Facultad
de
Ciencias

ESTUDIO E IMPLEMENTACIÓN DE UNA HERRAMIENTA PROBABILÍSTICA PARA LA CROSS-IDENTIFICACIÓN DE OBJETOS EXTRAGALÁCTICOS

(Study and Implementation of a Probabilistic Tool for the Cross-Identification of Extragalactic Objects)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Alberto Manjón García

Director: Diego Herranz Muñoz

Septiembre - 2015

Índice de Contenidos

Resumen

En este trabajo se ha implementado y estudiado la eficacia de un método probabilístico bayesiano de cross-identificación de objetos extragalácticos, basado en el formalismo teórico propuesto por Tamás Budavári y Alexander S. Szalay [1]. Esta herramienta es capaz de cotejar dos catálogos astronómicos obtenidos con instrumentos de distintas resoluciones angulares, como son el Observatorio Espacial Herschel y el proyecto GAMA, e incorpora no sólo información tanto de las distintas resoluciones angulares y las distancias relativas entre los candidatos, sino también otra información adicional de carácter astrofísico conocida a priori como es el corrimiento al rojo. Este enfoque bayesiano utiliza el factor de Bayes como medida fiable de la calidad de las asociaciones entre fuentes astronómicas de los catálogos estudiados, permitiendo excluir contrapartidas inverosímiles o poco probables.

La presente memoria se divide en cinco partes. En la primera se introducen extensamente los conceptos de inferencia estadística, inferencia bayesiana, teorema de Bayes y factor de Bayes. En la segunda parte se aplican estos conceptos para el caso concreto de disponer de medidas de las posiciones y corrimientos al rojo de diversos objetos extragalácticos, mostrándose las consideraciones y aproximaciones tenidas en cuenta para deducir las ecuaciones utilizadas por nuestro método probabilístico de cross-identificación. En la tercera parte se presentan los proyectos de investigación con los que se han obtenido los catálogos astronómicos cotejados por nuestra herramienta, así como una descripción de los mismos. En la cuarta parte se explican los programas y el script utilizados para implementar la herramienta y analizar los resultados. Finalmente, se muestra el análisis de los resultados y las conclusiones obtenidas.

Palabras clave: objetos extragalácticos, inferencia bayesiana, cross-identificación, Herschel, GAMA.

Abstract

In this work we have implemented and studied the efficiency of a Bayesian probabilistic method of cross-identification of extragalactic objects, based on the theoretical formalism proposed by Tamás Budavári and Alexander S. Szalay [1]. This tool is able to collate two astronomical catalogs obtained with instruments of different angular resolutions, such as Herschel Space Observatory and GAMA project, and incorporates not only information of both the different angular resolutions and the relative distances between the candidates, but also other additional astrophysical information known a priori as is the redshift. This Bayesian approach uses Bayes factor as a reliable measure of the quality of the partnerships between astronomical sources of the catalogs studied, allowing excluding implausible or unlikely counterparts.

This memory is divided into five parts. In the first one the concepts of statistical inference, Bayesian inference, Bayes theorem and Bayes factor are widely introduced. In the second part these concepts are applied to the specific case of having measures of positions and redshifts of many extragalactic objects, showing the considerations and approaches taken into account to derive the equations used by our probabilistic cross-identification method. In the third part the research projects used to obtain the astronomical catalogs checked by our tool and a brief description of them are presented. In the fourth part programs and the script used to implement the tool and analyze the results are explained. Finally, the analysis of the results and conclusions obtained are shown.

Key words: extragalactic objects, Bayesian inference, cross-identification, Herschel, GAMA.

1. Introducción

El progreso tecnológico que se lleva experimentando en los últimos años propicia la necesidad de procesar grandes cantidades de datos de forma rápida. Dentro del campo de la astrofísica y la cosmología esta labor se vuelve más ardua aún si cabe por la complejidad y numerosas variables de los datos obtenidos. Con la introducción de detectores de alta resolución y gran formato en todas las longitudes de onda del espectro electromagnético, los astrofísicos afrontan hoy una avalancha de datos procedente de estos nuevos instrumentos. Por este motivo, es importante desarrollar técnicas apropiadas de procesado de datos e implementarlas en entornos de supercomputación para facilitar este trabajo.

Una de las herramientas más poderosas de que disponemos hoy en día para comprender la formación y evolución de galaxias a lo largo de la historia del Universo es el estudio de su distribución espectral de energía (SED), es decir, de las propiedades de su espectro electromagnético en todo el rango de longitudes de onda observables. Sin embargo, uno de los grandes problemas a la hora de realizar el estudio de la distribución espectral de energía de estos objetos extragalácticos es que en la práctica no es posible observar simultáneamente y con un único instrumento todo el espectro electromagnético de un cuerpo celeste. Esto es debido a que la física de los detectores empleados es muy distinta a diferentes frecuencias. En lugar de ello, es preciso combinar conjuntos de datos o catálogos obtenidos en distintos observatorios, con características instrumentales muy diferentes, para así poder acceder a los distintos rangos espectrales de interés: rayos X, ultravioleta, óptico, infrarrojo, microondas, radio, etc.

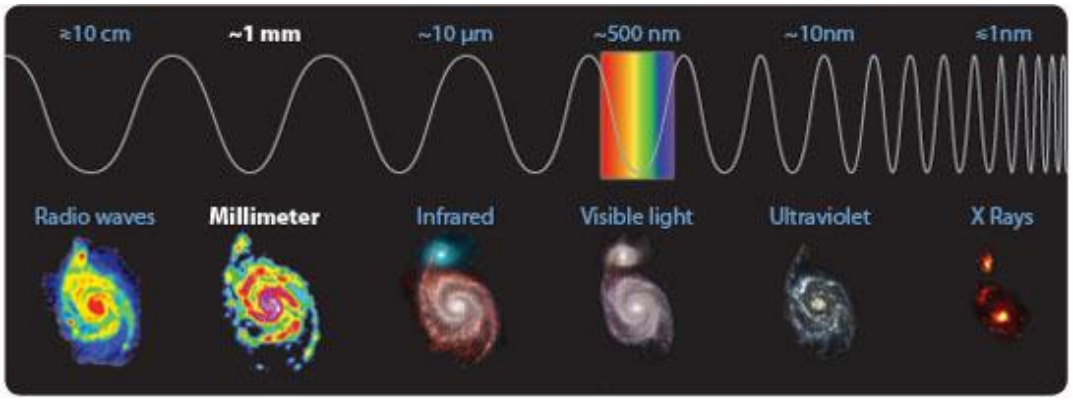

Figura 1. Ilustración del rango de longitudes de onda observables correspondientes al espectro electromagnético de una cierta galaxia.
(http://www.iram-institute.org/medias/uploads/image/ScienceAndTechnology/Wavelengths.jpg)

Los diversos instrumentos de observación utilizados normalmente tienen una resolución angular y sensibilidad diferentes por lo que resulta necesario realizar una cross-identificación, es decir, asociar a cada uno de los objetos en uno de los catálogos parciales su contrapartida más fiable en el resto de catálogos. Esta tarea puede volverse muy difícil cuando la discrepancia en resolución angular y/o sensibilidad de los distintos experimentos involucrados es muy grande. Frecuentemente esto conduce a situaciones en las que para un objeto observado con un instrumento de baja resolución angular o sensibilidad pueden presentarse múltiples posibles contrapartidas observadas con otro instrumento de mayor precisión, entre las que se hace necesario elegir cuál es la más probable.

De este modo, el objetivo de este trabajo ha sido implementar y estudiar la eficacia de un método probabilístico bayesiano de cross-identificación de objetos extragalácticos, basado en el formalismo teórico propuesto por Tamás Budavári y Alexander S. Szalay [1], capaz de cotejar dos catálogos astronómicos obtenidos con instrumentos de distintas resoluciones angulares, como son el Observatorio Espacial Herschel y el proyecto GAMA, y que incorpora no sólo información tanto de las distintas resoluciones angulares y las distancias relativas entre los candidatos, sino también otra información adicional de carácter astrofísico conocida a priori como es el corrimiento al rojo. En este método probabilístico, la cantidad o magnitud utilizada para determinar cuál es la contrapartida más probable a un objeto extragaláctico se denomina factor de Bayes.

En este trabajo de fin de grado se ha familiarizado con un problema astrofísico que, con la llegada de los nuevos grandes cartografiados automáticos de galaxias, cada vez está adquiriendo mayor importancia: la necesidad de procesar mediante cross-matching catálogos de millones de objetos extragalácticos de la forma más automática posible; y se ha implementado una forma sencilla de afrontar este problema para dos catálogos astronómicos. Así mismo, en el transcurso del mismo se han introducido y tratado conceptos estadísticos novedosos, como son la inferencia bayesiana y el factor de Bayes, se ha familiarizado con el manejo e interpretación de catálogos astronómicos, se ha profundizado en la naturaleza de los instrumentos y observatorios que han hecho posibles estos catálogos y se han manejado programas de tratamiento de gran cantidad de datos que serán útiles en el futuro.

Figura 2. Fotografía de la Vía Láctea sobre el Gran Telescopio de Canarias (GTC) (http://www.iac.es/blog/vialactea/wp-content/uploads/2014/06/GTC-Pano-Lacteapeq.jpg)

2. La Inferencia Bayesiana

2.1. Inferencia Estadística

La inferencia estadística es el proceso de deducción de propiedades sobre una cierta población estadística, el conjunto de elementos sobre el que se realizan las observaciones y sobre el que queremos sacar conclusiones, mediante el análisis de datos. Al ser normalmente demasiado grande para poder abarcarse, los elementos o datos observados son sólo una muestra de la población estudiada. De este modo, un problema de inferencia estadística es aquel en el que se han de analizar datos que han sido generados de acuerdo con alguna distribución de probabilidad desconocida, planteando determinadas cuestiones sobre tal distribución que permitan, tras analizar los datos, conocer la distribución que los ha generado. En muchas ocasiones esta distribución de probabilidad es conocida salvo para los valores de uno o más parámetros.

Cualquier inferencia estadística requiere algunas suposiciones previas. Así, un análisis estadístico inferencial hace proposiciones sobre la población, a partir de datos extraídos de la misma con algún tipo de muestreo. Dada una hipótesis sobre la población, la inferencia estadística primero selecciona un modelo estadístico, que no es más que un conjunto de supuestos acerca de la generación de los datos observados, y después deduce proposiciones estadísticas de ese modelo, en forma de estimaciones puntuales, estimaciones de intervalos de confianza, intervalos de credibilidad, clasificación de los datos en grupos o el rechazo de la hipótesis de partida.

En contraposición con la inferencia estadística se encuentra la estadística descriptiva, la cual se enfoca exclusivamente en las propiedades de los datos observados y no asume que estos puedan provenir de una población estadística mayor. La estadística descriptiva suele utilizarse como paso preliminar a la obtención de conclusiones por la inferencia estadística.

Existen principalmente dos modelos de inferencia estadística: la estadística clásica de errores o inferencia frecuentista y la inferencia bayesiana. La inferencia frecuentista es un tipo de inferencia estadística en la que las conclusiones se extraen del estudio de la frecuencia o propensión de algún fenómeno. Por su lado, la inferencia bayesiana es un tipo de inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta.

2.1.1. Inferencia Frecuentista

La inferencia frecuentista consiste en extraer conclusiones a partir de una muestra de datos atendiendo a la frecuencia o proporción de los mismos. Este modelo de inferencia se ha asociado con la interpretación frecuentista de la probabilidad alegando que sólo es aplicable en términos de probabilidad de frecuencia; es decir, en términos del muestreo repetido de una población. Específicamente en cuanto a que cualquier experimento dado puede ser considerado como uno de una secuencia infinita de posibles repeticiones del mismo, cada uno capaz de producir resultados estadísticamente independientes. Desde este punto de vista, el enfoque de la inferencia frecuentista de sacar conclusiones a partir de datos es efectivo en cuanto a que la conclusión correcta debería obtenerse con una probabilidad dada alta, entre este conjunto teórico de repeticiones.

No obstante, estos mismos procedimientos se pueden desarrollar bajo una formulación sutilmente diferente. Desde otro punto de vista, se puede argumentar que el diseño de un experimento debe incluir, antes de su realización, decisiones sobre las medidas a tomar para poder alcanzar conclusiones a partir de los datos obtenidos. De este modo, siguiendo estas reglas hay una alta probabilidad de llegar a una conclusión correcta, donde la probabilidad se refiere a un conjunto de eventos al azar aún por ocurrir y, por lo tanto, no se basa en la interpretación de frecuencia de la probabilidad.

La diferencia fundamental entre la inferencia frecuentista y la bayesiana es el concepto de probabilidad. Aunque las probabilidades están involucradas en ambas aproximaciones a la inferencia, éstas están asociadas con tipos de cosas diferentes. Para la estadística clásica es un concepto objetivo que se encuentra en la naturaleza, mientras que para la estadística bayesiana se encuentra en el observador, siendo así un concepto subjetivo. La inferencia frecuentista sólo toma como fuente de información las muestras obtenidas. En el caso bayesiano, sin embargo, a parte de la muestra de datos obtenida, también juega un papel fundamental la información previa o externa que se posee en relación a los fenómenos que se tratan de modelizar. El resultado de la inferencia bayesiana suele ser una distribución de probabilidad sobre lo que se conoce de los parámetros dados los resultados del experimento o estudio. El resultado de la inferencia frecuentista es o bien una conclusión de veracidad o falsedad o una conclusión en forma de intervalo de confianza que incluye el valor verdadero para una muestra dada.

Otra diferencia importante entre los enfoques frecuentista y bayesiano de la inferencia es que para la perspectiva clásica los parámetros de las distribuciones de probabilidad son desconocidos pero cantidades fijas pertenecientes a un determinado espacio, denominado espacio paramétrico. Por su lado, la perspectiva bayesiana parte de la premisa de que los parámetros de interés de las distribuciones de probabilidad, aunque desconocidos, son variables aleatorias, lo que permite modelar todas las fuentes de incertidumbre en los modelos estadísticos. En este sentido, en la inferencia frecuentista no es posible asociar probabilidades a estos parámetros desconocidos puesto que no son susceptibles de ser tratados como variables aleatorias. En contraste, la inferencia bayesiana permite asociar probabilidades a sus parámetros al sí tratarse de variables aleatorias, donde estas probabilidades pueden tener a veces tanto una interpretación frecuentista de la probabilidad como una bayesiana.

