se profundizará en la configuración de `SExtractor` empleada para la detección de fuentes, pero cabe destacar que la alta flexibilidad a la hora de configurar el programa supone una de sus mayores ventajas, y a la vez representa una tarea delicada. Los parámetros que se han personalizado en el archivo de configuración de `PSFEX` para asegurar un funcionamiento adecuado conforme a las imágenes estudiadas están reflejados en la tabla 5.

Parámetro	Modo	Significado
`BASIS_TYPE`	`PIXEL`	Indica que los vectores que definen el modelo de la PSF están escritos en una base de píxeles, la opción con más resolución de las disponibles.
`PSF_SIZE`	(29, 29)	El tamaño final en píxeles del modelo de PSF, elegido para ser consistente con el tamaño de las viñetas y para no ser mucho más grande del tamaño recomendado por `PSFEx` (25, 25).
`SAMPLEVAR_TYPE`	`NONE`	Este parámetro indica si se han de esperar variaciones de la FWHM de la gaussiana de ajuste entre objetos provocadas por el *seeing*. Como las imágenes del SDC1 no incluyen errores de calibración, no se producen variaciones consecuencia del *seeing*.

Tabla 5: *Parámetros modificados en el archivo de configuración de* ***`PSFEx`****.*

Para el resto de parámetros se ha mantenido la configuración por defecto. Con el catálogo de fuentes y el archivo de configuración ya se puede obtener un modelo de la

PSF con `PSFEx` empleando un comando de tipo:

```
psfex catálogo.cat -c config_file.psfex
```

Se obtiene así un archivo que contiene el modelo de la PSF extraída en formato *.psf*, una tabla FITS binaria [28]. Este archivo se puede usar para la elaboración del catálogo de fuentes final con `SExtractor`, que lo utiliza para corregir los efectos de la PSF para diferentes parámetros. Para esto `SExtractor` sólo acepta archivos en formato *.psf*, que es el motivo por el cuál no se puede trabajar directamente con la imagen en formato FITS que proporciona el SDC1 y es necesario modelar la PSF con `PSFEx` primero.

2.2.2. Corrección del *Primary Beam*

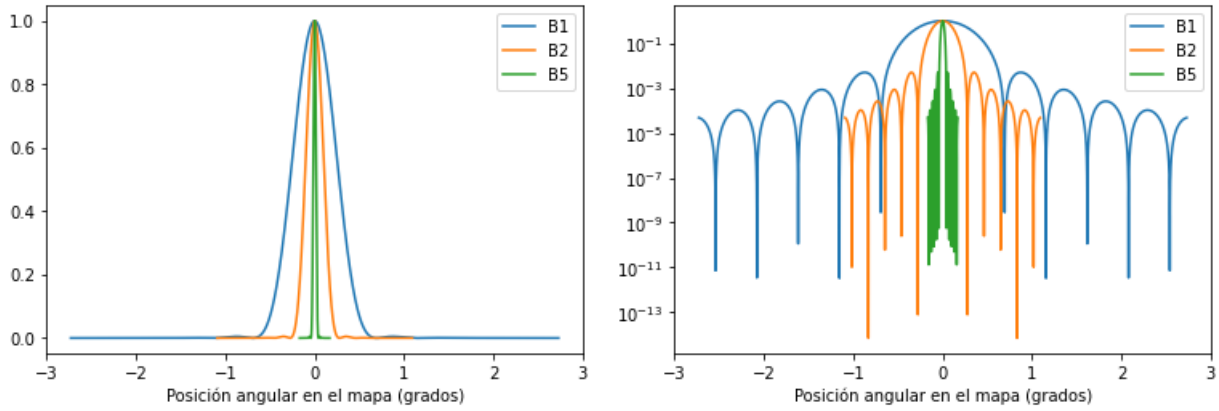
La corrección de los efectos del *Primary Beam* en el flujo medido para el catálogo es uno de los objetivos que exije el *challenge* dentro de la caracterización de las fuentes. Debido a las limitaciones de la simulación, sólo se puede corregir la imagen de forma aproximada, utilizando el *beam* estático y simplificado que se proporciona en el SDC1. Para corregir los efectos de dicho *beam*, basta con dividir la imagen entre el mismo como paso posterior a la detección de fuentes. Sin embargo, las imágenes de los *Primary Beams* a tres frecuencias proporcionadas para el *challenge* tienen un tamaño de 257 píxeles de lado, mientras que las imágenes simuladas de cielo son de un tamaño de 32768 píxeles, por lo que se ha de modelar la forma del *Primary Beam* para poder simular un *beam* del tamaño necesario para corregir la imagen. Dicho modelado parte de la ecuación 11, que modula la forma del *beam* como una función de tipo cosenoidal.

Se ha empleado `Python` para el modelado de los *Primary Beams*. En el anexo se incluye una descripción del código empleado, además de un enlace a un repositorio que lo contiene. Como parámetro libre se ha empleado la FWHM de los *Primary Beams*, medida directamente de las imágenes proporcionadas por el *challenge*. Los beams de las bandas B1, B2 y B5 tienen una FWHM de 0.55 grados, 0.22 grados y 0.034 grados respectivamente. Además, cada *Primary Beam* se ha simulado a la escala dictada por el tamaño de píxel en cada frecuencia. En la figura 6 se pueden ver imágenes de los tres *Primary Beams* simulados.

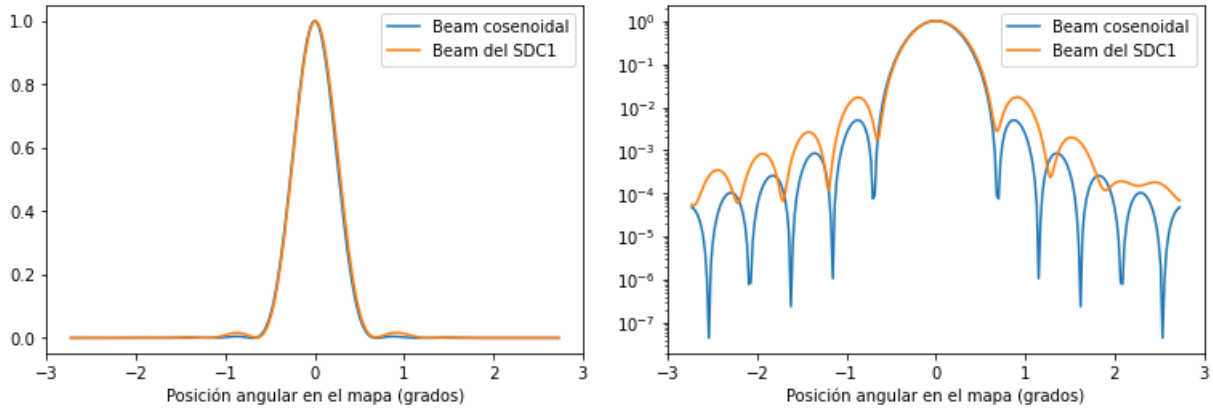
Cabe destacar que, pese a que en la documentación del SDC1 se indica que las imágenes del *Primary Beam* están escaladas dependiendo de la frecuencia, las tres imágenes proporcionadas son del mismo tamaño en píxeles y contienen el mismo *beam* simulado: son las tres iguales. El proceso de reescalado se ha llevado a acabo con el modelado del *Primary Beam* para cada frecuencia.

Una vez simulada la imagen del *Primary Beam*, para corregir sus efectos basta con dividir las imágenes simuladas del cielo a cada frecuencia entre el *beam* de tamaño adecuado, empleando para ello también `Python`. Para que la corrección del *Primary Beam* quede efectuada después de la detección de fuentes, tal y como es preferible para imágenes de un solo *pointing*, se va a trabajar con `SExtractor` en modo dual: una opción de

funcionamiento en la que la detección y análisis de fuentes pueden hacerse en imágenes diferentes, en este caso las imágenes simuladas del cielo con el *Primary Beam* con y sin corregir. Las imágenes del cielo con el *Primary Beam* corregido se incluyen en la figura 7.



(a) Imagen y perfil del *Primary Beam* para las tres bandas.



(b) Comparación entre el *Primary Beam* modelado y el original.

Figura 6: *Primary Beams modelados a partir de una función cosenoidal. En la figura (a) se incluyen los perfiles de los beams ajustados para las tres bandas en escalas lineal y logarítmica. Para poder apreciar las diferencias de tamaño dependientes del tamaño angular del campo de visión a cada frecuencia y el consiguiente tamaño angular de los píxeles en cada banda, se han representado los beams sobre una escala angular. De hacerse lo mismo en una escala de píxeles, los tres beams aparecerían superpuestos e iguales (todos los beams tienen el mismo tamaño en píxeles, pero los píxeles tienen tamaños diferentes para cada frecuencia porque las imágenes de distintas bandas no tienen las mismas dimensiones angulares). En la figura (b) están representados el beam original del SDC1 y su equivalente modelado para la banda B1. Con los tamaños angulares correctos, las figuras equivalentes para las bandas B2 y B5 son iguales (por eso no se han incluido). La imagen en escala logarítmica facilita el estudio de la calidad del ajuste que hace la función para los lóbulos a los lados del pico principal. Aunque no se trate de un ajuste perfecto, la morfología completa del beam queda mejor descrita que con funciones de otro tipo, como por ejemplo una gaussiana.*

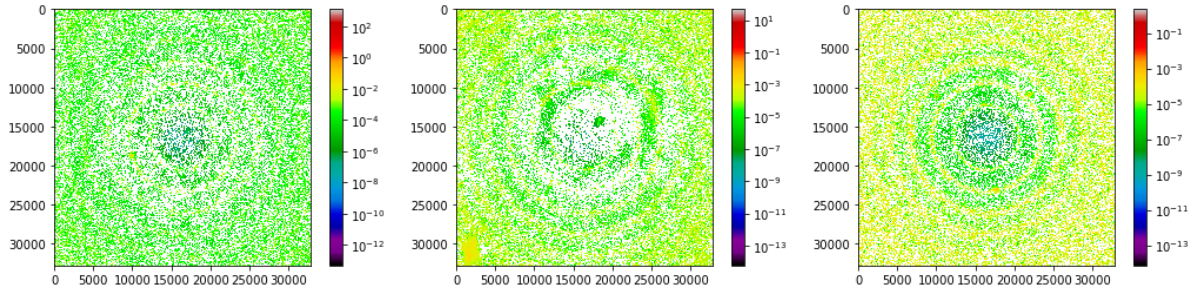


Figura 7: *Imágenes del cielo en las bandas B1, B2 y B5 con la corrección del Primary Beam correspondiente aplicada. Se observa que las imágenes quedan marcadas con un patrón en forma de anillos. Esto es consecuencia del modelado del Primary Beam, y se estudiará en profundidad en la discusión.*

2.3. Detección de fuentes: uso de `Sextractor`

El primer objetivo del *challenge* es la detección de fuentes en las imágenes simuladas del cielo, y para atacarlo se va a emplear el programa `SExtractor`. Como se ha introducido brevemente al hablar de `PSFEx`, `SExtractor` es un programa que elabora un catálogo con diferentes categorías a partir de una imagen astronómica [29]. `SExtractor` puede trabajar sobre una imagen en formato *.fits* en la que detecta y mide las características de las fuentes presentes, o puede emplearse en modo dual que, como se ha establecido previamente, es más adecuado dada la naturaleza del proceso de corrección del *Primary Beam*.

`SExtractor` lleva a cabo la detección y el análisis de fuentes de forma totalmente automática, el usuario sólo ha de configurarlo. Primero, el programa construye un modelo del fondo de la imagen y, posteriormente, extrae y filtra los píxeles que se corresponden a objetos que destacan contra dicho fondo mediante un proceso de segmentación a partir de un umbral. Las características pedidas en la configuración se devuelven en forma de catálogo. Para el modo dual en que se va a trabajar, `SExtractor` detectará las fuentes en la imagen del cielo con el *Primary Beam* sin corregir, y medirá los parámetros en la imagen con el *beam* corregido (las imágenes incluidas en la figura 7 para cada banda).

Las categorías que puede extraer `SExtractor` de cada objeto se clasifican en tres grupos: parámetros isofotales, *windowed positional parameters* o parámetros de posición 'aventanados', parámetros de fotometría de apertura y parámetros de ajuste a un modelo. Los parámetros isofotales derivan de la isofota del objeto, es decir, de una zona con el mismo brillo (en este caso, los píxeles detectados para un objeto por encima de un umbral determinado) y se miden directamente de la imagen detectada del objeto con la corrección de fondo. Incluyen parámetros tanto posicionales como fotométricos, además del parámetro `CLASS_STAR`, empleado en el catálogo para la elaboración del modelo de la PSF. Los *windowed parameters* son únicamente parámetros posicionales, y en vez de medir directamente sobre los píxeles detectados para un objeto, ajustan la fuente a una ventana circular gaussiana y los valores pedidos se integran dentro de esa zona. Los parámetros de

fotometría de apertura incluyen sólo parámetros fotométricos que intengran los parámetros pedidos dentro de una apertura de tamaño fijo o variable que puede ser definida por el usuario. Los parámetros de esta categoría sólo se han empleado para la obtención de la PSF. Por último, `SExtractor` ofrece una serie de parámetros posicionales y fotométricos que ajustan los píxeles detectados para cada objeto a diferentes modelos morfológicos. Entre las opciones se cuentan el ajuste a un modelo con el fondo plano, el ajuste a todos los objetos como fuentes puntuales, un ajuste a un modelo que describe los objetos como discos exponenciales y, por último, un ajuste que modela los objetos como esferoides con perfil de Sérsic. Este último es de especial interés en el presente trabajo. Además, se incluyen parámetros que corrigen los efectos de la PSF si se proporciona a `SExtractor` un modelo de la función en forma de archivo *.psf* [29]. Se ha estudiado la lista de parámetros de `SExtractor` con el objetivo de ajustar la detección de fuentes para que su caracterización sea lo más directa posible. El catálogo final ha de contener las entradas reflejadas en la tabla 2, y para ello se ha establecido la siguiente correspondencia.

Se ha identificado la categoría 1 (ID) con el parámetro de `SExtractor ID_PARENT`, se trata simplemente de una forma de identificar la fuente.

Para las coordenadas de posición del centroide (categorías 4 y 5) se han empleado los parámetros `ALPHAWIN_SKY` y `DELTAWIN_SKY`, correspondiendo con la ascensión recta y declinación respectivamente. Se trata de parámetros *windowed*, obtenidos de encapsular el objeto dentro de una ventana gaussiana. `SExtractor` ofrece también parámetros de posición dentro de las categorías de parámetros isofotales y de parámetros ajustados a un modelo. Pese a que existen parámetros posicionales tanto dentro del ajuste a un perfil Sérsic como con corrección de la PSF incluida cuyo uso pudiera parecer más adecuado, se ha comprobado que, para el caso de las imágenes del *challenge*, los *windowed parameters* proporcionan una posición más acertada al compararse con las posiciones de los objetos de los *true catallogs*. El sufijo `_SKY` indica que las posiciones se proporcionan en referencia al WCS (*Worlds Coordinate System*) de la imagen sobre la que se hacen las detecciones, por lo que su uso no implica transformaciones entre sistemas de coordenadas y facilita la elaboración del catálogo. `Sextractor` indica ambas magnitudes en grados, la misma unidad que han de llevar en el catálogo final.

La densidad de flujo integrada (entrada 6 del catálogo final) se ha identificado con la categoría `FLUX_PSF`, uno de los parámetros para los que `SExtractor` ofrece una corrección de la PSF integrada gracias al uso de `PSFEx`. Lamentablemente, esta es la única categoría de `SExtractor` con corrección de PSF que es adecuada para incluirse en el catálogo: el resto de parámetros disponibles son posicionales y, como se ha establecido, menos precisos que sus equivalentes *windowed*. `SExtractor` devuelve los flujos en las unidades de la imagen de entrada. Como las imágenes del *challenge* tienen unidades de Jy/beam [6], estas son también las unidades del flujo medido por `SExtractor`. En el catálogo final el flujo ha de estar en Jy, por lo que hay que cambiarlas teniendo en cuenta el área de la fuente y la resolución de su haz. Para el cambio de unidades de esta magnitud, es necesario conocer el área en píxeles que ocupa la isofota del objeto por encima del umbral de detección, que puede obtenerse directamente del parámetro `ISOAREA_IMAGE`, devuelto en píxeles al

cuadrado (tal y como india el sufijo _IMAGE).

La obtención de la categoría 7, la fracción de flujo del bulbo frente al flujo del objeto completo o *core_frac*, es un procedimiento más laborioso. No existe un parámetro de SExtractor que se pueda identificar directamente con esta magnitud, por lo que se ha ideado el siguiente proceso para estimarla. Uno de los parámetros de configuración de SExtractor (en los que se profundizará más adelante) es el DETECT_THRESH o *detection threshold*, que es el número de desviaciones estándar (σ) del valor de brillo de un píxel en relación al RMS (*Root Mean Square*) del fondo necesarias para que un pixel pueda ser parte de una detección. Es un parámetro que proporciona el usuario. Para un mismo objeto, su análisis y detección con distintos *thresholds* supone que el área y flujos delimitados no sean iguales: con un *threshold* más alto sólo se detecta y analiza la región más brillante del mismo, que se puede identificar con el bulbo. Por tanto, pasando SExtractor sobre una imagen con un *threshold* bajo primero y con uno más alto después, se pueden obtener flujos para la galaxia completa y para su bulbo aislado de forma estimada, e identificar su cociente como el *core_frac*. Ejecutando SExtractor con este *threshold* más alto, se pueden obtener también la ascensión recta y declinación del *core*, las categorías 2 y 3. Cabe destacar que esta es solo una de las formas posibles de estimar los parámetros del catálogo asociados al *core* como elementos morfológicos separados. Se trata de una solución rápida y sencilla que ha proporcionado buenos resultados, pero viene motivada por las limitaciones en el método, y se trata en realidad de un problema complejo.

Para construir el catálogo final, por tanto, se han de obtener dos catálogos como paso previo: uno con un *threshold* más alto para los parámetros asociados al *core*, y uno con el umbral más bajo para obtener las categorías correspondientes al centroide. Esto presenta un problema, dado que al subir el umbral de detección no sólo se confina la región de flujo medido para un objeto, sino que también pueden perderse las fuentes menos brillantes. Para armar el catálogo final hay que combinar las categorías de ambos catálogos, y si estos no tienen el mismo número de entradas porque detectan una cantidad diferente de objetos este proceso es muy laborioso. Por suerte, SExtractor ofrece un modo de trabajo que permite relacionar ambos catálogos mediante una serie de parámetros de correlación: el modo ASSOC.

Primero, se elabora el catálogo con el umbral más alto, pues el número de fuentes detectadas está limitado por el *threshold* más alto. Dicho catálogo ha de incluir los parámetros XWIN_IMAGE y YWIN_IMAGE, que indican la posición del objeto detectado en píxeles. A partir de estas dos entradas, se ha de crear un archivo de texto (en formato *.txt*) que las contenga como columnas. Al generar el catálogo de umbral más bajo, dicho archivo se puede incluir en la configuración (de una forma similar a cómo se incluye el modelo de la PSF obtenido de PSFEx) para indicar a SExtractor que sólo ha de buscar objetos en las posiciones contenidas en ese archivo. Este segundo catálogo ha de incluir el parámetro VECTOR_ASSOC para que la correlación funcione. Se trata de un vector que relaciona los objetos de ambos catálogos, por lo que además será útil a la hora de ordenar todas las categorías para el catálogo final. De esta forma, se puede relacionar ambos catálogos fácilmente, aunque no es un modo de funcionamiento perfecto y en ocasiones el

número de objetos de ambos catálogos no se corresponde de forma exacta. El parámetro VECTOR_ASSOC permite, sin embargo, seleccionar las entradas de una forma eficiente.

Las categorías relacionadas con la morfología de las detecciones, es decir, los semiejes mayor y menos de la elipse junto con su ángulo de posición (categorías 8, 9 y 10 en la tabla 2) se obtienen del ajuste a un modelo esferoide con perfil de Sérsic. Dado que la mayoría de los objetos del catálogo son SFGs y su forma para el SDC1 se ha simulado a partir de un perfil de Sérsic, es el ajuste más preciso para un gran número de fuentes. A partir de los parámetros SPHEROID_REFF_WORLD y SPHEROID_ASPECT_IMAGE, se pueden obtener los semiejes mayor y menor de la elipse que describe la forma del objeto. SPHEROID_REFF_WORLD está en grados, y las categorías 8 y 9 han de indicarse en segundos de arco. El parámetro SPHEROID_THETA_SKY se puede identificar directamente con el ángulo de posición (categoría 10), y se obtiene en grados y en coordenadas del WCS.

Como la morfología de los objetos se obtiene de un ajuste a un perfil exponencial de Sérsic, la categoría 11 vale 3 para todos los objetos del catálogo. La clasificación de los objetos en clases (categoría 12) se efectúa de forma posterior a la detección y caracterización de fuentes y cuenta con su propia sección en este trabajo. De esta forma, los parámetros que han de obtenerse de SExtractor en los dos catálogos de umbrales diferentes necesarios para construir el catálogo final quedan reflejados en la tabla 6.