2.1.2. Inferencia Bayesiana

El principal enfoque alternativo a la inferencia frecuentista es la inferencia bayesiana. Este enfoque alternativo se fundamenta en que en ocasiones, antes de disponer de las observaciones de la variable objeto de estudio, el experimentador dispone de información adicional sobre donde es probable que se encuentre el valor del parámetro y esa información se puede expresar en términos de una distribución de probabilidad en el espacio paramétrico. Por lo tanto, esta información previa permite al investigador tener la creencia de que es más probable que el parámetro se encuentre en una determinada región del espacio paramétrico que en otra.

La lógica establece unas reglas de inferencia a partir de las cuales se construye el sistema de razonamiento deductivo, en el que la conclusión cierta se infiere necesariamente de las premisas. Una determinada proposición es considerada cierta o falsa sin que se admitan grados entre estos dos extremos. La lógica tradicional no contempla que tanto la información de entrada como las propias reglas puedan no ser ciertas con carácter absoluto. No obstante, la incertidumbre y la imprecisión son inherentes al proceso de razonamiento humano. En este

sentido, existen métodos de razonamiento aproximado que aportan modelos teóricos que simulan la capacidad de razonamiento en condiciones de inexactitud o incertidumbre, siendo capaces de obtener conclusiones útiles a partir de información incompleta o que admite un rango de variación.

Entre estos métodos de razonamiento aproximado se encuentran los métodos bayesianos, basados en el conocido teorema de Bayes. Todos ellos tienen en común la asignación de una probabilidad como medida de credibilidad de las hipótesis. En este contexto, la inferencia se entiende como un proceso de actualización de las medidas de credibilidad al conocerse nuevas evidencias. Mediante la aplicación del teorema de Bayes se busca obtener las probabilidades de las hipótesis condicionadas a las evidencias que se conocen. Las probabilidades previas a la inferencia se conocen como probabilidades a priori, y las probabilidades posteriores a la inferencia se conocen como probabilidades a posteriori.

La inferencia bayesiana utiliza aspectos del método científico, lo que implica recolectar la evidencia que se considera consistente o inconsistente con una hipótesis dada. Incluso antes de haberla observado ya calcula un estimador numérico del grado de creencia en la hipótesis y tras disponer de la nueva evidencia vuelve a calcular este estimador numérico. A medida que la evidencia se acumula, el grado de creencia en una hipótesis se va modificando. Con evidencia suficiente, el grado de creencia en la hipótesis se hará muy grande o muy pequeño. En ese sentido se dice que la probabilidad bayesiana proporciona un método racional para la actualización de creencias. Así, los que sostienen la inferencia bayesiana dicen que puede ser utilizada para discriminar entre hipótesis en conflicto: las hipótesis con un grado de creencia muy alto deben ser aceptadas como verdaderas y las que tienen un grado de creencia muy bajo deber ser rechazadas como falsas. Sin embargo, los detractores dicen que este método de inferencia puede ser prejuicioso debido a las creencias iniciales que se deben sostener antes de comenzar a recolectar cualquier evidencia.

La denominación de bayesiana se debe al matemático y estadístico británico Thomas Bayes (1702-1761) quién proporcionó un primer tratamiento matemático para un problema no trivial de inferencia bayesiana, demostrando un caso especial de lo que hoy se conoce como Teorema de Bayes, en un artículo titulado *"An Essay towards solving a Problem in the Doctrice of Chances"*. No obstante, no fue él sino Pierre-Simon Laplace (1749-1827) quien popularizó la probabilidad bayesiana, introduciendo una versión general de dicho teorema y utilizándolo para abordar problemas de la mecánica celeste, estadística médica, fiabilidad, y jurisprudencia.

Más tarde, en el siglo XX, las ideas de Laplace fueron desarrolladas aún más divergiendo en dos direcciones distintas y dando origen a las corrientes objetivista y subjetivista de la práctica bayesiana. En la corriente objetivista, la probabilidad mide objetivamente la plausibilidad de las proposiciones, es decir la probabilidad de una proposición corresponde a una razonable creencia de que todos los que comparten los mismos conocimientos deben compartir esa creencia. El análisis estadístico depende sólo del modelo asumido y los datos analizados, no creyendo necesario involucrar decisiones subjetivas, y las reglas de la estadística bayesiana están justificadas por razones de racionalidad y coherencia interpretadas como una extensión de la lógica. Para los subjetivistas, en cambio, la probabilidad cuantifica una opinión personal y niegan la posibilidad de que se pueda realizar un análisis estadístico totalmente objetivo. En consecuencia, las corrientes objetiva y subjetiva de la probabilidad bayesiana difieren principalmente en su interpretación y construcción de la distribución de probabilidad a priori.

Uno de los principales impulsores del renacimiento de la visión bayesiana de la probabilidad fue Harold Jeffreys con la publicación de su libro *"Theory of Probability"* en 1939. Posteriormente, ya en la década de 1980, se produjo un espectacular crecimiento en la investigación y aplicación de los métodos bayesianos, en su mayoría atribuidos al descubrimiento de la Cadena de Márkov y los métodos de Monte Carlo, que eliminaron muchos de los problemas de cálculo.

Análogamente a lo que sucedía con la inferencia frecuentista, la inferencia bayesiana se ha asociado a menudo con la interpretación bayesiana de la probabilidad y, en consecuencia, se ha generalizado que la diferencia esencial entre la inferencia frecuentista y la inferencia bayesiana es la misma que la diferencia entre las dos interpretaciones, frecuentista y bayesiana, del significado de la probabilidad. Sin embargo, donde es apropiado, la inferencia bayesiana, entendida en este caso como una aplicación del teorema de Bayes, también es utilizada en el marco de la interpretación frecuentista de la probabilidad.

La interpretación bayesiana del concepto de probabilidad o probabilidad bayesiana es, en contraste con la interpretación de la probabilidad como la frecuencia o propensión de algún fenómeno, una cantidad que asignamos con el propósito de representar un estado de conocimiento, o estado de creencia. Desde el punto de vista bayesiano, una probabilidad es asignada a una hipótesis, mientras que bajo el punto de vista frecuentista, una hipótesis es habitualmente probada sin tener asignada una probabilidad. Puede verse como una extensión de la lógica proposicional que permite razonar con hipótesis, es decir, las proposiciones cuya verdad o falsedad es incierta. Desde este punto de vista, la probabilidad bayesiana pertenece a la categoría de probabilidades probatorias proporcionando un conjunto estándar de procedimientos y fórmulas para realizar este cálculo. Para evaluar la probabilidad de una hipótesis, especifica alguna probabilidad a priori, que se actualiza a continuación, a la luz de nuevos datos relevantes, que conforman la evidencia.

2.2. Descripción Discreta de la Inferencia Bayesiana

En general, se usan probabilidades de modo informal para expresar la información o la incertidumbre que se tiene sobre las observaciones de cantidades desconocidas. Sin embargo, el uso de probabilidades para expresar la información se puede hacer de modo formal. Desde el punto de vista matemático se puede demostrar que con el cálculo de probabilidades el conjunto racional de creencias puede representarse de modo numérico, de modo que existe una relación entre probabilidad e información, y la regla de Bayes proporciona un modo natural de actualización de las creencias cuando aparece nueva información. Este proceso de aprendizaje inductivo por medio de la regla de Bayes es la base de la inferencia bayesiana.

El concepto básico en estadística bayesiana es el de probabilidad condicional. Séanse dos sucesos, A y B, tal que $P\ (B) > 0$, la probabilidad condicional se enuncia como la probabilidad de que ocurra el suceso A dado que también tenga lugar el suceso B y se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

donde $P(A \cap B) = P(A, B)$ es la probabilidad conjunta de los sucesos A y B, la intersección de ambos. Como puede verse, en este caso la probabilidad condicional se denota $P(A|B)$ y se lee "la probabilidad de A dado B". Una interpretación de este concepto es que, tomando los mundos en los que se cumple el suceso B, $P(A|B)$ es la fracción de los mismos en los que también se

cumple el suceso A. No tiene por qué haber una relación causal o temporal entre A y B. A puede preceder en el tiempo a B, ser posterior o pueden ocurrir ambos simultáneamente. A puede causar B, viceversa o pueden no tener relación causal. Las relaciones causales o temporales son nociones que no pertenecen al ámbito de la probabilidad, pero pueden desempeñar un papel o no dependiendo de la interpretación que se le dé a los sucesos.

A partir de la ecuación anterior y de la expresión recíproca para la probabilidad condicional $P(B|A)$ se deduce la relación:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{2}$$

para el caso en que los sucesos A y B son estadísticamente dependientes.

En teoría de probabilidades, se dice que dos variables estadísticas son estadísticamente independientes cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra. Aplicado a dos sucesos aleatorios, se dice que son independientes entre sí cuando la probabilidad de cada uno de ellos no está influida porque el otro suceso ocurra o no, es decir, cuando ambos sucesos no están relacionados. Intuitivamente, el suceso A será independiente del suceso B si $P(A|B) = P(A)$. Si el suceso A es independiente del suceso B, automáticamente, el suceso B será independiente del suceso A, cumpliéndose que $P(B|A) = P(B)$. En otras palabras, si A y B son independientes, la probabilidad condicional de A dado B es simplemente la probabilidad de A, y viceversa. De este modo, la ecuación anterior se convierte en:

$$P(A \cap B) = P(A) \cdot P(B) \tag{3}$$

cumpliéndose así la regla de la multiplicación según la cual la probabilidad de ocurrencia simultánea de dos o más sucesos estadísticamente independientes es igual al producto de sus probabilidades individuales.

Desde el punto de vista bayesiano, todas las probabilidades son condicionales porque casi siempre existe algún conocimiento o experiencia previa. Así, a partir de la igualdad mostrada en la expresión (2) se puede derivar el teorema de Bayes para los sucesos A y B:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4}$$

Considerando una partición del suceso A, tal que $A = \{A_1, \dots, A_n\}$, y el mismo suceso B anterior del que se conocen las probabilidades condicionales $P(B|A_i)$, el teorema de la probabilidad total define la probabilidad del suceso B mediante la siguiente expresión:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i) \tag{5}$$

Finalmente, aplicando la expresión anterior a la ecuación del teorema de Bayes (4) se tiene:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i)\, P(A_i)}{\sum_{i=1}^{n} P(B|A_i) P(A_i)} \propto P(B|A_i)\, P(A_i) \tag{6}$$

2.3. Descripción Continua de la Inferencia Bayesiana

La anterior descripción de la nomenclatura y conceptos básicos de la estadística bayesiana puede aplicarse también a variables aleatorias continuas. En probabilidad y estadística, una variable aleatoria es una variable estadística cuyo valor está sujeto a variaciones debidas al azar. Puede tomar diferentes posibles valores, de manera similar a una variable matemática, pero se diferencia de ella en que cada uno de estos valores lleva una probabilidad asociada. Estos posibles valores pueden ser, por ejemplo, resultados de un experimento realizado.

Dada una variable aleatoria no es posible conocer con certeza el valor que tomará esta al ser medida o determinada, aunque sí se sabe que existe una distribución de probabilidad asociada al conjunto de posibles valores que puede tomar. Esta distribución de probabilidad es una función que asigna a cada suceso definido sobre la variable aleatoria, la probabilidad de que dicho suceso ocurra. Se usa para describir la probabilidad de que se den los diferentes valores.

Por su lado, la función de densidad de probabilidad (PDF) de una variable aleatoria continua, es una función que describe la probabilidad relativa de que esta variable aleatoria tome un valor dado. La probabilidad de que la variable aleatoria caiga dentro de un intervalo particular de valores está dada por la integral de esta función a lo largo de ese intervalo. Esta función de densidad de probabilidad es no negativa en todo su dominio, y su integral sobre todo el espacio es igual a uno, está normalizada. En teoría de la probabilidad, la función de densidad de probabilidad de una variable aleatoria y cualquiera aparece denotada como $p(y)$. La propiedad de normalización de esta función viene definida como:

$$\int_{-\infty}^{\infty} p(y)\, dy = 1 \tag{7}$$

Si en lugar de una única variable, se tienen dos variables aleatorias y e z, al igual que sucedía en el caso de los sucesos A y B, la función densidad de probabilidad conjunta de ambas se denota como $p(y, z)$. Por su lado, la función densidad de probabilidad condicional de que y ocurra si se da z se escribe $p(y|z)$. Igualmente, la función densidad de probabilidad condicional de que ocurra z si se da y es $p(z|y)$.

Si las variables aleatorias y e z son además estadísticamente independientes, entonces la función densidad de probabilidad conjunta de ambas es simplemente el producto de sus respectivas funciones de densidad de probabilidad:

$$p(y, z) = p(y) \cdot p(z) \tag{8}$$

Sin embargo, si no son estadísticamente independientes, la función densidad de probabilidad conjunta resulta:

$$p(y, z) = p(y|z) \cdot p(z) = \ p(z|y) \cdot p(y) \tag{9}$$

Nuevamente, a partir de esta última igualdad se deriva el teorema de Bayes:

$$p(y|z) = \frac{p(z|y) \cdot p(y)}{p(z)} \tag{10}$$

De acuerdo con la definición del teorema de la probabilidad total, la función densidad de probabilidad de la variable continua z viene dada por la expresión:

$$p(z) = \int p(z|y)\, p(y)\, dy \tag{11}$$

Aplicando la expresión anterior a la ecuación del teorema de Bayes (10) se tiene:

$$p(y|z) = \frac{p(z|y)\, f(y)}{p(z)} = \frac{p(z|y)\, p(y)}{\int p(z|y)\, p(y)\, dy} \propto p(z|y)\, p(y) \tag{12}$$

2.4. El Teorema de Bayes

La herramienta utilizada por la inferencia bayesiana para actualizar la probabilidad de que una hipótesis sea cierta a medida que se adquiere evidencia sobre la misma es el teorema de Bayes. Es una importante técnica no sólo en estadística, sino también especialmente en estadística matemática y en el análisis dinámico de una secuencia de datos. La inferencia bayesiana deriva la probabilidad a posteriori como consecuencia de dos antecedentes, una probabilidad a priori y una función de verosimilitud procedente de un modelo estadístico para los datos observados, de acuerdo con:

$$P(H_0|E) = \frac{P(E|H_0)\, P(H_0)}{P(E)} \tag{13}$$

donde:

- E es la nueva evidencia o datos observados y no usados al calcular la probabilidad a priori.