Parámetros catálogo de umbral bajo	Parámetros catálogo de umbral alto
ID_PARENT	ID_PARENT
ALPHAWIN_SKY [grados]	ALPHAWIN_SKY [grados]
DELTAWIN_SKY [grados]	DELTAWIN_SKY [grados]
ISOAREA_IMAGE [píxel^2]	ISOAREA_IMAGE [píxel^2]
FLUX_PSF [Jy/beam]	FLUX_PSF [Jy/beam]
SPHEROID_REFF_WORLD [grados]	XWIN_IMAG [grados]
SPHEROID_ASPECT_IMAGE	YWIN_IMAGE [grados]
SPHEROID_THETA_SKY [grados]	
VECTOR_ASSOC	

Tabla 6: *Parámetros que se han incluido en la configuración de* ***SExtractor*** *para la obtención de los dos catálogos necesarios en la elaboración del catálogo final.*

Una de las características más representativas de SExtractor es su gran versatilidad y la libertad que ofrece a la hora de ser configurado. Los parámetros de configuración son muy personalizables, y así su ajuste influye mucho en la calidad de los resultados. Determinar la mejor forma de configurar SExtractor ha sido una de las partes clave del trabajo, y se han hecho numerosas pruebas hasta dar con una configuración que proporcionase unos buenos resultados. Aunque inicialmente se basara la configuración del programa en el procedimiento descrito por los participantes del *challenge* de equipo del IPM, la versión final del archivo de configuración es muy diferente y, como se expondrá en la sección correspondiente, proporciona mejores resultados.

CATALOG_NAME	*catalogo.cat*	El nombre del catálogo de salida.
CATALOG_TYPE	ASCII_HEAD	El formato del catálogo de salida. Para construir el catálogo descrito en la sección que detalla el uso de PSFEx, el formato ha de ser FITS_LDAC
PARAMETERS_NAME	*parametros.param*	Nombre del archivo en formato *.param* en que se incluyen la lista de parámetros pedidos para el catálogo, los que se indican en la tabla 6. Se ha de correr SExtractor dos veces: una con los parámetros de umbral alto y otra con los de umbral bajo.
DETECT_MINAREA	3	El número mínimo de píxeles por encima del umbral necesarios para considerar una detección. Dado que las imágenes del SDC1 incluyen objetos compactos con un tamaño cuyo eje mayor puede ser más pequeño de tres píxeles, este área es adecuada para asegurar que las fuentes compactas quedan incluidas en el catálogo, pero minimizando los efectos del ruido.
DETECT_THRES	Diferentes valores, se discutirán más adelante	Umbral de detección de brillo, en σs. Para el catálogo con las entradas de los *cores*, tiene un valor más alto, y para el de los centroides más bajo. Se ha escogido diferentes valores tras estudiar el comportamiento de catálogos obtenidos con umbrales distintos con respecto a los *true catalogues*. Sus valores se detallarán al discutir los resultados.
FILTER	Y	Si se ha de aplicar un filtro o no (Y/N). Se ha comprobado que la presencia de un filtro gaussiano con una FWHM de 2 píxeles atenúa los efectos del ruido de fondo simulados en la imagen y mejora la precisión del catálogo.
FILTER_NAME	*gauss_2.0_5x5.conv*	Nombre del filtro proporcionado. Se ha utilizado uno de los filtros por defecto que proporciona SExtractor, un kernel de dimensiones 5x5 con una gaussiana de 2 píxeles de ancho.
DEBLEND_MINCONT	0.005	El contraste mínimo para la selección de fuentes. Se ha elegido conforme a la configuración descrita por el IPM para la elaboración del catálogo y participación en el SDC1 [9].
SEEING_FWHM	1.5	FWHM del *Seeing* presente en las imágenes en arcsec. Aunque las imágenes del SDC1 no incluyen errores de calibración, la documentación de SExtractor recomienda establecer en la configuración un *seeing* similar al ancho de la PSF, por lo que se ha configurado con valores de 1.5, 0.6 y 0.0913 para las bandas B1, B2 y B5 respectivamente.
BACKPHOTO_TYPE	LOCAL	De nuevo, emulando la configuración del IPM, se ha elegido que el modelado del fondo se haga de manera local frente a una forma global debido a la densidad de fuentes presente en las imágenes [9].
ASSOC_NAME	*assoc.txt*	Archivo que contiene las posiciones de los objetos en el catálogo de umbral más alto y garantiza el funcionamiento del modo ASSOC. Se ha de incluir al configurar SExtractor para la obtención del catálogo con *threshold* bajo. Para el catálogo con umbral alto, esta categoría se ha de dejar vacía.
ASSOC_PARAMS	1,2	Columnas del archivo *assoc.txt* en las que se indica la posición de los objetos (en este caso, las dos únicas columnas).
ASSOCCOORD_TYPE	PIXEL	Unidades de la posición de los objetos en el archivo *assoc.txt*.
ASSOC_RADIUS	4	Radio de búsqueda máximo para la localización de objetos en el catálogo a partir del archivo *assoc.txt*. Se ha escogido tras comprobarse que era el tamaño más adecuado para que el modo de búsqueda no contase más que unos pocos objetos de más.
ASSOC_TYPE	NEAREST	Indica cómo discriminar entre objetos cuando SExtractor detecta más de una fuente en la posición propuesta por *assoc.txt* (en este caso, la más cercana en posición).
PSF_NAME	*modelo.psf*	Nombre del archivo *.psf* elaborado por PSFEx que contiene el modelo de la PSF para su corrección.

Tabla 8: *Modelo del archivo de configración de* ***SExtractor*** *para la obtención de los dos catálogos necesarios en la elaboración del catálogo final.*

En la tabla 8 se expone de forma detallada cómo ha de configurarse SExtractor para

el correcto funcionamiento de los mecanismos descritos: el modo ASSOC de correlación entre catálogos, el modo dual para la corrección del *Primary Beam*, la corrección de la PSF, y la obtención de las categorías deseadas. La tabla incluye un ejemplo de archivo de configuración tal y como se ha empleado. Los parámetros que no se reflejan en la misma es porque se ha mantenido su configuración por defecto.

Una vez editado el archivo de configuración, se puede utilizar `SExtractor` en modo dual con un comando del tipo:

```
sex imagenSDC1.fits,imagenconprimarybeamcorregido.fits -c config_file.sex
```

y el programa devuelve un catálogo del nombre especificado en formato *.cat* que contiene las categorías pedidas.

2.4. Caracterización de fuentes

Una vez detectadas las fuentes, es necesario juntar los dos catálogos con umbrales diferentes, reordenarlos y cambiar las unidades de algunas categorías para que el catálogo final tenga el formato pedido para participar en el SDC1, con las categorías finales de la tabla 2. El manejo de los catálogos se ha llevado a cabo mediante `Python`. Los detalles referentes al código utilizado se pueden ver en el anexo: en esta sección se discute su uso desde su significado físico. A continuación, se detalla el proceso que ha de seguirse para elaborar cada categoría.

Empleando los parámetros `VECTOR_ASSOC` en el catálogo de umbral bajo y `XWIN_IMAGE` en el catálogo de umbral alto, se asocian las entradas de un catálogo con otro a partir de su posición en la imagen, para tener ordenados todos los datos correspondientes al bulbo y centroide de un mismo objeto. Una vez hecho esto, se puede proceder a la elaboración de las entradas del catálogo.

La categoría 1, el ID de la fuente, se identifica directamente con el parámetro `ID_PARENT`.

Las entradas 2 y 4 del catálogo, que se corresponden con la ascensión recta (RA) del bulbo y el centroide, se obtienen de los parámetros `ALPHAWIN_SKY` del catálogo con el *threshold* más alto y más bajo respectivamente. El parámetro es registrado en grados por `SExtractor`, así que no es necesario cambiar las unidades en esta categoría. No obstante, las imágenes y catálogos del SDC1 están centrados en una posición tal que $RA = 0$, de forma que hay que desplazar el origen en la categoría obtenida por `SExtractor`. Se ha de cambiar el rango de distribución de la ascensión recta, contenido en el intervalo $[0, 360]$, por uno que abarque $[-180, 180]$, para que las posiciones vayan dadas con respecto al centro de la imagen y se puedan comparar con los valores de los *true catalogues*.

Las entradas correspondientes a las declinaciones del *core* y el centroide son los parámetros `DELTAWIN_SKY` de los catálogos con umbrales alto y bajo respectivamente. El parámetro está en grados y en el sistema de referencia correcto, por lo que ambas categorías se

pueden identificar con su entrada correspondiente en el catálogo final sin modificaciones.

El flujo (categoría 6) se obtiene del parámetro `FLUX_PSF`, que está en las unidades de la imagen, Jy/beam. El *beam* hace referencia al área de la función a la que se puede ajustar el objeto una vez deconvolucionada la PSF. Con el parámetro `ISOAREA_IMAGE` se puede obtener la superficie en píxeles detectada por encima del umbral para el objeto, es decir, el área de su *beam* en píxeles. De esta forma, se puede obtener el flujo en las unidades deseadas (Jy) de la forma:

$$\text{flujo}\,[\text{Jy}] = \texttt{FLUX\_PSF}\,[\text{Jy/beam}] \cdot \texttt{ISOAREA\_IMAGE}\,[\text{beam/píxeles}^2] \cdot A_p\,[\text{sr}], \tag{12}$$

donde A_p es el área de 1 píxel en estereorradianes. Este factor varía dependiendo de la banda de frecuencia, pues el tamaño del mapa y, por consecuencia, el de cada píxel individual, es diferente para las tres bandas. Así, para cada banda este tiene un valor de:

$$\begin{aligned} A_{p1} &= \left(\frac{2\pi \cdot 0{,}6}{3600}\right)^2 \\ A_{p2} &= \left(\frac{2\pi \cdot 0{,}24}{3600}\right)^2 \\ A_{p5} &= \left(\frac{2\pi \cdot 0{,}037}{3600}\right)^2 \end{aligned} \tag{13}$$

El flujo que procede del catálogo con el umbral más bajo y con las unidades correctas se añade al catálogo final. La corrección de unidades para el flujo del catálogo de umbral alto, es decir, el flujo del *core*, se hace de forma análoga, y ambas cantidades se pueden dividir para añadir su cociente al catálogo como la categoría 7 (el *core_frac*).

A partir de los parámetros `SPHEROID_REFF_WORLD` y `SPHEROID_ASPECT_IMAGE` del catálogo con umbral más bajo, se pueden obtener las categorías 8 y 9: los semiejes mayor y menor de la elipse que describe la forma del objeto. Como son parámetros que se corresponden con el ajuste exponencial a un perfil de Sérsic, se tiene que el área total del objeto:

$$A = \pi R_{eff}^2, \tag{14}$$

donde R_{eff} es el radio efectivo del esferoide descrito por el modelo de perfil de Sérsic. Esta cantidad se puede identificar con el parámetro `SPHEROID_REFF_WORLD` que proporciona `SExtractor`, que es el radio efectivo del esferoide ajustado por el modelo. Sabiendo que el área de una elipse:

$$A = \pi\, bmaj \cdot bmin \tag{15}$$

Se pueden igualar ambas cantidades. Conociéndose también la razón de tamaño entre los dos semiejes ($bmaj$ y $bmin$), que viene dada por el parámetro `SPHEROID_ASPECT_IMAGE`, se puede plantear el sistema:

$$\begin{aligned} (\texttt{SPHEROID\_REFF\_WORLD} * 3600)^2 &= bmaj \cdot bmin \\ \texttt{SPHEROID\_ASPECT\_IMAGE} &= \frac{bmaj}{bmin}, \end{aligned} \tag{16}$$

que permite obtener las categorías bmaj y bmin en arcsecs, las unidades requeridas en el catálogo final. El ángulo de posición (categorías 10) se puede identificar directamente con el parámetro `SPHEROID_THETA_SKY` sin necesidad de hacer cambios de unidades. Se ha empleado un ajuste a perfil exponencial de Sérsic para modelar la morfología de todos los objetos, y eso se ha de reflejar en el catálogo señalando con un 3 todos los objetos en la categoría 11.