- H_0 es la hipótesis a estudiar, llamada a veces hipótesis nula en alusión a que postula que no existe una relación entre dos fenómenos medidos, inferida antes de considerar la nueva evidencia y cuya probabilidad es afectada por la misma.

- $P(H_0)$ es la probabilidad a priori de que la hipótesis H_0 sea cierta antes de observar la nueva evidencia E.

- $P(E|H_0)$ es la verosimilitud, la probabilidad condicional de haber observado la evidencia E si la hipótesis H_0 es cierta.

- $P(E)$ es verosimilitud marginal, la probabilidad de haber observado la evidencia E independientemente de que la hipótesis H_0 sea cierta o no. Este término es igual sea cual sea la hipótesis H_0 que se esté considerando, sólo depende de los datos observados.

- $P(H_0|E)$ es la probabilidad a posteriori de que la hipótesis H_0 sea cierta una vez que se ha tenido en cuenta la evidencia E observada.

Cabe destacar que, en el caso de considerar diferentes hipótesis H_0, sólo los factores $P(H_0)$ y $P(E|H_0)$ afectarían a la distribución final $P(H_0|E)$. En este sentido, la verosimilitud marginal $P(E)$ se interpreta como una constante de integración que asegura que la distribución de probabilidad final $P(H_0|E)$ esté normalizada. $P(E)$ no depende de la hipótesis H_0 considerada, no proporcionando ninguna información adicional sobre la probabilidad a posteriori. Así, como ambos factores, $P(H_0)$ y $P(E|H_0)$, aparecen en el numerador, la probabilidad a posteriori $P(H_0|E)$ puede expresarse, al igual que se ha mostrado en las ecuaciones (6) y (12), únicamente como:

$$P(H_0|E) \propto P(E|H_0)\, P(H_0) \tag{14}$$

Esta expresión es la probabilidad a posteriori sin normalizar considerando únicamente que la probabilidad a posteriori de una hipótesis está determinada por una combinación de la probabilidad inherente a la hipótesis (probabilidad a priori) y la compatibilidad de la evidencia observada con la hipótesis (verosimilitud).

En el teorema de Bayes de la ecuación (13), el factor $P(E|H_0)/P(E)$ representa el impacto que la evidencia E tiene sobre la creencia en la hipótesis H_0. De este modo, si es posible observar la evidencia E cuando la hipótesis H_0 considerada es verdadera, entonces este factor será muy grande:

$$\frac{P(E|H_0)}{P(E)} > 1 \implies P(E|H_0) > P(E) \tag{15}$$

Esto quiere decir que, si la hipótesis H_0 fuese cierta, la evidencia E sería más probable de observar que lo predicho inicialmente sin tener en cuenta la hipótesis. Al multiplicar la probabilidad a priori de la hipótesis $P(H_0)$ por este factor se va a tener una probabilidad a posteriori muy grande dada la evidencia. Lo opuesto se aplica para una disminución en la creencia. Por su lado, si la creencia no cambia:

$$\frac{P(E|H_0)}{P(E)} = 1 \implies P(E|H_0) = P(E) \tag{16}$$

Esto quiere decir que la evidencia E es independiente de la hipótesis H_0. Si la hipótesis fuese cierta, la observación de la evidencia sería exactamente igual de probable que lo predicho inicialmente sin tener en cuenta la hipótesis.

En la inferencia bayesiana, por lo tanto, el teorema de Bayes mide cuánto la nueva evidencia es capaz de alterar la creencia en la hipótesis. Si la evidencia no se corresponde con una hipótesis, se debe rechazar la hipótesis. Pero si una hipótesis es muy poco probable a priori, también hay que rechazarla, incluso si la evidencia parece coincidir.

El punto crítico de la inferencia bayesiana es, entonces, que proporciona una forma de combinar nuevas evidencias con las creencias a priori, a través de la aplicación de la regla de Bayes. Esto contrasta con la inferencia frecuentista, que confía únicamente en la evidencia como un todo, sin ninguna referencia a las creencias anteriores. No obstante, a pesar de sus diferencias, los resultados obtenidos bajo las inferencias bayesiana y frecuentista son similares bajo ciertas circunstancias, concretamente cuando:

1. Se usan distribuciones de probabilidad a priori objetivas.
2. El tamaño muestral es muy grande y la influencia de la distribución de probabilidad a priori es muy pequeña en comparación con la influencia de la función de verosimilitud.

Si la evidencia está formada por un conjunto de observaciones independientes e igualmente distribuidas, tal que $\mathrm{E} = \{e_1, \dots, e_n\}$, y se tiene un conjunto de hipótesis $\mathrm{H} = \{H_1, \dots, H_m\}$ mutuamente excluyentes y exhaustivas, el teorema de la probabilidad total afirma que, conocidas las probabilidades condicionales $P(\mathrm{E}|H_i)$, entonces la probabilidad $P(E)$ de observar la nueva evidencia bajo todas las hipótesis mutuamente excluyentes, también llamada probabilidad marginal de E, viene dada por la expresión:

$$P(\mathrm{E}) = \sum_{i=1}^{m} P(\mathrm{E}|H_i)P(H_i) \tag{17}$$

Aplicando la expresión anterior al teorema de Bayes mostrado en la ecuación (13) se tiene que:

$$P(H_i|\mathrm{E}) = \frac{P(\mathrm{E}|H_i)}{P(\mathrm{E})}P(H_i) = \frac{\prod_{k=1}^{n} P(e_k|H_i)}{\sum_{i=1}^{m} P(\mathrm{E}|H_i)P(H_i)}P(H_i) \tag{18}$$

donde

$$P(\mathrm{E}|H_i) = \prod_{k=1}^{n} P(e_k|H_i) \tag{19}$$

2.5. Pasos de la Inferencia Bayesiana

Considerando el problema general de inferir una distribución para un determinado parámetro θ, que en la estadística bayesiana no es un valor fijo sino una variable aleatoria, a partir de una cierta distribución de datos $\mathrm{E} = \{e_1, \dots, e_n\}$, la metodología a seguir por la inferencia bayesiana es la siguiente:

2.5.1. Distribución de Probabilidad a Priori

1. Especificar una distribución de probabilidad a priori para el parámetro θ, denotada $p(\theta|\alpha)$, que exprese el grado de creencia de uno o conocimiento previo sobre θ antes de observar la distribución de datos E.

Esta distribución refleja la incertidumbre sobre θ antes de tener en cuenta evidencia alguna. A menudo, una distribución de probabilidad a priori es la evaluación puramente objetiva de un experto con experiencia. Los parámetros α de una distribución de probabilidad a priori se denominan hiperparámetros para distinguirlos de los parámetros θ del modelo de la distribución de datos E subyacente.

La clave del éxito del razonamiento bayesiano reside en disponer de una asunción previa apropiada. Con la correcta distribución de probabilidad a priori, incluso con pocos datos, se pueden hacer predicciones bayesianas con sentido. Por el contrario, una visión frecuentista hace menos asunciones a priori sobre la distribución de probabilidad, lo que otorga a este método una robustez mayor respecto al bayesiano, pero a la vez es impráctico a la hora de tomar decisiones en base a información limitada o incompleta, algo bastante frecuente en el campo de la investigación. Los estadísticos bayesianos sostienen que aun cuando distintas personas puedan proponer probabilidades a priori muy diferentes, la nueva evidencia que surge de nuevas observaciones va a lograr que las probabilidades subjetivas se aproximen cada vez más. Otros, sin embargo, sostienen que cuando distintas personas proponen probabilidades a priori muy diferentes, las probabilidades subjetivas a posteriori pueden no converger nunca, por más evidencias nuevas que se recolecten.

2.5.1.1. Distribuciones a Priori Informativas

Las distribuciones de probabilidad a priori en las que cada posible valor del parámetro θ tiene asignada una determinada probabilidad reciben el nombre de distribuciones a priori informativas. Este tipo de distribuciones expresa información específica y definida sobre una variable.

La evidencia preexistente que ya se ha tenido en cuenta es parte de la distribución a priori y, a medida que más evidencia se va acumulando, la distribución a priori se determina más en gran medida por la evidencia que por cualquier suposición original, siempre que la suposición original admitiese la posibilidad de lo que la evidencia está sugiriendo.

2.5.1.2. Distribuciones a Priori no Informativas

Las distribuciones de probabilidad a priori en las que la masa de probabilidad inicial se reparte por igual en el espacio paramétrico, indicando que no hay ninguna preferencia a priori sobre algunos valores del parámetro θ, reciben el nombre de distribuciones a priori no informativas. Este caso es frecuente ya que no siempre se tiene información a priori sobre la probabilidad de cada parámetro. Este tipo de distribuciones expresa información vaga o general sobre una variable. Se suelen denominar también distribuciones a priori objetivas puesto que suelen expresar información objetiva tal como que la variable es positiva o que la variable es menor que un cierto límite. El uso de una distribución a priori no informativa típicamente produce resultados que no son muy diferentes de los obtenidos con un análisis estadístico convencional.

La regla más simple y más antigua para la identificación de una distribución a priori no informativa es el principio de indiferencia, que asigna probabilidades iguales a todas las posibilidades. Como ejemplo de una distribución a priori no informativa, se puede pensar en una situación en la que se sabe que una bola ha sido escondida bajo uno de tres vasos, A, B o C, pero no se tiene ninguna otra información sobre la localización de la bola. En este caso, una distribución a priori uniforme de la forma $P(A) = P(B) = P(C) = 1/3$ parece intuitivamente la única opción razonable. Más formalmente, podemos ver que el problema sigue siendo el mismo si intercambiamos las etiquetas "A", "B" y "C" de los vasos. Por tanto, sería extraño que elegir una distribución a priori para la cual una permutación de las etiquetas de los vasos causara un cambio en nuestras predicciones sobre qué vaso oculta la pelota; la distribución a priori no informativa es la única que conserva esta invarianza. Si se acepta este principio de invariancia, entonces se puede ver que la distribución a priori uniforme es la distribución a priori lógicamente correcta para representar este estado de conocimiento. Cabe señalar que esta distribución a priori es objetiva en el sentido de ser la elección correcta para representar un estado particular de conocimiento, pero no es objetiva en el sentido de ser una función independiente del observador del mundo: en realidad la pelota existe bajo un vaso concreto, y sólo tiene sentido hablar de probabilidades en esta situación si hay un observador con un conocimiento limitado sobre el sistema.

2.5.1.3. Distribuciones a Priori Impropias

Los problemas prácticos asociados con las distribuciones a priori no informativas incluyen el requisito de que la distribución a posteriori sea propia. Las distribuciones a priori no informativas habituales sobre variables continuas y sin confinar son impropias. Esto no tiene por qué ser un problema, siempre que la distribución a posteriori resultante sea propia.

Atendiendo a la definición del teorema de Bayes mostrada en las ecuaciones (6) y (12), se puede apreciar que se obtendría el mismo resultado aunque las probabilidades a priori estuviesen multiplicadas por una constante dada o una variable continua aleatoria. Si la suma o integral en el denominador converge, las probabilidades a posteriori seguirán estando normalizadas incluso si los valores a priori no lo hacen, y por lo tanto las probabilidades a priori sólo necesitan ser especificadas en la proporción correcta. Llevando esta idea más lejos, en muchos casos la suma

o la integral de los valores a priori no necesita incluso ni ser finita para conseguir respuestas sensatas para las probabilidades a posteriori. Cuando este es el caso, la distribución a priori recibe el nombre de distribución a priori impropia. Esta denominación deriva de la definición de integral impropia como aquella integral definida en la que uno o ambos extremos del intervalo de integración se acercan a ∞ o a $-\infty$, o en la que la función a integrar no es continua en todo el intervalo de integración, pudiéndose presentar ambas situaciones a la vez. Las distribuciones a priori impropias permiten no imponer información subjetiva a priori.

2.5.2. Función de Verosimilitud

2. Séase el conjunto de datos E, elegir un modelo estadístico que describa su distribución de probabilidad dado el parámetro θ. Este modelo estadístico se denota $p(\mathrm{E}|\theta)$ y se conoce como distribución muestral o función de verosimilitud cuando se expresa como una función de θ.

En estadística, una función de verosimilitud es una función de los parámetros de un modelo estadístico que permite realizar inferencias sobre su valor a partir de un conjunto de observaciones. No debe confundirse verosimilitud con el término probabilidad. La probabilidad permite, a partir de una serie de parámetros conocidos, realizar predicciones acerca de los valores que toma una variable aleatoria. En contextos formales, verosimilitud se utiliza a menudo como sinónimo de probabilidad, pero en un contexto estadístico, se hace una distinción entre ambos términos dependiendo de los roles de los resultados o parámetros. Probabilidad se utiliza para describir una función del resultado dado un parámetro de valor fijo. Por ejemplo, si una moneda es lanzada 10 veces y está equilibrada, ¿cuál es la probabilidad de que en cada lanzamiento salga cara? Verosimilitud es usado para describir una función de un parámetro dado un resultado. Por ejemplo, si una moneda es lanzada 10 veces y en todas ellas ha salido cara, ¿cuál es la verosimilitud de que la moneda esté equilibrada?