La última entrada del catálogo esta reservada para la clasificación de fuentes en distintas categorías. Para distinguir los objetos entre ellos se han estudiado sus flujos a lo largo de las tres bandas de observación, con el objetivo de identificar qué mecanismos de emisión predominan en el espectro de cada objeto a partir de su presencia en cada una de las bandas y de la variación de su flujo. El método propuesto se detalla en la sección que sigue.

2.5. Clasificación de fuentes

El último objetivo del *challenge* es la clasificación de los objetos previamente detectados y caracterizados, y se ha de incluir como la entrada número 12 de los catálogos finales para las tres bandas. Las clases posibles son 1, 2 y 3 para SS-AGNSs, FS-AGNs y SFGs respectivamente.

Se ha simplificado el proceso de clasificación de fuentes usándose como parámetro de estudio su flujo a lo largo de las diferentes bandas. Cada clase de objeto tiene un espectro de emisión diferente, y su presencia a lo largo de las diferentes frecuencias y cómo el flujo que emite varía entre las mismas puede proporcionar información sobre los mecanismos que lo dominan. Para llevar a cabo la clasificación, primero se han agrupado las detecciones en función de en qué bandas aparecen. Así, se han formado grupos de objetos que tienen presencia en las tres bandas de frecuencia, en dos de ellas o en sólo una. A partir de los datos de flujo que se tienen para cada objeto, se han definido grupos de diferentes 'colores' con una clase probable asociada a cada grupo. Esto ha sido posible mediante el uso de los datos del *true catalogue*: se emplean con la misma finalidad con la que que los participantes del *challenge* disponían de los *training sets*, pero de una forma más completa. En un contexto diferente al del *challenge*, podrían emplearse datos reales de simulaciones previas o modelos. El método planteado se trata de un procedimiento muy general y con posibilidades de mejora, se plantea como posible vía a desarrollar para conseguirse en un futuro una implementación más eficiente.

Para desarrollar el procedimiento, se han agrupado los objetos de los *true catalogues* según su aparición en las bandas. Estudiando los flujos de cada objeto y sabiéndose además cuál es su clase, el objetivo es detallar toda la casuística que relaciona el flujo con la clase basándose en estos datos. Se ha empleado `Python` (el código y su descripción van incluidos en el anexo) para agrupar los objetos del *true catalogue* según en qué bandas de frecuencia aparecen, empleando para ello su posición y tamaño. Una vez agrupados, se han estudiado sus flujos. A continuación, se detalla la distribución de los objetos en función de su clase y flujo para cada subgrupo.

El primer grupo está conformado por objetos que aparecen en las tres bandas: B1, B2 y B5. Para estos objetos se conocen datos del flujo en tres frecuencias, por lo que para estudiar su distribución se definen los colores:

$$color\ 1 = \frac{flujo(B1)}{flujo(B2)}$$
$$color\ 2 = \frac{flujo(B2)}{flujo(B5)} \tag{17}$$

Representando para cada objeto del *true catalogue* sus colores junto con la clase que tiene asignada se puede estudiar la distribución de las clases con respecto a los colores de flujo, y así definir grupos de clases en función de su posición en el diagrama. Delimitando las zonas en las que se concentra la presencia de cada clase para los objetos del *true catalogue*, se pueden colocar posteriormente los objetos detectados para el *challenge* en el mismo diagrama y asignarles la clase que se corresponde con la zona que ocupan en el diagrama de color. En la figura 8 se incluyen los objetos de este grupo y cómo se han definido las zonas en las que predomina la presencia de cada clase.

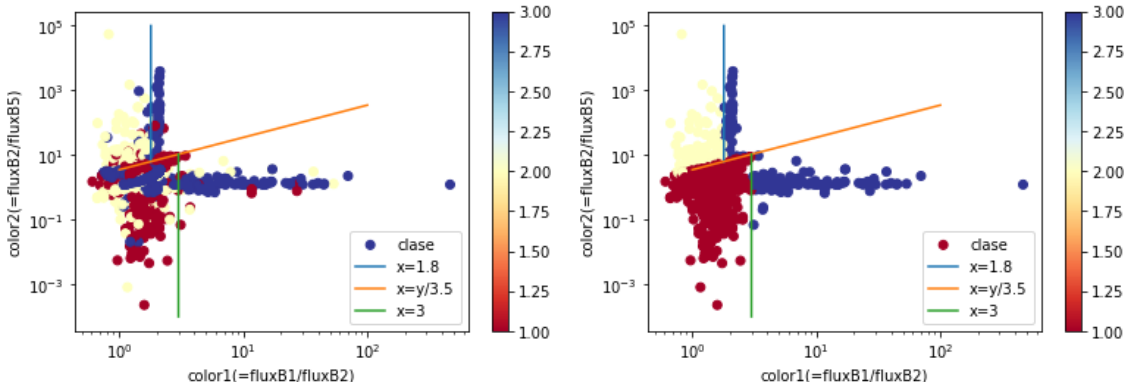
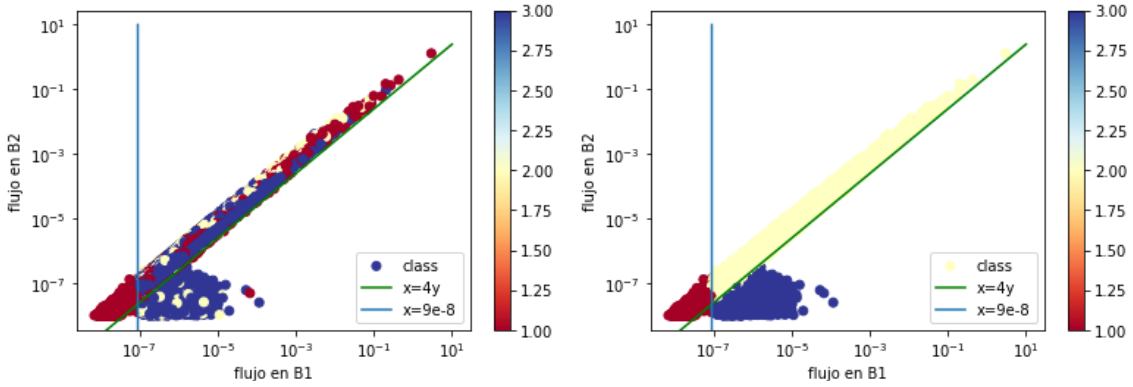
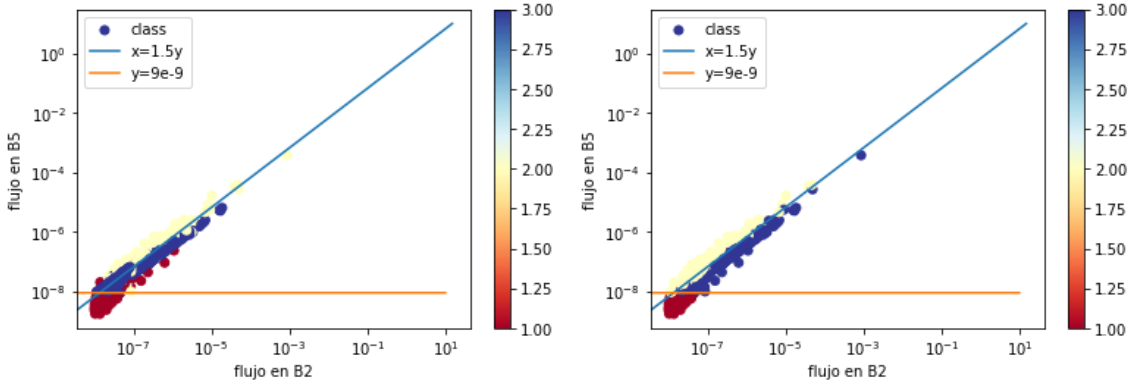


Figura 8: *Distribución de objetos que aparecen en las tres bandas en el diagrama de color. Se representan para cada objeto los dos colores definidos en la ecuación 17. La barra de color indica la clase de cada objeto. La figura de la izquierda muestra los datos con sus clases asociadas en el true catalogue. Las rectas dibujadas definen las zonas en las que se ve mayor presencia de objetos de una misma clase. Las delimitaciones no son perfectas, se trata de una clasificación aproximada. Una vez identificadas las regiones en las que se dividen los grupos, se puede asignar una clase a cada objeto en función de su zona. El diagrama de la segunda figura muestra los mismos objetos pero con la clase asignada en función de los límites de zona definidos para la figura anterior.*

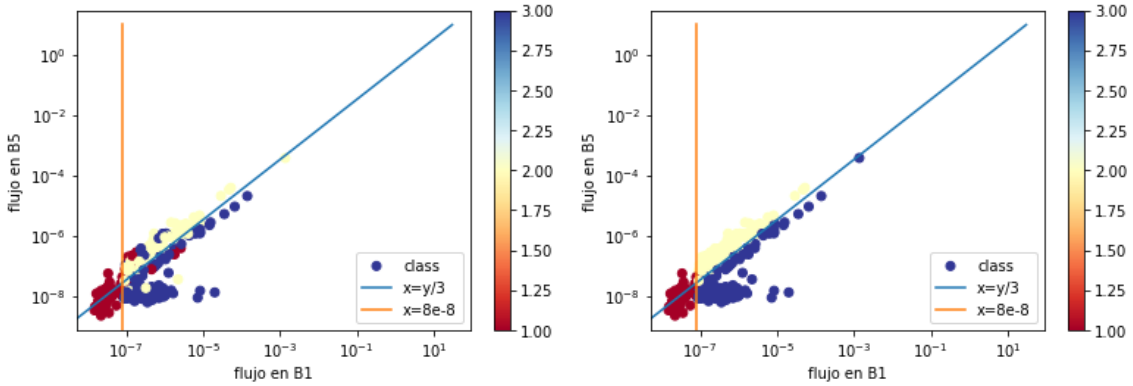
Para los objetos que aparecen en dos de las bandas se ha seguido un procedimiento similar. Como para estos grupos sólo se tienen datos del flujo en dos frecuencias, en el diagrama de colores se representa el flujo en una banda frente al flujo en la otra. De forma análoga al procedimiento descrito para los objetos que están en tres bandas, se estudia la distribución de las clases en función de su posición en el diagrama de flujos y se delimitan las zonas en las que predomina cada clase para así definir los grupos. Los grupos establecidos se pueden estudiar en la figura 9.



(a) Objetos que aparecen en las bandas B1 y B2



(b) Objetos que aparecen en las bandas B2 y B5.



(c) Objetos que aparecen en las bandas B1 y B5.

Figura 9: *Diagramas de flujo para los diferentes grupos de objetos. Se representa el flujo del objeto en una banda frente al flujo del mismo objeto en la otra banda (en* Jy*). La barra de color indica la clase de cada objeto. Las figuras de la izquierda muestran los objetos de cada grupo con las líneas que delimitan las áreas de distribución de cada clase, y las de la izquierda cómo queda el diagrama con la clase que se le reasigna a cada objeto una vez definidas las zonas. Para las figuras (b) y (c), se ve la tendencia de los datos de forma clara: los objetos de clase 1 se agrupan en la parte inferior del diagrama, mientras que los objetos de clases 2 y 3 se sitúan encima y debajo de una recta con una pendiente dada. Para los datos de la figura (a) la disposición no está tan clara, por lo que se han delimitado las zonas usando como guía la tendencia general de las figuras (b) y (c): los objetos de clase 1 en la parte inferior del diagrama, con una recta que separa las zonas en que se concentran 2 y 3.*

Una vez asignada una clase a todas las posibles colocaciones para un objeto que aparece en tres bandas y en dos bandas, queda por determinar el tratamiento de los objetos que aparecen en una sola banda.

Como los mapas de los tres canales tienen tamaños diferentes, sólo los objetos contenidos en el área en común de dos o tres de las bandas aparecen en los grupos definidos anteriormente. Por ejemplo, la mayoría de objetos de la banda B1 (la más grande) sólo aparecen en esta banda, porque únicamente los objetos situados en la zona central del mapa pueden aparecer también en las otras dos. Como resultado, un gran porcentaje de objetos de las tres bandas aparecen en una sola banda. Para estos objetos sólo se tienen datos de flujo a una frecuencia, por lo que no se puede construir un diagrama de colores como en los procedimientos descritos hasta ahora, y esto presenta la limitación principal de este método. Se ha determinado que, debido a los mecanismos de emisión de las diferentes clases de objetos, en cada banda es más probable la aparición de algunos de ellos:

- Los objetos de clase SS-AGN tienen un espectro de emisión con un flujo que decae rápidamente con el aumento de la frecuencia, por lo que los objetos que aparecen sólo en la banda B1 se han clasificado como tal, la clase número 1.
- El espectro de emisión de las SFGs está dominado por la radiación que proviene de las nubes de polvo que emiten con espectro de cuerpo negro con pico en el infrarrojo alto, por lo que se observan sobre todo en los canales con frecuencia alta. Así, los objetos que aparecen sólo en la banda B5 se han etiquetado como SFGs (clase 3).
- Por último, los objetos que aparecen sólo en la banda B2 se han identificado con la clase 2, FS-AGNs, dado que su espectro de emisión es más plano con respecto a la variación de frecuencia.

Esta clasificación es muy aproximada, la asignación de clases puede mejorarse teniendo en cuenta otros parámetros además del flujo.

2.6. Funcionamiento del *scorer*

Con la finalización del SDC1 y la publicación de los resultados obtenidos por los grupos participantes, se hicieron también públicos los *true catalogues* y se reveló el método empleado para la evaluación de los catálogos. El método de puntuación o *scorer* se puede utilizar como librería de `Python` [10].

El *scorer* funciona de la siguiente manera. Se compara cada catálogo de entrada con el *true catalogue* de la frecuencia correspondiente y se determina cuántos objetos se han detectado, caracterizado y clasificado con éxito. Primero, se hace un *match* posicional entre ambos catálogos. Dada la densidad de objetos en las imágenes, para cada objeto del *true catalogue* se obtienen múltiples candidatos dentro del catálogo propuesto. Para determinar cuál es el más adecuado, se estudian el tamaño y flujo de las fuentes posibles

para seleccionar el *match*. Se define un umbral máximo que los errores de ambas magnitudes no pueden superar para ser seleccionados [9].

Una vez hechas las correspondencias entre catálogos, se otorga una puntuación al catálogo. Evaluando los errores de cada uno de los parámetros pedidos, se determina lo bien que se ajustan como entradas del catálogo, y a cada *match* se le otorga un peso final en el cómputo total del catálogo, que vale un máximo de 1 y un mínimo de 0. La suma ponderada de todos los *matches* indica la puntuación final del catálogo o *score* [27].

Del *scorer* se pueden también obtener diferentes parámetros, como por ejemplo, el número de detecciones falsas, el número total de matches o el número de entradas del catálogo que se han considerado no válidas. También puede devolver un *dataframe* con todos los objetos que tienen un *match* en el *true catalogue* y sus características en ambos catálogos.

3. Resultados y análisis

Siguiendo el procedimiento descrito en las secciones anteriores, se han elaborado catálogos de las características pedidas con pares diferentes de umbrales para estudiar qué valores producían un catálogo con mejor comportamiento. Ha sido necesario construir catálogos distintos para familiarizarse con el comportamiento de los datos y mejorar la técnicas de análisis. Finalmente, se ha escogido estudiar los catálogos con las características propuestas en la tabla 9. Los catálogos se han seleccionado por su amplio rango de umbrales, además de por su comportamiento frente al proceso de evaluación. Se han estudiado y evaluado emulando el proceso seguido por SKAO para puntuar las participaciones del SDC1.

	Umbral alto (parámetros del *core*)	Umbral alto (parámetros del centroide)
Catálogo 1	10σ	7σ
Catálogo 2	7σ	5σ
Catálogo 3	5σ	$2{,}5\sigma$

Tabla 9: *Thresholds de los catálogos seleccionados para estudiar su comportamiento según las directrices del SDC1.*

A partir de ahora, se evalúa el comportamiento de los catálogos 1, 2 y 3 emulando en gran medida el método de puntuación del SDC1 a partir del *scorer*. Se estudian tres categorías directamente relacionadas con los objetivos del *challenge*: se evalúan la detección de fuentes, su caracterización y la clasificación de objetos. Se incluyen además material complementario a los ítems propuestos para la evaluación del SDC1.

3.1. Resultados de la detección de fuentes

Una vez obtenidos los 9 catálogos (3 por cada par de umbrales indicados en la tabla 9, como corresponde a los tres canales de frecuencia), se han evaluado mediante el *scorer* del SDC1. De esta forma se obtienen los parámetros necesarios para el estudio de su comportamiento, que aparecen reflejados en la tabla 10.

	Frecuencia	N_d	N_m	N_f	$\tilde{N}_m$	B
Catálogo 1	560MHz	126409	118607	7802	58147.63	50345.63
	1400MHz	45196	43961	1235	21756.84	20521.84
	9200MHz	102	93	9	42.45	33.45
Catálogo 2	560MHz	188657	173064	15593	84179.40	68586.40
	1400MHz	66789	64506	2283	31792.53	29509.53
	9200MHz	164	155	9	72.59	63.59
Catálogo 3	560MHz	266840	234622	32218	113641.96	81423.96
	1400MHz	93520	88924	4596	43726.59	39130.59
	9200MHz	291	282	9	136.92	127.92

Tabla 10: *Parámetros del scorer utilizados para estudiar y puntuar los catálogos. Las categorías se corresponden con: el número de detecciones encontradas en el catálogo propuesto (N_d), el número de matches entre el catálogo de entrada y el true catalogue de la frecuencia correspondiente (N_m), y el número de falsos positivos dentro del catálogo (N_f). $\tilde{N}_m$ se refiere a la suma ponderada de matches (en la que cada uno tiene un peso determinado por lo bien que está caracterizado). Por último, el parámetro B, también denominado score o puntuación, es la suma ponderada de matches, $\tilde{N}_m$, menos el número de detecciones falsas, N_f. En caso de que haya más detecciones falsas que matches con parámetros correctos, el score del catálogo puede ser negativo.*

Se estudia la calidad de la detección de fuentes mediante el análisis de la *completeness* y la *reliability*, que indican cómo de completo es el catálogo basándose en el volumen de detecciones, y cómo de correcto es en función de la precisión de las mismas. Son dos magnitudes que están relacionadas, y es deseable encontrar un catálogo con un buen equilibrio entre su cantidad de detecciones y la calidad de las mismas. Se definen ambas magnitudes como:

$$\begin{aligned} C_{tot} &= \frac{N_{m,\nu1}}{FoV_{\nu1}} + \frac{N_{m,\nu2}}{FoV_{\nu2}} + \frac{N_{m,\nu5}}{FoV_{\nu5}} \\ R_{tot} &= \frac{1}{3}\left[\frac{N_{m,\nu1}}{N_{d,\nu1}} + \frac{N_{m,\nu2}}{N_{d,\nu2}} + \frac{N_{m,\nu5}}{N_{d,\nu5}}\right] \end{aligned} \tag{18}$$

A partir de los parámetros de la tabla 10 se pueden obtener la *completeness* y *reliability* totales para los tres catálogos. Los campos de visión necesarios para la correcta obtención de estos parámetros son de 30.25, 4.84 y 0.112 grados cuadrados para las bandas B1, B2 y B5 [9]. En la tabla 11 se pueden comparar los resultados obtenidos con los de diferentes grupos participantes del *challenge*. Los resultados obtenidos son comparables con los obtenidos por los equipos de ICRAR e IPM por su orden de magnitud.

	C_{tot}	R_{tot}
IPM2	141801	0.07
JLRAT	107705	0.58
ARCIt-CACAO	47766.1	0.83
EngageSKA	45734.9	0.80
Shanghai	33288.6	0.98
Catálogo 3	28646.7	0.9331
Catálogo 2	20432.7	0.9427
ICRAR	18020.1	0.71
Catálogo 1	13834.1	0.9409
IPM	9937.96	0.05
hs	2235.29	0.07
RADGK	0.9256	0.08

Tabla 11: *Completeness y reliability de los catálogos con diferentes umbrales propuestos. Se han integrado en la lista de resultados de los equipos participantes en el SDC1 [9], ordenados en orden decreciente de completeness. Como se ha establecido previamente, el uso de* `SExtractor` *en el presente trabajo emula el procedimiento seguido por la participación del IPM. Para los tres catálogos, se obtiene una completeness un orden de magnitud mayor que la del IPM, y una reliability mejor.*

Comparando los tres catálogos entre sí, se tiene que los catálogos con un umbral más bajo tienen una *completeness* más alta, como corresponde a bajar el *threshold* de detección. A su vez, la *reliability* no disminuye de forma drástica al bajar el umbral de detección del catálogo, de forma que los catálogos 2 y 3 tienen un mejor equilibrio con respecto a los dos parámetros.