La verosimilitud de un conjunto de parámetros θ dados los resultados E es igual a la probabilidad de esos resultados observados dados los parámetros, o expresado formalmente:

$$L(\theta|\mathrm{E}) = P(\mathrm{E}|\theta) \tag{20}$$

En este sentido, la verosimilitud es una versión inversa de la probabilidad condicional. Conocidos los datos E, la probabilidad condicional del parámetro θ es $P(\theta|\mathrm{E})$, pero si se conoce el parámetro θ pueden realizarse inferencias sobre el valor de los datos E a través del teorema de Bayes, según el cual:

$$P(\mathrm{E}|\theta) = \frac{P(\theta|\mathrm{E})\,P(\mathrm{E})}{P(\theta)} \tag{21}$$

2.5.3. Función de Verosimilitud Marginal

3. Conocidas la función de verosimilitud $p(\mathrm{E}|\theta)$ y la distribución de probabilidad a priori $p(\theta|\alpha)$, se procede a calcular la distribución de verosimilitud marginal o distribución marginal de los datos observados:

$$p(\mathrm{E}|\alpha) = \begin{cases} \sum_{\theta} p(\mathrm{E}|\theta)\,p(\theta|\alpha) & \text{en el caso discreto} \\ \int p(\mathrm{E}|\theta)\,p(\theta|\alpha)\,d\theta & \text{en el caso continuo} \end{cases} \tag{22}$$

2.5.4. Distribución de Probabilidad a Posteriori

4. Finalmente, se aplica el teorema de Bayes para determinar la distribución de probabilidad a posteriori, o distribución de probabilidad condicional del parámetro θ una vez tenida en cuenta la evidencia E, y así actualizar las creencias o conocimientos sobre el parámetro θ expresadas en la distribución a priori:

$$p(\theta|\mathrm{E},\alpha) = \frac{p(\mathrm{E}|\theta)\,p(\theta|\alpha)}{p(\mathrm{E}|\alpha)} = \begin{cases} \dfrac{p(\mathrm{E}|\theta)\,p(\theta|\alpha)}{\sum_\theta p(\mathrm{E}|\theta)\,p(\theta|\alpha)} & \text{en el caso discreto} \\ \dfrac{p(\mathrm{E}|\theta)\,p(\theta|\alpha)}{\int p(\mathrm{E}|\theta)\,p(\theta|\alpha)\,d\theta} & \text{en el caso continuo} \end{cases} \tag{23}$$

Cabe destacar que la función de verosimilitud marginal $p(\mathrm{E}|\alpha)$ es una constante de integración que asegura que la distribución de probabilidad a posteriori $p(\theta|\mathrm{E},\alpha)$ esté normalizada. $p(\mathrm{E}|\alpha)$ no depende del parámetro θ considerado, no proporcionando ninguna información adicional sobre la distribución de probabilidad a posteriori. Así, al igual que se explicó en la ecuación (14), la distribución de probabilidad a posteriori puede escribirse:

$$p(\theta|\mathrm{E},\alpha) \propto p(\mathrm{E}|\theta)\,p(\theta|\alpha) \tag{24}$$

Esta expresión es la distribución de probabilidad a posteriori sin normalizar y es un resultado útil para los cálculos porque implica que se pueden olvidar las constantes multiplicativas hasta el final del cálculo en modelos complicados.

Una característica importante de la inferencia bayesiana es el uso secuencial o iterativo de la fórmula de Bayes. Esta distribución de probabilidad a posteriori se convertirá en la nueva distribución de probabilidad a priori en cuanto una nueva evidencia o distribución de datos E sea observada y una nueva distribución de probabilidad a posteriori será calculada teniendo en cuenta esta evidencia, y así sucesivamente. Este procedimiento recibe el nombre de actualización bayesiana. Del mismo modo, la distribución a posteriori de un problema puede convertirse en la distribución a priori de otro problema distinto. Por último, sólo restaría evaluar el ajuste del modelo $p(\theta|\mathrm{E},\alpha)$ a los datos y la sensibilidad de las conclusiones a cambios en los supuestos iniciales del modelo.

En definitiva, los métodos bayesianos proporcionan: estimadores de los parámetros que tienen buenas propiedades estadísticas, una descripción simple de los datos observados, estimaciones de los datos faltantes y predicciones de futuras observaciones y una metodología computacional potente para la estimación, selección y validación de modelos.

2.5.4.1. Familias Conjugadas

La principal dificultad que surge en los problemas de inferencia bayesiana es tanto la licitación de la distribución a priori como el cálculo de la distribución a posteriori. La primera cuestión es importante ya que la inferencia que se realice posteriormente puede depender de la elección hecha de la distribución inicial, razón por la cual en muchos casos se recurre a distribuciones a priori no informativas, que no imponen unas condiciones muy fuertes sobre el parámetro θ, o bien se puede aprovechar parte de la información muestral para mejorar la distribución inicial, dando origen a las denominadas distribuciones intrínsecas a priori, de gran auge en la actualidad. En cuanto a la segunda opción, el cálculo de la distribución a posteriori no tiene por

qué conducir a una distribución tratable y, en ocasiones, hay que recurrir a métodos numéricos para poder trabajar con ellas.

Centrándonos en la segunda cuestión, interesa considerar familias de distribuciones a priori cuyas distribuciones a posteriori asociadas sean fáciles de calcular. En este sentido surge el concepto de distribuciones o familias a priori conjugadas. En teoría de probabilidad bayesiana, si la distribución de probabilidad a posteriori $p(\theta|\mathrm{E},\alpha)$ pertenece a la misma familia que la distribución de probabilidad a priori $p(\theta|\alpha)$, ambas distribuciones son entonces llamadas distribuciones conjugadas, y la distribución a priori es llamada distribución a priori conjugada para la función de verosimilitud $p(\mathrm{E}|\theta)$. Definido más formalmente, una familia de distribuciones a priori $F = p(\theta|\alpha)$ se dice conjugada de la familia de funciones de verosimilitud $p(\mathrm{E}|\theta)$ cuando para cualquier distribución a priori perteneciente a F, la distribución a posteriori $p(\theta|\mathrm{E},\alpha)$ también pertenece a F.

Un ejemplo es la familia gaussiana, la cual es conjugada a sí misma o autoconjugada respecto a una función de verosimilitud gaussiana. En este caso, si se dispone de una función de verosimilitud gaussiana, la mejor elección sería tomar una distribución a priori gaussiana porque nos garantizará que la distribución a posteriori también sea gaussiana.

Una distribución a priori conjugada es una conveniencia algebraica, dando una expresión de forma cerrada para la distribución a posteriori; de lo contrario sería necesario llevar a cabo una integración numérica compleja. Además, las distribuciones a priori conjugadas son más intuitivas, mostrando de forma más transparente cómo la función de verosimilitud actualiza la distribución a priori. La forma de la distribución a priori conjugada puede ser determinada normalmente mediante la inspección de la función de densidad de probabilidad.

2.6. El Factor de Bayes

En un artículo de 1935 y en su libro *"Theory of Probability"*, Harold Jeffreys desarrolló una metodología para cuantificar la evidencia en favor de una hipótesis o teoría científica. La pieza central era un número, hoy conocido como factor de Bayes. Este estudio aborda la comparación de las predicciones hechas por dos teorías científicas competidoras, introduciendo modelos estadísticos para representar la probabilidad de concordancia de los datos con cada una de ambas teorías y utilizando el teorema de Bayes para calcular la probabilidad a posteriori de que una de ambas teorías sea correcta.

2.6.1. Definición

Partiendo de una cierta distribución de datos $\mathrm{E} = \{e_1, \dots, e_n\}$, que se supone han surgido bajo una de las dos hipótesis o modelos H_1 o H_2, parametrizados por los parámetros vectoriales θ_1 y θ_2, de acuerdo con una función densidad de probabilidad, o verosimilitud, $p(\mathrm{E}|H_1)$ o $p(\mathrm{E}|H_2)$, y dadas unas funciones densidad de probabilidad a priori $p(H_1)$ y $p(H_2) = 1 - pr(H_1)$, los datos producen unas funciones densidad de probabilidad a posteriori $p(H_1|\mathrm{E})$ y $p(H_2|\mathrm{E}) = 1 - p(H_1|\mathrm{E})$. A partir del teorema de Bayes, se obtiene que:

$$p(H_k|\mathrm{E}) = \frac{p(\mathrm{E}|H_k)p(H_k)}{p(\mathrm{E})} = \frac{p(\mathrm{E}|H_k)p(H_k)}{p(\mathrm{E}|H_1)p(H_1) + p(\mathrm{E}|H_2)p(H_2)} \qquad (25)$$

donde $k = 1$ ó 2 en función de qué función densidad de probabilidad a posteriori se esté calculando, de modo que:

$$\frac{p(H_1|\mathrm{E})}{p(H_2|\mathrm{E})} = \frac{p(\mathrm{E}|H_1)p(H_1)}{p(\mathrm{E}|H_2)p(H_2)} \tag{26}$$

y la transformación entre el cociente de las funciones densidad de probabilidad a posteriori de ambas y el cociente de las funciones densidad de probabilidad a priori se obtiene multiplicando por:

$$B_{12} = \frac{p(\mathrm{E}|H_1)}{p(\mathrm{E}|H_2)} = \frac{\int\, p(\mathrm{E}|\theta_1, H_1)\, p(\theta_1|H_1)\, d\theta_1}{\int\, p(\mathrm{E}|\theta_2, H_2)\, p(\theta_2|H_2)\, d\theta_2} \tag{27}$$

A este número se le conoce como factor de Bayes y se define como el cociente o razón de las funciones de verosimilitud de ambas hipótesis, aunque naturalmente se ve afectado por la distribución de probabilidad a priori elegida inicialmente.

En el caso de tener varias medidas o distribuciones de datos independientes, por ejemplo $\mathrm{E} = \{e_1, \dots, e_n\}$ y $\mathrm{A} = \{a_1, \dots, a_n\}$ ambas surgidas bajo una de las dos hipótesis H_1 o H_2, se pueden calcular factores de Bayes separados para cada una de estas mediciones, y la combinación de ambas es simplemente su producto:

$$B_{12} = \frac{p(\mathrm{E}|H_1)}{p(\mathrm{E}|H_2)} \frac{p(\mathrm{A}|H_1)}{p(\mathrm{A}|H_2)} = B_{\mathrm{E}} B_{\mathrm{A}} \tag{28}$$

En estadística, el uso del factor de Bayes es una alternativa bayesiana a la comprobación clásica de hipótesis para evaluar la evidencia en favor de una determinada hipótesis. Entre sus ventajas, destaca que proporciona una forma de incorporar información externa a la evaluación de la evidencia sobre una hipótesis.

2.6.2. Interpretación

El factor de Bayes es un compendio de la evidencia aportada por los datos en favor de una hipótesis o teoría científica, representada por un modelo estadístico, en contraposición de otra. Generalmente, un valor de $B_{12} > 1$ significa que la hipótesis H_1 es sustentada con mayor fuerza por los datos en consideración que la hipótesis H_2. Harold Jeffreys proporcionó una escala para la interpretación del valor del factor de Bayes, tal que:

B_{12}	$log_{10}(B_{12})$ / bans	Solidez de la Evidencia
< 1	< 0	Negativa (Apoya H_2)
$1 - 10^{1/2}$	$0 - 1/2$	A penas vale la pena mencionar
$10^{1/2} - 10$	$1/2 - 1$	Substancial
$10 - 10^{3/2}$	$1 - 3/2$	Fuerte
$10^{3/2} - 100$	$3/2 - 2$	Muy Fuerte
> 100	> 2	Decisiva (Apoya H_1)

Tabla 1. Escala de referencia para comprobar la plausibilidad o validez de una hipótesis H_1 frente a otra hipótesis H_2 donde B_{12} es el factor de Bayes obtenido mediante la definición de la ecuación (27) y $log_{10}(B_{12})$ es el logaritmo en base 10 del factor de Bayes, también conocido como peso de la evidencia.

3. Método Probabilístico de Cross-Identificación

3.1. La Evidencia Observacional

A continuación, se proceden a aplicar los conocimientos y metodología de inferencia bayesiana mostrados previamente para desarrollar un formalismo probabilístico general para la cross-identificación de fuentes puntuales astronómicas atendiendo a las observaciones astrométricas de sus posiciones angulares en la esfera celeste.

Cuando se presentan una serie de posiciones observadas, a uno le interesaría saber si estas observaciones proceden realmente de la misma fuente astronómica o si, por el contrario, corresponden a objetos extragalácticos distintos. Si las coordenadas de estas observaciones se encuentran esparcidas por toda la esfera celeste, parece muy poco probable que sean medidas del mismo objeto astronómico, pero cuando estas coordenadas distan sólo una pequeña fracción de segundo de arco, se puede asegurar razonablemente que se ha encontrado una buena coincidencia. En estas circunstancias corresponde preguntarse cómo de buena es esta coincidencia o qué evidencia se tiene de que se haya logrado una coincidencia.

3.1.1. Modelado de la Astrometría

La astrometría es una disciplina experimental o técnica de la astronomía que se encarga de medir y estudiar la posición, paralaje y el movimiento propio de los astros u objetos astronómicos. Dentro de ella, la astrometría global se ocupa de la catalogación de las posiciones de los objetos sobre amplias regiones del cielo dando lugar a grandes catálogos astronómicos.