Para terminar de evaluar la calidad de las detecciones, se incluyen en la figura 10 imágenes de las fuentes detectadas con su posición en el mapa para las tres bandas con el fin de ilustrar su distribución.

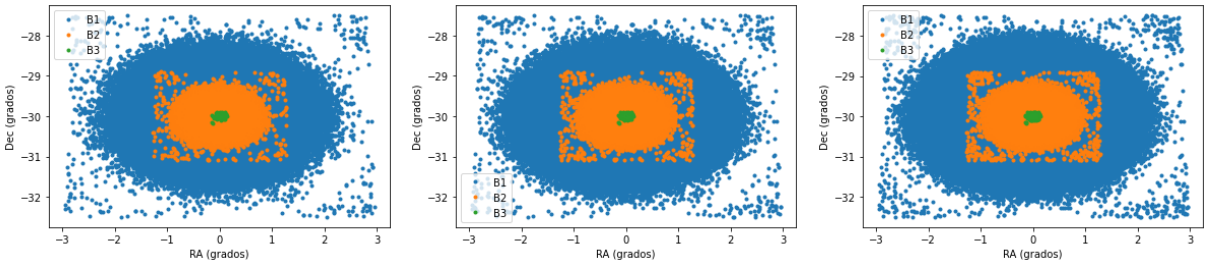


Figura 10: *Posición de los objetos detectados en la imagen para los tres catálogos elaborados. En orden, se corresponden con los catálogos 1, 2 y 3. Para cada una, se observa cómo las detecciones para cada banda están confinadas a la región del espacio contenida por el mapa en cada frecuencia, de forma que sólo los objetos que están en el centro de la imagen pueden ser detectados en las tres bandas. Comparando las tres imágenes entre sí, se puede apreciar cómo aumenta la densidad de detecciones al bajar el umbral del catálogo.*

Del estudio de la figura 10 se aprecia que la densidad de objetos detectada no es homogénea en toda la superficie del mapa. Observando los objetos de cada banda se puede ver que, además, es algo que afecta a las tres frecuencias. Esto se analizará la discusión, pero está relacionado con el proceso de corrección del *Primary Beam*.

3.2. Resultados de la caracterización de fuentes

El segundo objetivo del SDC1 es la caracterización de fuentes. Para evaluar la calidad de los parámetros incluidos en el catálogo, se definen los siguientes errores [9]:

$$\begin{aligned} D_{pos} &= \sqrt{(x-x')^2+(y-y')}/S' \\ D_{size} &= |S-S'|/S' \\ D_{flux} &= |f-f'|/f' \\ D_{cf} &= |cf-cf'|/cf' \\ D_{PA} &= |PA-PA'|/10^\circ \end{aligned} \tag{19}$$

donde los parámetros con una marca de prima se refieren a parámetros del *true catalogue*. S, el tamaño promedio de la fuente, se puede obtener como $S = (bmaj + bmin)/2$ [27]. Además de proporcionar parámetros necesarios para la puntuación de los catálogos como los que se enumeran en la tabla 10, el *scorer* puede elaborar una tabla de datos con todas las entradas de cada objeto en el catálogo que se está evaluando y en el *true catalogue*. Haciendo uso de esa tabla, se puede obtener el error asociado a cada entrada medida a partir de las ecuaciones en 19. En la figura 11 se incluyen histogramas con la distribución de errores estimados para cada magnitud, presentados del mismo modo que los resultados evaluados para el SDC1 de los distintos equipos [9].

Los errores de posición, tamaño y flujo son especialmente relevantes, pues son los parámetros que el *scorer* utiliza para seleccionar el *match* para cada objeto del *true catalogue*, y para discriminar entre todos los objetos candidatos. Por cómo están definidos los errores, todas las distribuciones son asimétricas, y los errores toman sólo valores positivos. La distribución ideal tiene un pico estrecho en el cero, indicando una mayoría de errores pequeños y con muy poca contribución estadística [9].

La distribución del error de posición tiene la forma deseada para una magnitud con errores pequeños: el pico se encuentra sobre cero para los tres catálogos y es estrecho, decreciendo rápidamente al alejarse de ese valor e indicando que las posiciones contienen pocos errores estadísticos.

Para los errores de tamaño, el pico de la distribución se encuentra desplazado del origen. Esto indica un mayor error en las magnitudes medidas, aunque es coherente con los errores del resto de grupos participantes del SDC1, que presentan una distribución del error de tamaño más ancha que la de la posición y alejada del valor cero. En la figura 12 se puede ver la correspondencia entre los tamaños de los semiejes mayor y menor en los tres catálogos obtenidos con su valor en el *true catalogue*. Para un conjunto de *matches* perfectos la correlación entre ambas magnitudes debería ser lineal. Pese a que ambas magnitudes se relacionan de forma coherente, del estudio de la figura se puede ver que, en algunos casos, ambos semiejes son hasta un orden de magnitud más pequeños que su equivalente en el *true catalogue*, lo que se traduce en un error asociado más grande que el de otras categorías.

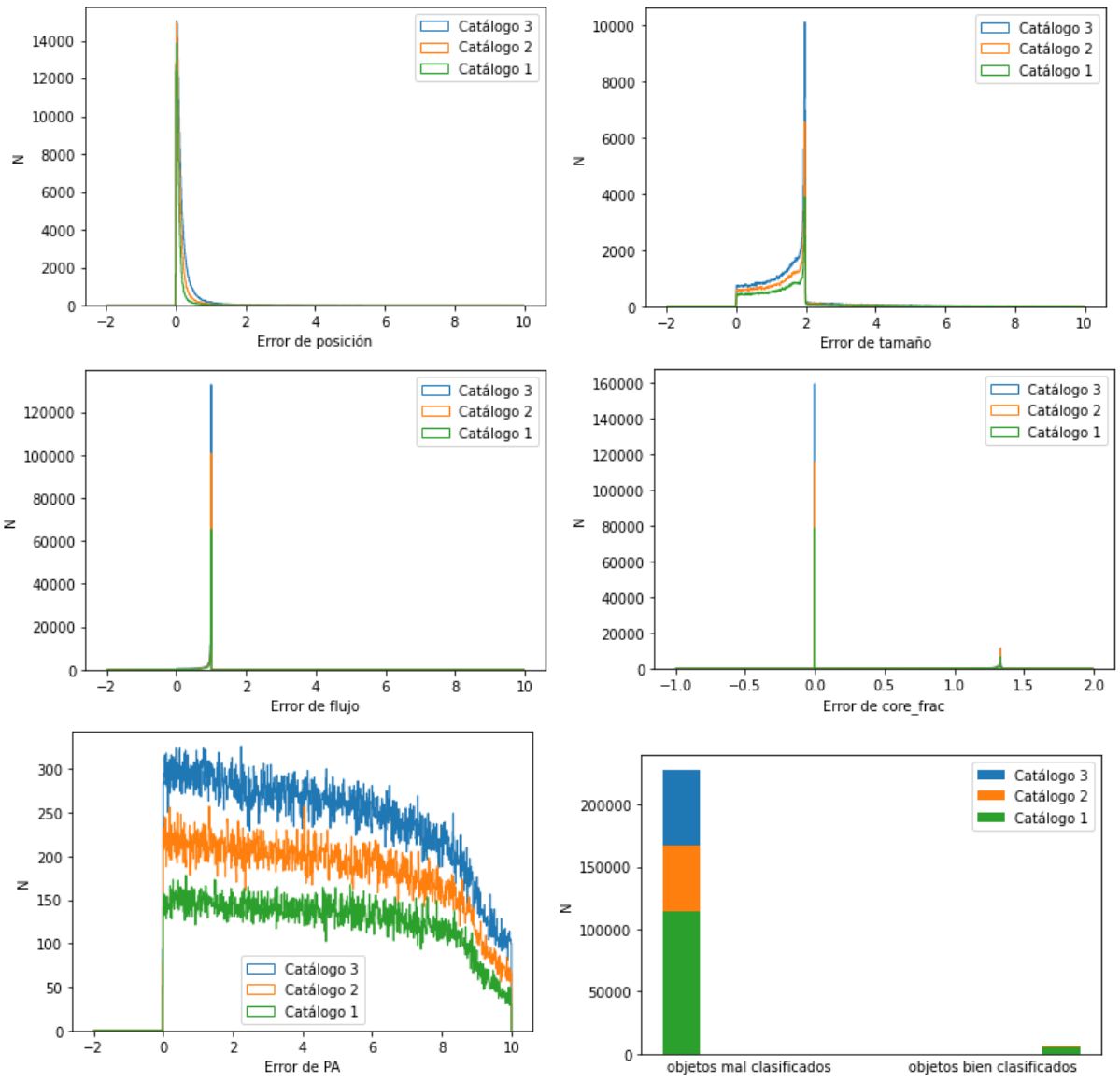


Figura 11: *Distribución de errores de los distintos parámetros de caracterización, obtenidos según las ecuaciones 19. Los datos se corresponden con la banda B1 para los tres catálogos: al ser la frecuencia con imágenes de mayor tamaño contiene más fuentes y es una muestra más grande para obtener distribuciones más completas y precisas, tal y como se ha hecho para evaluar a los participantes del SDC1. Se incluyen también en la última figura los errores de la clasificación de objetos, que se estudiarán en la sección siguiente.*

Pese a no estar situado en el origen, el pico de distribución del error del flujo también es estrecho y toma valores pequeños, por lo que presenta una distribución razonable, similar a las del resto de participantes. La figura 12 incluye también la correspondencia entre valores de flujo para los catálogos elaborados y el *true catalogue*. Igual que en el caso de los tamaños, el flujo adquiere valores un poco más pequeños de los que corresponderían según su *match* en el *true catalogue*, lo que explica el desplazamiento del pico de la distribución de errores.

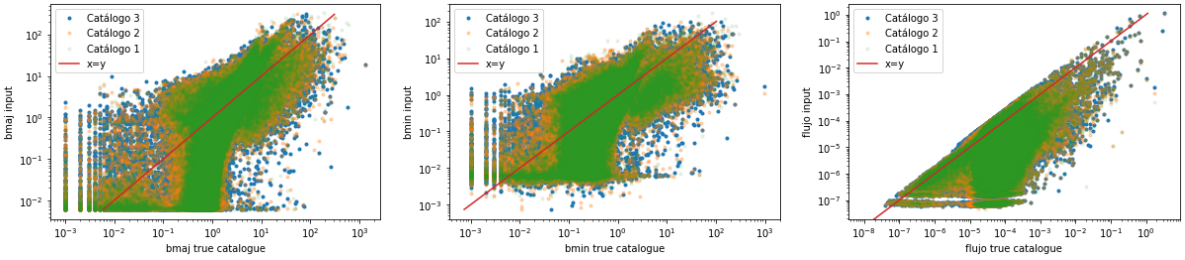


Figura 12: *Correspondencia entre los matches que hace el scorer: valores que toman diferentes parámetros en los tres catálogos propuestos y su equivalente en el true catalogue. Las figuras representan la correlación para los semiejes mayor y menor, y para el flujo de las fuentes.*

El error del ángulo de posición presenta la distribución menos adecuada de todos. Los errores se distribuyen de forma uniforme sin ningún pico claro, lo que indica que su origen es principalmente estadístico. Salvo uno de ellos, todos los participantes del *challenge* obtienen distribuciones similares para el error del PA. Uno de los temas comentados en la evaluación de los equipos del SDC1 por SKAO es, precisamente, que el ángulo de posición es difícil de recuperar para los datos de SKAMid, puesto que la distribución obtenida para los errores es, en general, plana. Se discutirá este tema en profundidad más adelante, pero el estudio de los valores para el ángulo de posición en las entradas del *true catalogue* indica que puede haber algún problema derivado del criterio elegido para establecer el origen a la hora de simular este parámetro.

La distribución de errores para el *core_frac* presenta dos picos diferentes, aunque el que se encuentra sobre el origen es el de mayor tamaño. Esto está relacionado también con la naturaleza de los datos simulados: en el *true catalogue* los diferentes objetos toman 1 o 0 como valor de *core_frac* según sean AGNs o SFGs, por lo que el error de esta magnitud presenta una distribución bimodal para varios de los participantes, como es el caso aquí [9].

Para evaluar la precisión y otorgar una puntuación final a la calidad de la caracterización, se definen los parámetros:

$$\begin{aligned} A_{tot} &= \frac{\tilde{N}_{m,\nu1}}{FoV_{\nu1}} + \frac{\tilde{N}_{m,\nu2}}{FoV_{\nu2}} + \frac{\tilde{N}_{m,\nu5}}{FoV_{\nu5}} \\ G_{tot} &= \frac{B_{\nu1}}{FoV_{\nu1}} + \frac{B_{\nu2}}{FoV_{\nu2}} + \frac{B_{\nu5}}{FoV_{\nu5}} \end{aligned} \tag{20}$$

siendo A_{tot} la precisión y G_{tot} el *score* total para cada participación. A partir de los valores reflejados en la tabla 10, se pueden obtener ambos parámetros para los tres catálogos propuestos y compararlos con los resultados del resto de equipos participantes del *challenge*. Los resultados se recogen en la tabla 12. Comparando los resultados obtenidos para los tres catálogos con las puntuaciones de los equipos que tomaron parte en el *challenge*, se pueden situar en un lugar intermedio entre las puntuaciones de los equipos EngageSKA y ICRAR. De nuevo, en comparación con los resultados del IPM, a partir de cuyo planteamiento se ha desarrollado el presente trabajo, se han obtenido resultados mejores.

	G_{tot}	A_{tot}
Shanghai	19112.8	19419.7
ARCIt-CACAO	17361.3	24684.6
EngageSKA	16914.9	20551.5
Catálogo 3	11918.68	14013.68
Catálogo 2	8932.1	9999.62
Catálogo 1	6203.03	6796.47
ICRAR	5265.56	11691.1
RADGK	-4.58427	0.746315
hs	-9325.29	684.933
JLRAT	-10625.9	64752.6
IPM	-196237	4356.57
IPM2	-533625	28973.2

Tabla 12: *Score global y precisión de los catálogos con diferentes umbrales propuestos, integrados en los resultados de los equipos participantes en el SDC1 [9] y ordenados en orden decreciente de G_{tot}. Si el número de falsos positivos excede la cantidad de matches, la puntuación global puede ser negativa. Que para los tres catálogos propuesto A_{tot} y G_{tot} sean similares indica que los catálogos contienen un porcentaje bajo de falsos positivos, tal y como indican lo altos índices de reliability reflejados en la tabla 11.*

3.3. Resultados de la clasificación de fuentes

No todos los grupos participantes en el SDC1 completaron la clasificación de fuentes, por lo que el procedimiento para evaluar y puntuar esta categoría no está definido con tanta precisión, de forma que se va a hacer un análisis cualitativo de los resultados obtenidos de implementar el método previamente descrito.

Se han clasificado los objetos de los tres catálogos en grupos en función de su aparición en las diferentes bandas, para así asociarles una clase según su posición en los diagramas de flujo y color correspondientes. Como se ha establecido, la principal limitación del procedimiento reside en la diferencia de tamaño que hay entre las imágenes de las diferentes bandas. En la tabla 10 se puede ver que, para la banda B5, se tienen sólo 102, 164 y 291 detecciones para los catálogos 1, 2 y 3 respectivamente. Con este número limitado de fuentes, hay grupos de bandas para los que no se ha podido construir el diagrama de flujos, porque no hay objetos con presencia en todas las bandas necesarias a la vez.

El catálogo 1 tiene la pareja de umbrales más altos y, por tanto, el menor número de detecciones: no se han podido construir todos los diagramas de flujo. El grupo de objetos que aparecen en las tres bandas incluye una sola detección, por lo que se ha asignado a ese objeto la clase que le corresponde según la zona que ocupa dentro del diagrama, pero no se incluye dicho diagrama porque no presenta interés. Dentro de la categoría de objetos que aparecen en dos de las bandas para los objetos en B2 y B5 ocurre igual, y no existen objetos que ocupen a la vez las bandas B1 y B5, por lo que en la figura 13 se incluye simplemente el diagrama de flujo correspondiente a los objetos que están en B1 y B2.

Pese a que los catálogos 2 y 3 tienen una banda B5 con un mayor número de detecciones, tal y como corresponde a sus *thresholds* más bajos, presentan los mismos problemas que el catálogo 1, con la diferencia de que para el catálogo 3 tampoco existe el grupo de objetos en tres bandas. De esta forma, los diagramas de flujo para el grupo de objetos que están en B1 y B2 se incluye también en la figura 13.

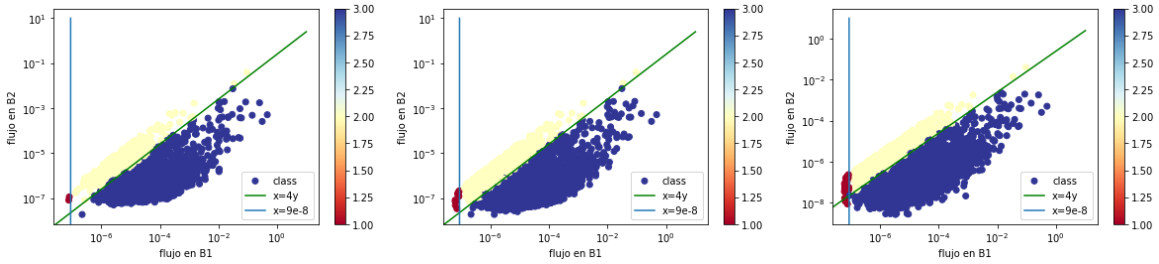


Figura 13: *Diagramas de flujo para los objetos que se encuentran en las bandas B1 y B2 de los catálogos 1, 2 y 3 respectivamente. Se han asociado las clases de los objetos en función de la región que ocupan sobre el diagrama, tal y como se ha establecido previamente.*

Para evaluar la calidad de la clasificación, se incluyen en la última imagen de la figura 11 la cantidad de objetos correcta e incorrectamente clasificados. Se ve que la clasificación ha tenido poco éxito. Pese a que la cantidad de objetos mal clasificados cambia mucho para los tres catálogos, la cantidad de objetos bien clasificados permanece constante para los tres catálogos. La diferencia principal entre los tres catálogos radica en que el número de objetos detectados y, por tanto, la cantidad de objetos en cada banda, aumenta considerablemente al bajar el *threshold*. Al aumentar el número de detecciones a lo largo de los catálogos el número de objetos que se puede clasificar empleando los diagramas de color y flujo no aumenta significativamente: lo que sí aumenta es el número de objetos que queda confinado a la banda B1. Como se ha establecido previamente, a esas fuentes se les ha asignado una clase basándose en las características del espectro de emisión de cada tipo de objeto. El aumento de clasificaciones incorrectas con el aumento de tamaño sugiere que esto es demasiado simple y, para una gran mayoría de casos, incorrecto. De hecho, de acuerdo con el *true catalogue*, el objeto más abundante en las imágenes del *challenge* son las SFGs, pero en esta clasificación son las SS-AGNs, porque son la clase que se ha asociado a los objetos que aparecen sólo en la banda B1.

Pese a que esto muestra que el método tiene fallos, y se beneficiaría de un análisis multiparámetro, la clasificación para el resto de participantes del SDC1 tampoco ha resultado satisfactoria. La simulación con un solo *pointing* y la falta de datos a lo largo de un rango más amplio de frecuencias, junto con el flujo invariante dentro de cada banda presentan una gran limitación a la hora de discriminar entre clases [9]. Sólo 5 de los equipos participantes han llevado a cabo la clasificación de fuentes, con diferentes grados de éxito.

4. Discusión y conclusiones

En esta sección se discuten cuestiones relacionadas con los resultados y problemas derivados de la implementación del método, además de diferentes posibilidades a la hora de afrontar algunos problemas y sugerencias en caso de querer perfeccionarse la técnica aplicada para resolver ciertas cuestiones.

Como se ha reiterado a lo largo del documento, el modelado y corrección del *Primary Beam* es un área de estudio muy amplia, y el proceso puede ser todo lo complejo que se desee. A la hora de modelarlo se barajaron varias opciones: una interpolación de diferentes grados de su forma a partir de la imagen, y un ajuste a una función de disco de Airy antes de decantarse por el ajuste a la función cosenoidal de la expresión 10. Un ajuste gaussiano se descartó rápidamente, porque pese a que describe bien el pico central no es útil a la hora de modelar los lóbulos laterales.

Tal y como se puede observar en la figura 7, al dividir las imágenes del cielo entre el *Primary Beam* modelado aparece en la imagen un patrón en forma de anillos. Esto se debe a que, incluso con el ajuste cosenoidal que describe los *sidelobes* con bastante precisión, su modelado es complejo y los pequeños desajustes se reflejan mucho al dividir la imagen final, dejando el patrón de la figura 7 como rastro. De hecho, el ajuste cosenoidal se ha escogido frente a los otros porque marca un patrón de anillos más débil, garantizando así una mejor obtención de los parámetros para el catálogo. Los *Primary Beams* interpolado y ajustado a una función de Airy dejan un rastro en la imagen mucho más marcado, que tiene como consecuencia una peor caracterización de las fuentes.

Debido a la naturaleza del *Primary Beam*, tal y como se ha discutido, la corrección ha de hacerse como paso posterior a la detección de fuentes. Esto implica que la detección se hace sobre la imagen sin corregir, y de esta forma se pierden fuentes cerca de los bordes de la imagen. No obstante, detectando sobre la imagen corregida también surgen problemas. Debido al patrón de anillos, hay zonas de la imagen en que no se detecta ninguna fuente, y las detecciones también se colocan sobre la imagen reflejando el mismo dibujo, perdiéndose así más información que de la otra manera.