En primer lugar, es necesario examinar el significado de precisión astrométrica. En el proceso de calibración de las posiciones en un catálogo de fuentes astronómicas, se pueden caracterizar las propiedades de las observaciones mediante la comparación de las posiciones con valores astrométricos estándar e incluso corregir compensaciones sistemáticas. Sin embargo, sigue habiendo una dispersión aleatoria alrededor de las posiciones verdaderas. Esta incertidumbre es a menudo modelada como una distribución normal, y los catálogos astronómicos suelen indicar su precisión mediante un único valor σ. En general, nuestro entendimiento de la astrometría es descrito por una función de densidad de probabilidad (PDF) que incluso puede variar en el cielo. De este modo, se parametriza un modelo o hipótesis M según el cual un cierto objeto astronómico está en la esfera celeste usando un vector normal unitario tridimensional $\boldsymbol{m}$, y se escribe $p(\boldsymbol{x}|\boldsymbol{m}, M)$ para denotar la función densidad de probabilidad de que un objeto astronómico sea observado en la posición $\boldsymbol{x}$ cuando su verdadera localización viene dada por $\boldsymbol{m}$. Como cualquier otra función densidad de probabilidad, esta función está normalizada:

$$\int p(\boldsymbol{x}|\boldsymbol{m}, M)\, d^3x = 1 \tag{29}$$

A continuación, se considera una única fuente observada en la posición $\boldsymbol{x_1}$ y se aplica el teorema de Bayes para hallar la función densidad de probabilidad a posteriori de la verdadera localización $\boldsymbol{m}$ del objeto astronómico dados los datos obtenidos:

$$p(\boldsymbol{m}|\boldsymbol{x_1}, M) = \frac{p(\boldsymbol{x_1}|\boldsymbol{m}, M)\, p(\boldsymbol{m}|M)}{p(\boldsymbol{x_1}|M)} \tag{30}$$

donde $p(\boldsymbol{x_1}|\boldsymbol{m}, M)$ es la función de verosimilitud o función densidad de probabilidad de observar el objeto astronómico en $\boldsymbol{x_1}$ si su verdadera posición es $\boldsymbol{m}$, la función densidad de probabilidad a priori trivial $p(\boldsymbol{m}|M)$ de que $\boldsymbol{m}$ esté en la esfera celeste es expresada mediante una función delta de Dirac:

$$p(\boldsymbol{m}|M) = \frac{1}{4\pi}\,\delta(|\boldsymbol{m}| - 1) \tag{31}$$

y la función de verosimilitud marginal $p(\boldsymbol{x_1}|M)$ actúa como una constante de normalización, garantizando que se cumpla el teorema de la probabilidad total:

$$p(\boldsymbol{x_1}|M) = \int p(\boldsymbol{x_1}|\boldsymbol{m}, M)\, p(\boldsymbol{m}|M)\, d^3\, m \tag{32}$$

Esta es la mejor comprensión que se tiene de la posible ubicación de un objeto astronómico en el cielo (previamente a medir su posición real) cuando es visto en una posición aparente $\boldsymbol{x_1}$ asumiendo una precisión astrométrica $p(\boldsymbol{x_1}|\boldsymbol{m}, M)$ obtenida a partir de la calibración.

3.1.2. El Factor de Bayes Posicional

Con múltiples observaciones a través de diversos instrumentos de posiblemente diferentes precisiones astrométricas, se procede ahora a calcular la probabilidad de que todas estas observaciones procedan de la misma fuente astronómica. De este modo, se introduce el factor de Bayes para probar esta hipótesis H en contra del caso K en que estas observaciones procedan de fuentes astronómicas separadas. Después de que las $D = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\}$ ubicaciones en el cielo son obtenidas, se calcula la razón de las distribuciones de probabilidad a posteriori y a priori de cada hipótesis. El factor de Bayes se define como la razón de ambas posibilidades:

$$B(H, K|D) = \frac{p(H|D)/\,p(H)}{p(K|D)/\,p(K)} = \frac{p(H|D)\,p(K)}{p(K|D)\,p(H)} \tag{33}$$

el cual, después de aplicar el teorema de Bayes a $p(H|D)$ y $p(K|D)$, se convierte en la razón de las funciones de verosimilitud de ambas hipótesis, tal y como se definió en la ecuación (27):

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \tag{34}$$

El presente cálculo está hecho parametrizando los dos modelos H y K e integrando las funciones de verosimilitud para todo el espacio de configuración. La hipótesis H establece que las posiciones son de una única fuente astronómica y, por tanto, puede ser parametrizada por una única localización común $\boldsymbol{m}$. Debido a la independencia de las medidas en D, la función densidad de probabilidad conjunta es simplemente el producto de las precisiones astrométricas $p_1, \dots, p_n$, y la integral se simplifica en:

$$p(D|H) = \int\, p(\boldsymbol{m}|H)\, p(D|\boldsymbol{m}, H) d^3m = \int\, p(\boldsymbol{m}|H) \prod_{i=1}^{n} p_i\,(\boldsymbol{x_i}|\boldsymbol{m}, H)\, d^3m \tag{35}$$

Por otro lado, la hipótesis alternativa K es parametrizada por un conjunto de n vectores de posición $\{\boldsymbol{m_i}\}$, y la integral se factoriza en el producto de las componentes independientes:

$$p(D|K) = \prod_{i=1}^{n} \left[\int p(\boldsymbol{m_i}|K)\, p_i(\boldsymbol{x_i}|\boldsymbol{m_i}, K)\, d^3m_i \right] \tag{36}$$

De modo que el factor de Bayes, mostrado en la ecuación (34), se puede escribir como:

$$B(H,K|D) = \frac{\int\ p(\boldsymbol{m}|H)\ \prod_{i=1}^{n} p_i\ (\boldsymbol{x_i}|\boldsymbol{m},H)\ d^3m}{\prod_{i=1}^{n}[\int\ p(\boldsymbol{m_i}|K)\ p_i(\boldsymbol{x_i}|\boldsymbol{m_i},K)\ d^3m_i\ ]} \tag{37}$$

Cuando el factor de Bayes es grande, las observaciones apoyan la hipótesis H de que la asociación es una coincidencia para el mismo objeto astronómico, mientras que si el factor de Bayes es pequeño, las observaciones sustentan la hipótesis K de que se han identificado fuentes astronómicas separadas. Esto se puede esquematizar en la siguiente tabla:

$B(H,K\|D)$	Solidez de la Evidencia
$\gg 1$	Hipótesis H es cierta (1 sola fuente)
~ 1	No es convincente
< 1	Hipótesis K es cierta (distintas fuentes)

Tabla 2. Escala de referencia para comprobar la plausibilidad o validez de la hipótesis H según la cual las observaciones proceden de la misma fuente astronómica frente a la hipótesis K según la cual las observaciones proceden de distintas fuentes astronómicas, donde $B(H,K|D)$ es el factor de Bayes obtenido mediante la definición de la ecuación (37).

3.1.3. La Distribución Normal

Las distribuciones normales surgen con frecuencia en la naturaleza donde permiten ajustar numerosos fenómenos naturales cuyos efectos desempeñan un papel importante en modelar la función densidad de probabilidad. Al tratarse el cielo observado de una superficie esférica, se hace necesario trabajar sobre una esfera. Aunque muchos de los argumentos habituales no se sostienen sobre variedades topológicas cerradas, es posible introducir una función análoga a la distribución normal sobre la esfera. Esta distribución normal esférica es a menudo elegida para caracterizar la precisión de observaciones astronómicas; por lo tanto, es de gran importancia para comprender sus propiedades y aplicar el marco bayesiano descrito anteriormente.

La distribución normal esférica en su forma normalizada usando la notación vectorial tridimensional anterior se escribe:

$$N(\boldsymbol{x}|\boldsymbol{m},w) = \frac{w\delta(|\boldsymbol{x}|-1)}{4\pi \sinh w} e^{w\boldsymbol{m}\boldsymbol{x}} \tag{38}$$

donde el peso w es típicamente muy grande. Cuando este es el caso, el peso w se puede relacionar con el más intuitivo parámetro de precisión σ mediante la ecuación:

$$w = \frac{1}{\sigma^2} \tag{39}$$

Por ejemplo, cuando σ es del orden de un arcosegundo, el peso toma valores del orden de 10^{10}. Habiendo observado un conjunto de posiciones $D = \{\boldsymbol{x_i}\}$ de forma independiente con sus correspondientes pesos w_i y, debido a que la función $N(\boldsymbol{x}|\boldsymbol{m},w)\ p(\boldsymbol{m}|M)$ es simétrica en $\boldsymbol{x}$ y en $\boldsymbol{m}$ para la distribución de probabilidad a priori trivial $p(\boldsymbol{m}|M)$ de la ecuación (31) y las funciones de densidad de probabilidad están normalizadas, se puede calcular analíticamente el factor de Bayes para las dos hipótesis H y K mostrado en las ecuaciones (34) y (37), obteniendo que:

$$B(H,K|D) = \frac{\sinh w}{w} \prod_{i=1}^{n} \frac{w_i}{\sinh w_i} \tag{40}$$

con

$$w = \left| \sum_{i=1}^{n} w_i \boldsymbol{x_i} \right| \tag{41}$$

La deducción matemática detallada de la ecuación (40) a partir de la ecuación (34) puede encontrarse en el Apéndice I adjunto al final de la memoria.

La precisión astrométrica de las observaciones reales es casi siempre extremadamente alta en el sentido absoluto, por lo que merece la pena examinar la aproximación del factor de Bayes en este límite. Así, suponiendo una gran precisión astrométrica σ, puesto que los pesos w son típicamente muy grandes, y pequeñas separaciones angulares entre las fuentes astronómicas, el factor de Bayes de la ecuación (40) toma la forma de:

$$B = 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp\left[- \frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right] \tag{42}$$

donde todas las sumas y productos se ejecutan sobre todos los miembros de la lista de elementos de los n catálogos astronómicos utilizados. La deducción matemática detallada de la ecuación (42) a partir de la ecuación (40) puede encontrarse en [1]. En el caso concreto de comparar observaciones astrométricas procedentes de sólo 2 catálogos astronómicos, sustituyendo $w_i = 1/\sigma_i^2$ de acuerdo con la ecuación (39), la ecuación (42) se simplifica, obteniendo la siguiente expresión correspondiente a la parte puramente posicional del factor de Bayes:

$$B_{ij_{\mathrm{pos}}} = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp\left[- \frac{\psi_{ij}^2}{2\,(\sigma_1^2 + \sigma_2^2)} \right] \tag{43}$$

donde ψ_{ij} es la distancia angular entre la fuente astronómica i del catálogo 1 y la fuente j del catálogo 2, y, σ_1 y σ_2, son los parámetros de precisión o incertidumbres de posición de cada uno de los catálogos astronómicos comparados.

La incertidumbre en la posición de las fuentes astronómicas de un catálogo astronómico viene dada por las limitaciones instrumentales de los telescopios e instrumentos de medición utilizados y es proporcionada por el proyecto o estudio que haya desarrollado ese catálogo. Como se ha mencionado anteriormente, es habitual modelar la incertidumbre de la posición o precisión astrométrica mediante una distribución normal, también denominada distribución gaussiana o de Gauss, cuya forma genérica es:

$$g(\theta) = \frac{1}{2\pi\sigma^2} e^{-\frac{\theta^2}{2\sigma^2}} \tag{44}$$

donde el primer cociente es simplemente una constante de normalización, σ es la incertidumbre en la posición que interesa obtener para la ecuación (43) y θ es la separación angular entre las coordenadas del apuntado del telescopio y el punto del espacio que se está considerando. No obstante, los catálogos no suelen proporcionar directamente el valor de la dispersión σ, sino que habitualmente facilitan el valor de la achura a media altura (FWHM).

La FWHM es una medida de la extensión de una función, que viene dada por la diferencia entre los dos valores extremos de la variable independiente en los que la variable dependiente es igual a la mitad de su valor máximo. En otras palabras, tal y como se muestra en la siguiente representación gráfica de la ecuación (44), es la anchura de la campana de Gauss en el punto el que la altura ha caído a la mitad:

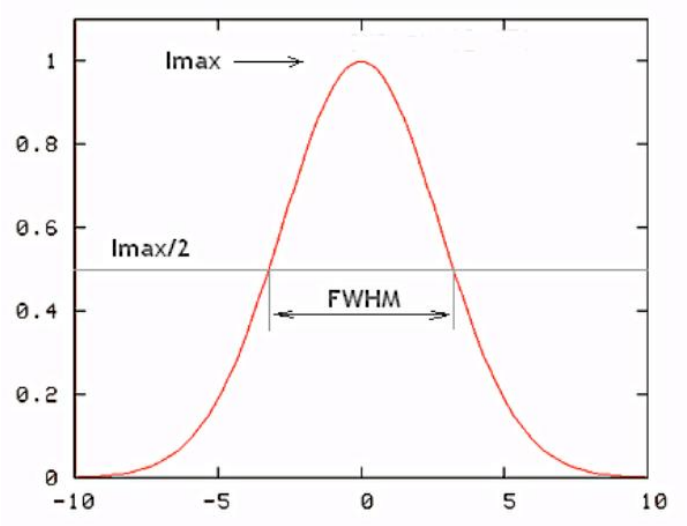

Figura 3. Representación gráfica de la distribución normal o gaussiana mostrada en la ecuación (44) para visualizar el concepto de anchura a media altura (FWHM) de una cierta función.

De este modo, conociendo la anchura a media altura FWHM del catálogo utilizado, se puede obtener el valor de su incertidumbre en la posición σ a través de la ecuación:

$$\sigma = \frac{\text{FWHM}}{2\sqrt{2\ln(2)}} \tag{45}$$

3.2. Información Adicional a Priori

Naturalmente, el formalismo bayesiano introducido al principio de esta memoria no es exclusivo de observaciones astrométricas. De hecho, es bastante sencillo incluir otras cantidades medidas en los cálculos. Esto es especialmente importante cuando se trata con múltiples coincidencias. Seleccionar la correcta combinación de fuentes astronómicas a partir de diversas configuraciones espacialmente similares es un problema degenerado que requiere información adicional para poder resolverse. El uso de información fotométrica es una elección natural debido a su amplia disponibilidad; sin embargo, su aplicación requiere mayores suposiciones sobre las distribuciones espectrales de energía (SED) de las fuentes astronómicas.

3.2.1. Modelado del Corrimiento al Rojo

En nuestro caso, como información adicional se han considerado las mediciones del corrimiento al rojo, tanto fotométricas como espectroscópicas, de los objetos extragalácticos de los catálogos astronómicos comparados. A continuación, se procede a aplicar la metodología de inferencia bayesiana mostrada previamente para desarrollar un formalismo probabilístico general para la cross-identificación de fuentes puntuales astronómicas atendiendo a las observaciones de sus corrimientos al rojo.