Otra cuestión a discutir es el ajuste a un modelo de perfil de Sérsic para describir la forma y tamaño de los objetos del catálogo. Se han determinado los tamaños de los semiejes mayor y menor de la elipse que da forma a cada objeto a través de este ajuste porque es el que ha ofrecido mejores resultados y el que devuelve los tamaños en el orden de magnitud correcto. Además, el perfil luminoso de una SFG viene descrito por un perfil de Sérsic en el *true catalogue*, por lo que es la mejor aproximación para la mayoría de objetos del catálogo. No obstante, si se quisieran mejorar los resultados en esta categoría, una vez determinada la clase de cada objeto se podría ajustar su tamaño con una distribución personalizada (gaussiana, por ejemplo) en función de su clase. Esto mejoraría los errores de tamaño reflejados en la figura 11.

A lo largo del documento se han discutido las limitaciones del método de clasificación, y se vuelve a incidir sobre ellas. El diferente tamaño angular de las bandas es la principal limitación a la hora de agrupar los objetos por su presencia en los distintos canales de frecuencia. Además, que dentro de cada banda el ancho de frecuencia no se refleje en variaciones de flujo a lo largo de la misma también es una limitación. Para mejorar los resultados de la clasificación el siguiente paso podría ser evaluar otras categorías además del flujo pero, de todas formas, en una situación real los datos de SKAMid tendrán un flujo variable a lo largo de la banda que proporcionará más información que su equivalente

simulado.

Por último, se han obtenido errores significativos a la hora de estimar el ángulo de posición (PA) de los diferentes objetos para caracterizarlos. Comparando los errores de los diferentes parámetros obtenidos, tal y como se ve en la figura 11, el PA es el que tiene la distribución menos deseable: no hay un pico definido y el error se extiende de manera uniforme, lo que indica una fuerte contribución estadística. Los errores del resto de participantes del *challenge* presentan una distribución similar. Un estudio de los ángulos de posición en el *true catalogue* revela que se extienden en un rango de valores contenidos en el intervalo [-90, 360], un intervalo un poco singular. Sin embargo, en los catálogos obtenidos y propuestos la magnitud se extiende entre los valores [-90, 90]. Dado que el error se extiende de forma homogénea por todos los valores y es así para la mayoría de participantes del *challenge*, es posible que se trate de un problema de definición en los datos simulados, derivado de un criterio inconsistente a la hora de establecer el origen.

El análisis de los datos del SDC1 dentro del presente trabajo representa dos objetivos principales: poder recrear una participación en el *challenge* de la forma más completa posible, y el desarrollo de habilidades desde un punto de vista más didáctico. En cuanto a los objetivos científicos, se han emulado con éxito la detección, caracterización y clasificación de fuentes propuestos en el *challenge*, con resultados comparables a los del resto de equipos científicos participantes. Teniendo en cuenta los recursos limitados con los que se cuenta en comparación, los resultados obtenidos son más que satisfactorios. Se han obtenido mejores resultados que el IPM, el otro equipo participante que empleó `SExtractor` para la detección y caracterización de fuentes, y los errores en las magnitudes e inconsistencias en los resultados tienen una explicación razonable y se discuten propuestas de mejora en caso de querer recrearse o retomar esta línea de investigación.

Debido a la naturaleza dinámica del proceso de trabajo, los objetivos del proyecto han ido evolucionando, para finalmente conformar los resultados obtenidos. La elaboración de los distintos catálogos ha implicado el desarrollo de habilidades multidisciplinares. La correcta realización del trabajo pasa por la comprensión del papel y la relevancia de SKAO en el contexto de desarrollo científico actual, y en los campos de la astrofísica y la cosmología. También se ha aprendido a manejar herramientas para la detección de fuentes con un uso muy extendido y de gran utilidad en el campo: `SExtractor` y `PSFEx`, haciendo especial hincapié en la comprensión de su funcionamiento y proceso de configuración. El volumen de los datos y su procedimiento de caracterización han permitido también un desarrollo en habilidades complementarias, como la programación mediante el uso de `Python`, que queda integrada como una herramienta muy útil.

El trabajo también ha servido como un primer contacto con el mundo de la investigación y el método de trabajo en un contexto científico real: la resolución de problemas de forma creativa, la comprensión de cuestiones complejas y una muestra de la colaboración internacional en el mundo de la ciencia.

Referencias

[1]SKAO, *The SKA Telescopes*, https://www.skao.int/en/explore/telescopes, (accedido: 10.02.2025).

[2]SKA, *SKA Telescope Specifications*, https://www.skao.int/en/science-users/118/ska-telescope-specifications#__otpm0, (accedido: 04.02.2025).

[3]SKAO, *History of the SKA project*, https://www.skao.int/en/about-us/91/history-ska-project, (accedido: 10.02.2025).

[4]A. Liu y J. R. Shaw, "Data Analysis for Precision 21cm Cosmology", Publications of the Atronomical Society of the Pacific **132**, 062001 (2020).

[5]SARAO, *MeerKAT Radio Telescope*, https://www.sarao.ac.za/science/meerkat/about-meerkat/, (accedido: 10.02.2025).

[6]SKAO Science Team, *SKA Science Data Challenge 1: Data Description*, inf. téc. (SKAO, 2018).

[7]SKAO, *Data Challenges*, https://www.skao.int/en/science-users/160/data-challenges, (accedido: 04.02.2025).

[8]SKAO, *SKA Science Data Challenge 1*, https://www.skao.int/en/464/ska-science-data-challenge-1, (accedido: 10.02.2025).

[9]A. Bonaldi et al, "Square Kilometre Array Science Data Challenge 1: analysis and results", Monthly Notices of the Royal Astronomical Society **500**, 3821–3837 (2020).

[10]SKA, *SKA Science Data Challenge Scoring's documentation*, https://developer.skatelescope.org/projects/sdc1-scoring/en/latest/index.html, (accedido: 17.01.2025).

[11]Y. S. Dai et al., "Is there a relationship between AGN and star formation in IR-bright AGNs?", Monthly Notices of the Royal Astronomical Society **478**, 4238–4254 (2018).

[12]A. Banerjee, B. Pandey y A. Nandi, *Clustering and physical properties of the star-forming galaxies and AGN: does assembly bias have a role in AGN activity?*, (2023) https://arxiv.org/abs/2310.12943.

[13]F. An et al, "Radio spectral properties of star-forming galaxies between 150 and 5000 MHz in the ELAIS-N1 field", Monthly Notices of the Royal Astronomical Society **528**, 5346–5363 (2024).

[14]B. L. Fanaroff y J. M. Riley, "The Morphology of Extragalactic Radio Sources of High and Low Luminosity", Monthly Notices of the Royal Astronomical Society **167**, 31P-36P (1974).

[15]SKAO, *SKA-Mid*, https://www.skao.int/en/explore/telescopes/ska-mid, (accedido: 10.02.2025).

[16]E. Hecht, *Optics*, 5.ª ed. (Pearson, 2017).

[17]A. R. Thompson, J. M. Moran y G. W. Swenson Jr., *Interferometry and Sythesis in Radio Astronomy*, 3.ª ed. (Springer Open, 2017).

[18]D. E. Gary, *Fourier Synthesis Imaging*, https://web.njit.edu/~gary/728/Lecture6.html, (accedido: 10.02.2025).

[19]D. E. Gary, *Primary Antenna Elements*, https://web.njit.edu/~gary/728/Lecture4.html, (accedido: 10.02.2025).

[20]H. Karttunen et al., *Fundamental Astronomy*, 5.ª ed. (Springer, 2007).

[21]SKA Organisation, *SKA Phase 1 Construction Proposal*, inf. téc. (SKAO, 2020).

[22]J. J. Condon y S. M. Ransom, *Essential Radio Astronomy*, https://www.cv.nrao.edu/~sransom/web/Ch3.html#S2.SS3, (accedido: 10.02.2025).

[23]U. Rau, *Imaging Algorithms in CASA*, https://www.aoc.nrao.edu/~rurvashi/ImagingAlgorithmsInCasa/ImagingAlgorithmsInCasa.html, (accedido: 10.02.2025).

[24]T. J. Cornwell, *EVLA Memo 62 Full primary beam Stokes I , Q , U , V imaging*, (2003) https://api.semanticscholar.org/CorpusID:933855.

[25]W. D. Cotton y M. de Villiers, *Primary Beam Corrections of MeerKAT: Getting it Right*, (2024) https://www.cv.nrao.edu/~bcotton/ObitDoc/MK_BeamCor_2.pdf.

[26]L. Schwardt y M. de Villiers, *Primary beam model library for the MeerKAT project*, https://github.com/ska-sa/katbeam, (accedido: 11.01.2025).

[27]SKAO Science Team, *SKA Science Data Challenge 1: Analysis Description*, inf. téc. (SKAO, 2019).

[28]E. Bertin y A. Moneti, *PSFEx User Manual*, https://psfex.readthedocs.io/en/latest/index.html, (accedido: 15.01.2025).

[29]E. Bertin, *SExtractor User Manual*, https://sextractor.readthedocs.io/en/latest/index.html, (accedido: 15.01.2025).

[30]A. Muñoz, *TFG-SDC1*, https://github.com/aliciamsm/TFG-SDC1, (accedido: 11.02.2025).

[31]J. Collinson, *Science Data Challenge 1 Scoring API*, https://gitlab.com/ska-telescope/sdc/ska-sdc/-/tree/master/ska_sdc/sdc1, (accedido: 09.02.2025).

Apéndice

El código empleado para la elaboración de los catálogos y el modelado del *Primary Beam* se ha escrito mediante `Python`. Se ha creado un repositorio de *Github* [30], donde se pueden consultar los programas escritos. Incluye el código usado para el modelado del *Primary Beam* y para la caracterización y clasificación de fuentes de los catálogos, con comentarios y una breve descripción del material proporcionado. Se ha trabajado con las librerías `numpy`, `matplotlib` y `astropy`, además de la librería desarrollada por SKAO para la evaluación del SDC1 [31].

`Numpy` es una librería con usos dentro del cálculo numérico, y `matplotlib` se ha empleado para generar gráficos y figuras como los que aparecen en este documento. `Astropy` es una librería de uso muy extendido dentro del ámbito de la astronomía: entre otras cosas, permite un cómodo manejo de archivos en formato *.fits*. La librería del *scorer* del SDC1 tiene una utilidad doble: puntuar la calidad de los catálogos y obtener datos como los que aparecen en la tabla 10, y comparar dos catálogos y determinar qué objetos aparecen en ambos. Comparando un catálogo con su *true catalogue* asociado se puede estudiar su calidad, pero esta función se ha empleado también para comparar catálogos de diferente bandas entre sí y agrupar los objetos según su aparición a diferentes frecuencias para desarrollar el método de clasificación de fuentes.