En primer lugar, partiendo de la definición del factor de Bayes de la ecuación (34), para la hipótesis H según la cual todas las observaciones proceden de la misma fuente astronómica frente a la hipótesis K según la cual las observaciones proceden de fuentes astronómicas separadas, en el caso de comparar observaciones procedentes de 2 catálogos se tiene, por ejemplo, para cada posible asociación de dos objetos, un par de valores observados del corrimiento al rojo $\{z_1, z_2\}$ con sus correspondientes incertidumbres $\{\sigma_{z_1}, \sigma_{z_2}\}$. Así, la hipótesis H queda parametrizada por un único valor del corrimiento al rojo z, correspondiente al objeto que se ha identificado de concernir ambas observaciones una coincidencia:

$$p(D|H) = \int p(z|H)\, p(D|z,H) d^3m = \int p(z|H) \prod_{i=1}^{n} p_i\,(z_i|z,H)\, dz \qquad (46)$$

Por su lado, la hipótesis K es parametrizada por un conjunto de $\{z_i^0\}$ valores del corrimiento al rojo, cada uno correspondiente a cada uno de las diferentes fuentes astronómicas medidas:

$$p(D|K) = \prod_{i=1}^{n} \left[\int p\left(z_i^0 \middle| K\right) p_i\left(z_i \middle| z_i^0, K\right) d^3 m_i \right] \qquad (47)$$

De modo que el factor de Bayes, mostrado en la ecuación (34), se puede escribir como:

$$B(H,K|D) = \frac{\int p(z|H) \prod_{i=1}^{n} p_i\,(z_i|z,H)\, dz}{\prod_{i=1}^{n} \left[\int p\left(z_i^0 \middle| K\right) p_i\left(z_i \middle| z_i^0, K\right) d^3 m_i \right]} \qquad (48)$$

donde $p(z|H)$ y $p\left(z_i^0 \middle| K\right)$ son las funciones densidad de probabilidad a priori y, $p_i(z_i|z,H)$ y $p_i\left(z_i \middle| z_i^0, K\right)$ son las funciones de verosimilitud. Las fórmulas anteriores se han escrito para el caso general de comparar n catálogos. En el caso de comparar las observaciones de sólo 2 catálogos (pares de objetos astronómicos), se toma $n = 2$.

3.2.1.1. Distribuciones de probabilidad a Priori

Las cantidades $p(z|H)$ y $p\left(z_i^0 \middle| K\right)$ mostradas en la ecuación (48) son las funciones de densidad de probabilidad a priori sobre las distribuciones del corrimiento al rojo del objeto (u objetos) astronómicos medidos. La hipótesis H contempla la existencia de un único objeto con una distribución de probabilidad a priori del corrimiento al rojo $p(z|H)$ mientras que en la hipótesis alternativa K cada objeto podría tener, aunque no necesariamente, una distribución de probabilidad a priori diferente. Por ejemplo, se podría considerar que las galaxias de uno de los catálogos son principalmente galaxias normales seleccionadas en el rango óptico con una distribución de probabilidad del corrimiento al rojo cuyos picos tienen valores bajos, mientras que el segundo catálogo podría estar formado por una muestra de cuásares seleccionados en el rango del radio con una distribución de probabilidad del corrimiento al rojo con máximos en torno a $z \sim 1$. Aquí se deberían utilizar modelos de distribuciones del corrimiento al rojo físicamente realistas, pero por simplicidad se va a asumir una distribución de probabilidad a priori no informativa plana entre $z_{min} = 0$ y un cierto valor z_{max}:

$$p(z|H) = p\left(z_i^0 \middle| K\right) = p_0 \qquad (49)$$

donde el valor de la constante p_0 viene dado por la condición de normalización:

$$\int_0^{z_{max}} p(z|H)\, dz = \int_0^{z_{max}} p_0 dz = p_0 \cdot z_{max} = 1 \tag{50}$$

Por lo tanto, $p_0 = 1/z_{max}$. En este trabajo se va a utilizar un valor $z_{max} = 2$. Es necesario recalcar que la adopción de esta distribución de probabilidad a priori plana es una simplificación y que el poder del método de cross-identificación sería mucho mayor si se usase un modelo más realista de la distribución de probabilidad del corrimiento al rojo de fuentes extragalácticas. Tales modelos pueden encontrarse, por ejemplo, en [3] y en sus respectivas referencias.

3.2.1.2. Funciones de Verosimilitud

Los factores $p_i(z_i|z,H)$ y $p_i\left(z_i\middle|z_i^0,K\right)$ mostrados en la ecuación (48) son la función de verosimilitud de observar un valor del corrimiento al rojo z_i cuando el verdadero valor del corrimiento al rojo es z bajo la hipótesis H, en la ecuación (46), y la función de verosimilitud de observar un valor del corrimiento al rojo z_i cuando el verdadero valor del corrimiento al rojo es z_i^0 bajo la hipótesis K, en la ecuación (47).

En este caso, se considera que ambas funciones de verosimilitud obedecen a una distribución gaussiana:

$$p_i(z_i|z,H) = \frac{1}{\sqrt{2\pi}\sigma_{z_i}} \exp\left[-\frac{1}{2}\frac{(z_i - z)^2}{2\sigma_{z_i}^2}\right] \tag{51}$$

$$p_i\left(z_i\middle|z_i^0,K\right) = \frac{1}{\sqrt{2\pi}\sigma_{z_i}} \exp\left[-\frac{1}{2}\frac{(z_i - z_i^0)^2}{2\sigma_{z_i}^2}\right] \tag{52}$$

Debe tenerse en cuenta que esta función de verosimilitud gaussiana es una buena aproximación sólo cuando las incertidumbres σ_{z_i} son más pequeñas que las estimaciones del corrimiento al rojo z_i. Este es siempre el caso para corrimientos al rojo espectroscópicos, pero puede no serlo para corrimientos al rojo fotométricos. Sin embargo, se procede con esta aproximación en aras de la simplicidad.

3.2.1.3. Factor de Bayes para 2 catálogos astronómicos

Para $n = 2$ catálogos astronómicos y las aproximaciones descritas anteriormente (función densidad de probabilidad a priori plana y funciones de verosimilitud gaussianas), se obtiene la siguiente expresión para el factor de Bayes debido exclusivamente a las mediciones del corrimiento al rojo:

$$B_z = \frac{1}{p_0\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}} \exp\left[-\frac{(z_1 - z_2)^2}{2(\sigma_{z_1}^2 + \sigma_{z_2}^2)}\right] = \frac{z_{max}}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}} \exp\left[-\frac{(z_1 - z_2)^2}{2(\sigma_{z_1}^2 + \sigma_{z_2}^2)}\right] \tag{53}$$

donde, en este trabajo, se considera $z_{max} = 2$. Al igual que el factor de Bayes puramente posicional de la ecuación (43), esta es una cantidad adimensional, y refleja la idea intuitiva que se tiene de la distancia entre medidas (el factor de Bayes decrece rápidamente si z_1 y z_2 son muy diferentes). Al utilizar está formula, es recomendable tener en mente todas las aproximaciones y simplificaciones que han conducido a ella.

Si uno de los catálogos astronómicos registra corrimientos al rojo exactos, es decir corrimientos al rojo espectroscópicos sin errores asociados, la ecuación (53) se simplifica considerablemente.

Asumiendo, por ejemplo, que el catálogo 2 tiene corrimientos al rojo exactos, $\sigma_{z_2} = 0$, la ecuación (53) se convierte en:

$$B_z = \frac{1}{p_0\sqrt{2\pi}}\frac{1}{\sigma_{z_1}} \exp\left[-\frac{(z_1 - z_2)^2}{2\sigma_{z_1}^2}\right] = \frac{z_{max}}{\sqrt{2\pi}\sigma_{z_1}} \exp\left[-\frac{(z_1 - z_2)^2}{2\sigma_{z_1}^2}\right] \qquad (54)$$

Si ambos catálogos tienen desplazamientos al rojo exactos, la regla bayesiana se simplifica a una operación binaria lógica: los objetos astronómicos coinciden si y sólo si tienen los mismos valores de corrimiento al rojo.

3.3. Factor de Bayes Conjunto

Como ya se ha comentado anteriormente, el análisis bayesiano es inherentemente recursivo. Tan pronto como se obtienen nuevas mediciones y se calcula la función densidad de probabilidad a posteriori, esta se convierte en la función densidad de probabilidad a priori en posteriores estudios. Esta es una propiedad muy potente y simplifica enormemente los cálculos. Una consecuencia de esto, como también se manifestó en la definición del factor de Bayes de la ecuación (28), es que el factor de Bayes combinado de las mediciones astrométricas y del corrimiento al rojo es simplemente el producto de ambas:

$$\mathrm{B} = \mathrm{B_{pos}}\mathrm{B_z} \qquad (55)$$

donde B_{pos} es la parte puramente posicional del factor de Bayes mostrada en la ecuación (43) y B_{z} es el factor de Bayes debido exclusivamente a las mediciones del corrimiento al rojo, para el cual puede tomarse la expresión de la ecuación (53), en caso de que ambos catálogos tengan incertidumbres asociadas a sus medidas de z, o la expresión de la ecuación (54) en caso de que las medidas de z de uno de los catálogos carezcan de errores asociados.

Esto significa que uno puede analizar simplemente las asociaciones espaciales primero y considerar mediciones adicionales y funciones de densidad de probabilidad a priori de las propiedades físicas de las fuentes astronómicas en los subsiguientes pasos, si fuese necesario. De este modo, nuestro enfoque bayesiano, simétrico en todas las observaciones, constituye la base de un marco de trabajo para la cross-identificación de fuentes astronómicas, donde no sólo se utiliza información astrométrica, sino que propiedades físicas, como es el corrimiento al rojo o redshift, también pueden ser consideradas de un modo natural.

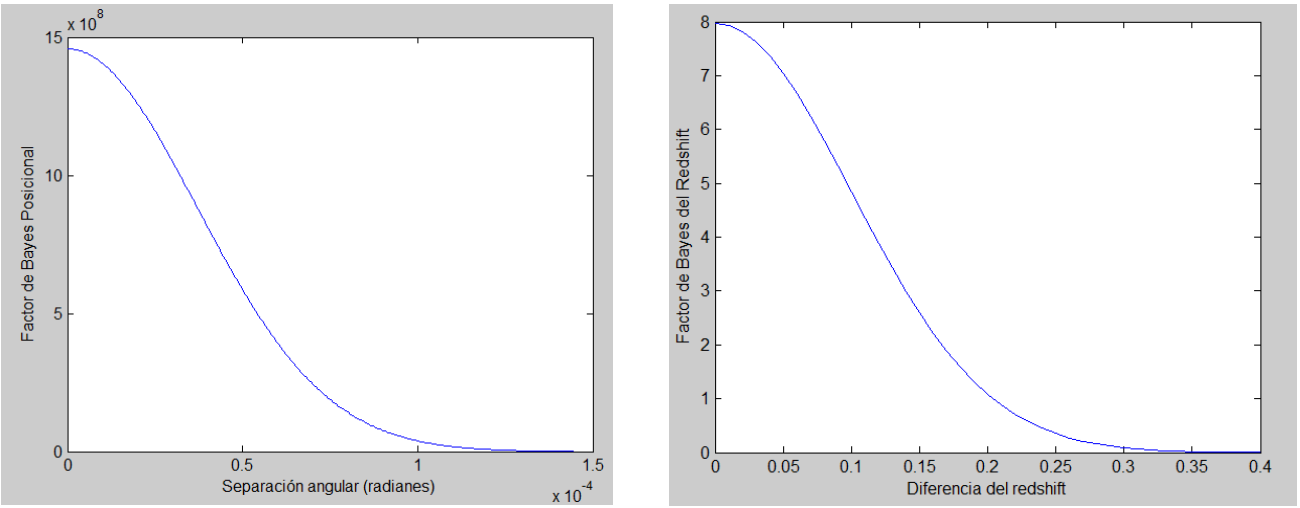

Figura 4. Representación gráfica del factor de Bayes posicional B_{pos} en función de la separación angular ψ_{ij} entre las fuentes astronómicas, de acuerdo con la ecuación (43), y del factor de Bayes del corrimiento al rojo B_z en función de la discrepancia $z_1 - z_2$ de los redshifts de las fuentes astronómicas, de acuerdo con la ecuación (54), para el caso de los catálogos Herschel ATLAS y GAMA cotejados en este trabajo. Ambas gráficas se han realizado usando MATLAB y se ha tomado $\sigma_1 = 3.70 \cdot 10^{-5}$ rad para Herschel ATLAS, $\sigma_2 = 1.44 \cdot 10^{-6}$ rad para GAMA y un valor $\sigma_{z_1} = 0.1$ genérico para la incertidumbre de los corrimientos al rojo.

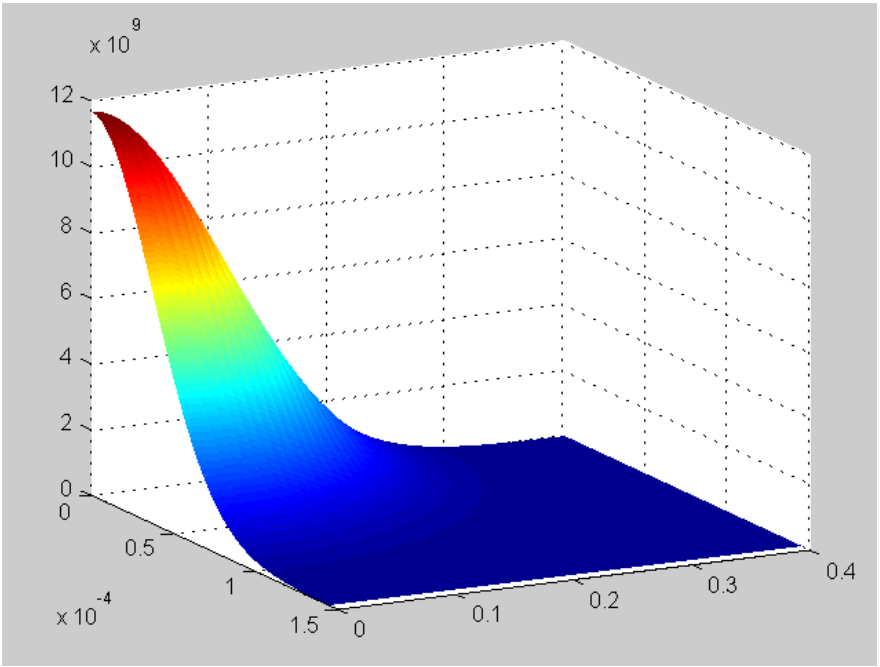

Figura 5. Representación gráfica del factor de Bayes combinado de las mediciones astrometricas y del corrimiento al rojo en función de la separación angular ψ_{ij} y de la discrepancia $z_1 - z_2$ de los redshifts entre las fuentes astronómicas, de acuerdo con la ecuación (55), para el caso de los catálogos Herschel ATLAS y GAMA cotejados en este trabajo. Esta gráfica ha sido realizada usando MATLAB y se ha tomado $\sigma_1 = 3.70 \cdot 10^{-5}$ rad para Herschel ATLAS, $\sigma_2 = 1.44 \cdot 10^{-6}$ rad para GAMA y un valor $\sigma_{z_1} = 0.1$ genérico para la incertidumbre de los corrimientos al rojo.

4. Catálogos Astronómicos

Como se ha mencionado, en este trabajo se ha implementado una herramienta para la cross-identificación de objetos extragalácticos pertenecientes a dos catálogos diferentes. A continuación se procede a la descripción de los dos catálogos astronómicos utilizados:

4.1. Catálogo Herschel-ATLAS

Herschel ATLAS (Astrophysical Terahertz Large Area Survey) es el mayor proyecto astronómico de estudio en tiempo abierto adjudicado al Observatorio Espacial Herschel de la Agencia Espacial Europea y es llevado a cabo por una colaboración internacional de instituciones que utilizan los instrumentos PACS y SPIRE de este observatorio para estudiar una amplia zona del cielo en 5 bandas fotométricas (100 μm, 160 μm, 250 μm, 350 μm y 500 μm) que abarcan desde el infrarrojo lejano hasta el rango submilimétrico. En total, se han concedido a este proyecto 600 horas de tiempo en Herschel para estudiar una vasta región de 550 grados cuadrados, cuatro veces mayor que todos los otros estudios extragalácticos de Herschel juntos.

4.1.1. Observatorio Espacial Herschel

El Observatorio Espacial Herschel de la Agencia Espacial Europea (anteriormente llamado Far Infrarred and Sub-milimetre Telescope o FIRST, y renombrado en honor de Sir William Herschel, quien en 1800 demostró la existencia de radiación infrarroja) fue llevado al espacio, junto con la misión Planck encargada de estudiar la radiación de fondo cósmico de microondas, por una lanzadora Ariane 5 ECA el 14 de mayo de 2009 y siguió funcionando a pleno rendimiento hasta que se agotó el helio líquido refrigerante necesario para mantenerse frio, el 29 de abril de 2013. Es hasta la fecha el mayor observatorio infrarrojo lanzado al espacio. Poco después del lanzamiento ambas misiones se separaron, operando de manera independiente. Tras un viaje aproximado de 60 días desde la Tierra, Herschel entró en su órbita operacional alrededor del segundo punto de Lagrange (L_2) del sistema Sol-Tierra, a unos 1.5 millones de kilómetros de la Tierra en una dirección diametralmente opuesta al Sol, con una amplitud media de 700000 km y un periodo orbital de unos 178 días, para una misión nominal de 3 años de vida con una posible prórroga de 1 año. De este modo, alrededor de 7000 horas de observación por año serán accesibles para los astrónomos de todo el mundo.

Figura 6. Ilustración de la configuración de lanzamiento de la lanzadera Ariane 5 ECA encargada de poner en órbita el Observatorio Espacial Herschel (parte superior) y el satélite Planck (parte inferior).

Los puntos de Lagrange, también denominados puntos L o puntos de libración, son las cinco posiciones en un sistema orbital donde un pequeño objeto de masa despreciable, sólo afectado por la gravedad, puede estar teóricamente estacionario respecto a dos objetos más grandes, como es el caso de un satélite artificial con respecto a la Tierra y la Luna. Estas posiciones surgen como soluciones estacionarias del problema de los tres cuerpos restringido a órbitas circulares. Concretamente, el punto L_2 se ubica en la línea definida por los dos objetos grandes, y más allá del de menor masa de ambos. El punto L_2 del sistema Sol-Tierra es un buen punto para los observatorios espaciales, porque un objeto alrededor de L_2 mantendría la misma orientación con respecto al Sol y la Tierra y la calibración de los instrumentos de medición y el blindaje del satélite serían más sencillos. La reducción de los efectos de la luz extraviada y la modulación térmica del Sol, la Tierra y la Luna, así como limitaciones técnicas de la nave espacial como las telecomunicaciones a tierra y el coste de transferencia a la órbita, impulsaron la elección del punto L_2 como la órbita de Herschel.

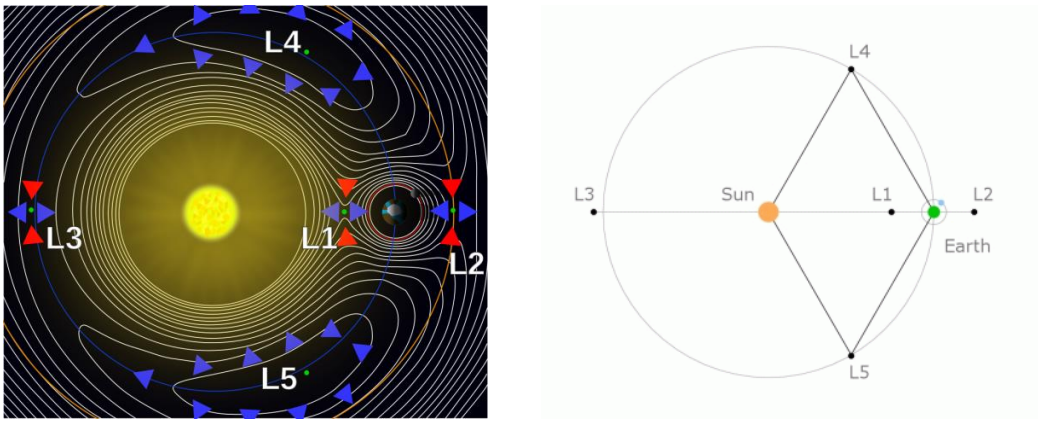

Figura 7. Ilustración de las curvas de potencial del sistema Sol-Tierra, mostrando los cinco puntos de Lagrange y en concreto el punto L_2 en que se ubica el Observatorio Espacial Herschel.

4.1.1.1. Objetivos de Herschel

Entre los objetivos básicos del Observatorio Espacial Herschel destacan:

- Estudiar la formación y posterior evolución de las galaxias en el Universo primitivo.
- Investigar la creación de estrellas y su interacción con el medio interestelar.
- Observar la composición química de las atmósferas y superficies de cometas, planetas y satélites.
- Examinar la química molecular del Universo.

La elección de Herschel para llevar a cabo este proyecto es debida a que procesos como el nacimiento de galaxias en el universo primitivo se pueden estudiar mejor con telescopios infrarrojos situados en el espacio y por lo tanto libres de las restricciones impuestas por la atmosfera terrestre. En este sentido, Herschel es el mayor y más potente telescopio infrarrojo jamás lanzado al espacio y es el primer observatorio espacial en poder cubrir desde la banda del infrarrojo lejano a la banda submilimétrica del espectro ($55\ \mu m - 672\ \mu m$), abriendo por primera vez una ventana a una parte casi inexplorada del espectro que no se puede observar bien desde el suelo, el misterioso y oculto universo frio.

Grandes partes del Universo son demasiado frías para irradiar en el rango de longitud de onda visible o en longitudes de onda más cortas. El estudio de estos objetos más fríos sólo es posible

mediante la observación en el espectro infrarrojo o incluso en longitudes de onda submilimétricas, rango de operación de Herschel. Además, muchos objetos de gran interés para los astrónomos están ocultos dentro o detrás de las nubes de gas y polvo. En las primeras etapas de su formación, las estrellas y los planetas están rodeados por las nubes de gas y polvo de las que se están creando. Y los núcleos galácticos y la mayor parte de los restos de los inicios del universo también están ocultos a la vista por nubes de polvo. Las partículas de polvo en estas nubes son comparables en tamaño a la longitud de onda de la luz visible y por lo tanto son eficientes en la dispersión o absorción de radiación en estas longitudes de onda. Por su lado, la radiación infrarroja se ve menos afectada por estas nubes, ya que la cantidad de dispersión producida decrece al aumentar la longitud de onda, es decir cuanto mayor sea la longitud de onda, más gruesa es la nube de polvo que puede atravesar.

El vapor de agua en la atmósfera terrestre presenta una gran absorbancia a la radiación en las bandas infrarroja y submilimétrica, haciendo las observaciones con base en tierra en estas longitudes de onda muy limitadas o imposibles. Por ello, la ubicación en el espacio de un observatorio como Herschel, que opera en este rango del espectro electromagnético, es la única solución verdaderamente satisfactoria a este problema.

En definitiva, el observatorio Herschel permite explorar en el infrarrojo lejano aún más que cualquier misión anterior, estudiando las regiones frías y polvorientas del cosmos que de otro modo serían invisibles, tanto en el infrarrojo lejano como en el cercano, a un nivel de detalle nunca antes visto.

Figura 8. Ilustración del Observatorio Espacial Herschel.

4.1.1.2. Objetivos de Herschel ATLAS

Uno de los principales propósitos de Herschel ATLAS era obtener el primer estudio imparcial del Universo local a longitudes de onda submilimétricas y del infrarrojo lejano, y como resultado fue diseñado para solaparse con los grandes estudios ópticos e infrarrojos existentes. El buque insignia de este proyecto científico es el estudio del polvo y la formación de estrellas de polvo oscurecido en ~ 10^5 galaxias en el universo cercano ($z < 0.3$). No obstante, persigue otros muchos objetivos científicos, entre los que destacan: estimar las contribuciones relativas de las galaxias con formación estelar y el efecto Sunyaev-Zeldóvich en grupos de elevado corrimiento al rojo, investigar la evolución de los perfiles de masas de las galaxias, la relación entre la formación de estrellas y agujeros negros en los quásares, la estructura a gran escala del universo en escalas $100 - 1000$ Mpc, y llevar a cabo el primer censo de núcleos preestelar y protoestrellas a latitudes galácticas elevadas.

Dentro del estudio Herschel ATLAS cabe destacar el Programa Universo Local, encargado de analizar la superposición de los datos en el infrarrojo lejano y rango submilimétrico obtenidos

por Herschel con los datos en el infrarrojo cercano y rango óptico obtenidos por el estudio GAMA. El objetivo principal de este análisis es derivar la función de luminosidad infrarroja de las galaxias del universo local ricas en gas. Otro de los objetivos es aplicar una herramienta de modelado de SED a un combinado de datos en los rangos óptico, infrarrojo cercano, infrarrojo lejano y submilimétrico proporcionados por las observaciones de GAMA y Herschel-ATLAS de galaxias ricas en gas para obtener por primera vez parámetros físicos intrínsecos como opacidades del polvo o tasas de formación estelar.

En conclusión, aunque se han realizado con éxito muchos estudios del Universo cercano en longitudes de onda ópticas, las cuales nos han proporcionado una comprensión detallada de las estrellas en las galaxias y de su evolución, Herschel ATLAS es el primer estudio de este tipo en longitudes de onda del infrarrojo lejano y submilimétricas, lo que proporcionará una nueva percepción de las galaxias e información de las estrellas ocultas por el polvo y el gas a partir de los cuales se pueden formar nuevas estrellas en el futuro.

4.1.1.3. Características Técnicas de Herschel

El observatorio Herschel tiene aproximadamente 7.5 metros de alto y 4 x 4 metros de sección transversal en conjunto, con una masa de lanzamiento de alrededor de 3.4 toneladas. Su estructura abarca dos componentes principales, un módulo de carga útil y un módulo de servicio. El módulo de carga útil está formado por:

- El telescopio, que recoge la radiación infrarroja y submilimétrica de fuentes astronómicas y la dirige, a través de un espejo secundario de 0.3 m de diámetro a los detectores de los tres instrumentos de medición de qué dispone Herschel dónde es detectada y registrada.

- El criostato, que contiene el banco óptico de los tres instrumentos, los componentes de los mismos que deben ser enfriados, es decir las unidades sensibles de los detectores, un sistema de refrigeración de helio líquido y enfriadores específicos para componentes que requieren condiciones aún más frías que las ofrecidas únicamente por el criostato.

- El parasol, que protege el telescopio y el criostato de la radiación visible e infrarroja solar, evita que la luz extraviada de la Tierra entre en el telescopio y dispone de células solares en la parte inferior de su superficie exterior que suministran energía eléctrica a la nave espacial.

- Estructuras de Apoyo, que conectan mecánicamente los diversos componentes del módulo de carga útil juntos y los montan en el módulo de servicio al tiempo que proporcionan aislamiento térmico donde se requiera.

Por su lado, el módulo de servicio alberga los sistemas necesarios para operar la nave espacial:

- Sistema de potencia, que se encarga de la generación, almacenamiento, acondicionamiento y distribución de energía eléctrica.

- Sistema de control de posición y órbita, que mide la posición de la nave espacial usando rastreadores de estrellas, giroscopios y sensores solares, y permite cambiar su posición u órbita por medio de ruedas de reacción y propulsores de hidracina.

- Sistema de control y gestión de datos, que se encarga de la recepción, almacenamiento y ejecución de comandos de tierra, el funcionamiento autónomo de la nave espacial en la ausencia de enlace con una estación terrestre, el almacenamiento y la gestión de la observación y la limpieza de datos.

- Sistema de comunicaciones, que se ocupa de la vinculación de la nave espacial con la estación de tierra para enviar de vuelta los datos y recibir comandos.

- El módulo de servicio también cuenta con las partes de los instrumentos que no requieren refrigeración.

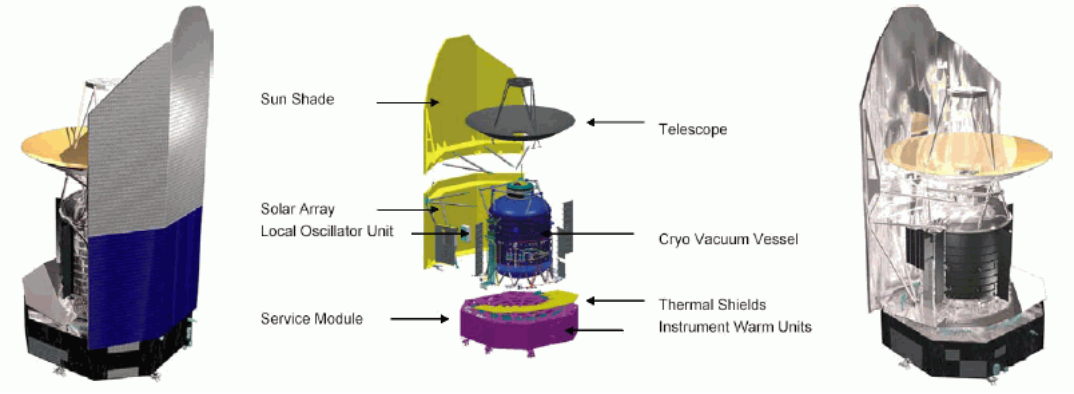

Figura 9. Esquema del diseño modular de Herschel en el que se muestran sus dos lados (caliente y frio) así como sus principales componente.

La nave espacial está diseñada para evitar cualquier problema causado por la interferencia de la radiación térmica infrarroja de la Tierra con las observaciones. La órbita de la misma en torno al punto L_2 también previene la aparición de cambios de temperatura debidos al movimiento de la nave espacial dentro y fuera del eclipse en una órbita terrestre, los cuales son un problema particular para los instrumentos infrarrojos que requieren estabilidad térmica extrema. No obstante, las órbitas sobre L_2 son dinámicamente inestables; pequeñas desviaciones del equilibrio crecen exponencialmente con una constante temporal de 23 días, razón por la cual Herschel usa su sistema de propulsión para realizar maniobras de mantenimiento de la órbita aproximadamente una vez al mes.

4.1.1.4. El Telescopio y los Instrumentos

Herschel está equipado con un telescopio reflector de diseño Cassegrain, caracterizado por utilizar 3 espejos, cuyo espejo primario tiene un diámetro de 3.5 metros, el espejo más grande jamás construido para un telescopio espacial, y capaz de recoger la radiación de onda larga de algunos de los objetos más fríos y distantes en el Universo.

La carga científica de Herschel comprende tres instrumentos de medición que realizan una combinación de espectroscopia, espectrometría de imágenes y fotometría de imágenes cubriendo un rango de longitudes de onda de 55 µm a 672 µm. Concretamente, se trata de dos cámaras (PACS y SPIRE) con capacidades de espectroscopia óptica adicionales, y un espectrómetro heterodino de muy alta resolución (HIFI):

Photodetector Array Camera and Spectrometer (PACS)

PACS consta de un fotómetro de imágenes (cámara) y un espectrómetro de campo integral de resolución media. Ambos subinstrumentos son independientes y ofrecen dos modos operativos básicos mutuamente excluyentes:

En el modo de fotometría de imágenes de banda dual, PACS fotografía un campo de visión de 1.75×3.5 minutos de arco, con 2560 píxeles y una resolución de 5 arcosegundos, en dos bandas simultáneamente, una de ellas a elegir entre $60 - 85$ μm y $85 - 125$ μm junto con una banda de $125 - 210$ μm, realizando un muestreo de haz completo en cada banda. Los detectores utilizados en este modo son dos arrays de bolómetros.

En el modo de espectroscopia de campo integral, PACS realiza espectroscopia entre 51 y 200 μm sobre un campo de visión de 47 x 47 segundos de arco, con 400 píxeles y una resolución de 10 arcosegundos. Este modo de PACS proporciona un poder de resolución entre 1000 y 4000, es decir una resolución espectral de alrededor de 75 a 300 km/s dependiendo de la longitud de onda, con una velocidad instantánea de cobertura de unos 1500 km/s. Los detectores de este modo son dos arrays fotoconductores de germanio/galio.

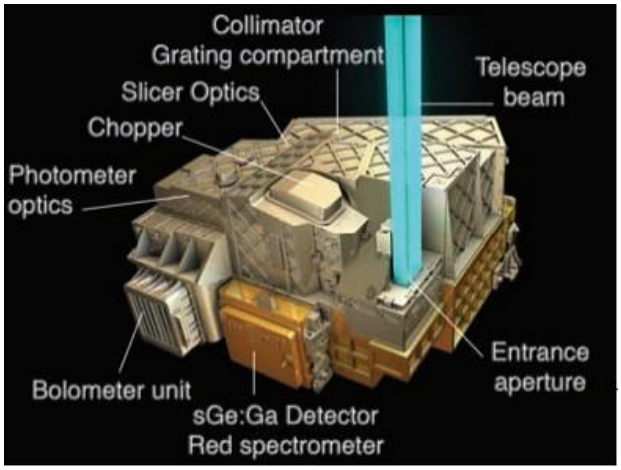

Figura 10. Esquema del instrumento PACS.

Spectral and Photometric Imaging Receiver (SPIRE)

SPIRE alberga un fotómetro de imágenes (cámara), que opera en tres bandas centradas en las longitudes de onda de 250, 350 y 500 μm, y un espectrómetro de imágenes con transformación de Fourier de resolución media. La cámara y el espectrómetro de ocupan compartimentos separados en el cuadro de instrumentos.

El compartimento de cada instrumento contiene espejos, filtros submilimétricos para definir las bandas de longitud de onda observadas, espejos móviles para controlar el haz, fuentes de calibración internas y, como detectores, utiliza arrays de bolómetros en tela de araña con sensores de temperatura de germanio dopado con transmutación de neutrones, que operan a una temperatura de 0.3 K alcanzada gracias a un enfriador interno. En total hay cinco arrays, tres dedicados a la fotometría y dos para la espectroscopia.

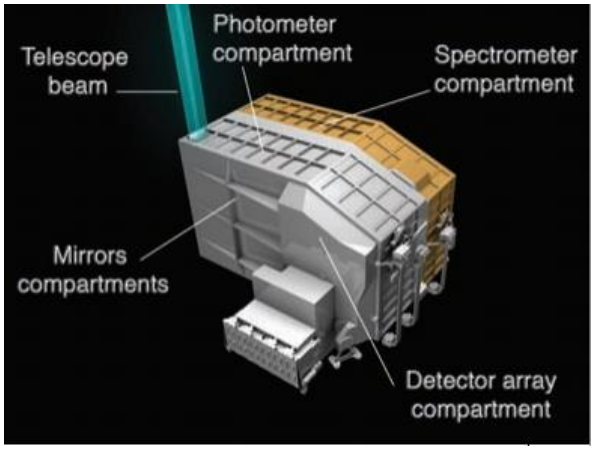

Figura 11. Esquema del instrumento SPIRE.

El fotómetro permite fotografiar un campo de visión en el cielo de 4×8 minutos de arco, con 270 píxeles y con una resolución de $20 - 30$ arcosegundos, en las tres bandas simultáneamente, cubriendo un rango de $200 - 670$ μm, y dispone de tres modos de observación:

- Fotometría de fuentes puntuales.
- Mapa de zonas pequeñas; para fuentes o áreas con diámetros menores de 5 minutos de arco.
- Mapa de zonas grandes; para cubrir grandes áreas del cielo o fuentes extensas mayores de 5 minutos de arco.

El espectrómetro de SPIRE se basa en una configuración Mach-Zehnder. Un puerto de entrada recibe el haz incidente procedente del telescopio mientras que un segundo puerto de entrada acepta una señal de una fuente de calibración. Los dos puertos de salida tienen cada uno un array de detectores, uno en el rango de 194 a 313 μm (37 detectores) y otro para $303 - 671$ μm (19 detectores), permitiendo medir las características espectrales de átomos y moléculas. El espectrómetro opera en modo de exploración continua cubriendo un campo de visión circular de 2.6×2.6 minutos de arco, con 56 píxeles, y una resolución de $20 - 50$ arcosegundos y una resolución espectral de 300 km/s.

Heterodyne Instrument For the Far Infrared (HIFI)

HIFI es un espectrómetro heterodino de muy alta resolución. El principio de detección heterodina implica trasladar el rango de frecuencias de la señal astronómica observada a una frecuencia menor, en la que es más fácil realizar las mediciones requeridas. Esto se consigue mediante la mezcla de la señal entrante con una señal monocromática muy estable, generada por un oscilador local, y extrayendo la diferencia de frecuencia para su posterior procesamiento. HIFI no permite obtener imágenes, sólo puede apuntar a una pequeña región del cielo cada vez, un único pixel, esparciendo la luz para que se pueda observar el espectro completo de la radiación infrarroja.

HIFI observa en 7 bandas cubriendo un rango de frecuencias 480 a 1910 GHz, o lo que es lo mismo un rango de longitudes de onda de 157 a 625 μm, con una resolución de $13 - 40$ segundos de arco y una resolución espectral de $0.02 - 0.7$ km/s. Las 5 primeras bandas, que cubren de forma continua desde los 480 a 1250 GHz, usan mezclas de superconductor-aislante-superconductor como detectores. Por su lado, las dos bandas restantes cubren el rango de 1410 a 1910 GHz y utilizan como detectores bolómetros de electrones calientes.

La diferencia de señal del proceso heterodino se envía a los espectrómetros de los instrumentos alojados en el módulo de servicio. Dentro de HIFI hay cuatro espectrómetros, dos espectrómetros acústico-ópticos de banda ancha (WBS) y dos espectrómetros de autocorrelación de alta resolución (HRS), estando disponible un espectrómetro de cada tipo para cada polarización de la luz (vertical y horizontal). Todos los espectrómetros se pueden utilizar tanto individualmente como en paralelo.

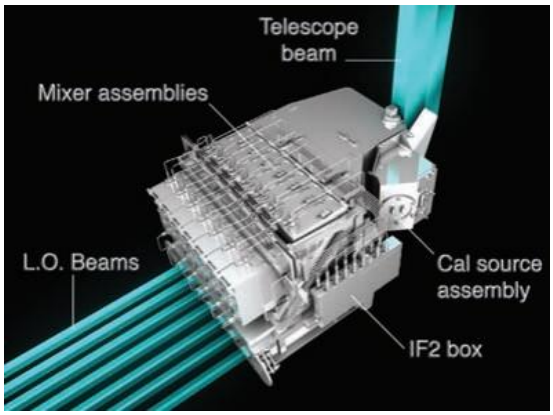

Figura 12. Esquema del instrumento HIFI.

Estos instrumentos han sido diseñados para aprovechar al máximo las características de la misión Herschel. Con el fin de realizar mediciones en longitudes de onda infrarrojas y submilimétricas, algunas partes de los instrumentos tienen que ser enfriadas a una temperatura cercana al cero absoluto, ya que un detector no puede estar a mayor temperatura que la radiación que intenta medir. El banco óptico, la estructura de montaje común de los tres instrumentos, está contenido dentro del criostato y más de 2.000 litros de helio líquido son utilizados para la refrigeración primaria. Los instrumentos de detección individuales están equipados con sistemas de refrigeración adicionales, especializados para alcanzar temperaturas muy bajas, de hasta 0.3 K para PACS y SPIRE.

En el proyecto Herschel ATLAS, se han usado tanto la cámara PACS como SPIRE puesto que juntos son capaces de ver todo el rango de luz desde el infrarrojo lejano hasta la región submilimétrica cubriendo un gran área y permitiendo estudiar tanto objetos astronómicos extremadamente raros como muchas galaxias normales cercanas. Las observaciones de Herschel ATLAS constan de dos escaneados de forma paralela alcanzando sensibilidades de 5σ de fuentes puntuales a densidades de flujo de 132, 126, 32, 36 y 45 mJy en las bandas de 100, 160, 250, 350 y 500 μm, respectivamente, con tamaños de haz aproximados de 9, 13, 18, 25 y 35 arcosegundos en las mismas bandas.

4.1.2. Publicación de Datos

En este trabajo se ha utilizado como catálogo astronómico la primera publicación de datos de Herschel ATLAS, conocida originalmente como *Science Demonstration Phase* (SDP). Esta publicación constituye la mayor difusión pública de datos de Herschel y una poderosa herramienta para el estudio de la evolución galáctica. Estos datos son el resultado del estudio de mapas del cielo observados en cinco bandas (100, 160, 250, 350 y 500 μm) y contienen 6876 fuentes astronómicas identificadas. Además proporcionan información sobre las contrapartidas ópticas de 2400 galaxias de estas fuentes astronómicas, las cuales tienen identificaciones confiables en la Sloan Digital Sky Survey (SDSS). Cabe destacar que los corrimientos al rojo espectroscópicos y las magnitudes ópticas de las bandas u − K utilizados en estos datos son cortesía del proyecto GAMA. Concretamente, el catálogo utilizado en este trabajo puede ser descargado en [6]. La información detallada de las columnas del catálogo puede encontrarse en [8]. A continuación se procede a describir las columnas del catálogo que se han utilizado:

- La columna **HATLAS_IAU_ID** es el identificador asignado por la Unión Astronómica Internacional a la correspondiente fuente astronómica del catálogo Herschel ATLAS a partir de su posición en la banda de 250 μm.

- La columna **SDP_ID** es el identificador asignado a la correspondiente fuente astrónomica del catálogo Herschel ATLAS dentro de la *Science Demonstration Phase*.

- La columna **RA_J2000** es la ascensión recta en grados de la correspondiente fuente astronómica determinada a partir de los datos de la banda de 250 μm.

- La columna **DEC_J2000** es la declinación en grados de la correspondiente fuente astronómica determinada a partir de los datos de la banda de 250 μm.

- La columna **PHOTOZ** es la estimación del corrimiento al rojo fotométrico de la correspondiente fuente astronómica, obtenido a partir de ANNz. (http://arxiv.org/abs/astro-ph/0311058)

- La columna **PHOTOZERR** es la estimación del error en el corrimiento al rojo fotométrico de la correspondiente fuente astronómica, obtenido a partir de ANNz.

- La columna **Z_SPEC** es la estimación del corrimiento al rojo espectroscópico de la correspondiente fuente astronómica, cedida por el proyecto GAMA.

- La columna **Z_QUAL** es una etiqueta de la calidad de la medida del corrimiento al rojo espectroscópico de la correspondiente fuente astronómica, obtenida por el proyecto GAMA. El catálogo Herschel ATLAS sólo utiliza los corrimientos al rojo espectroscópicos con valores de $z_{qual} \geq 3$ puesto que son los que tienen una fiabilidad de al menos el 90%.

- La columna **GAMA_IAU_ID** es el identificador asignado por la Unión Astronómica Internacional a la correspondiente fuente astronómica del catálogo GAMA, en este caso identifica las fuentes del catálogo GAMA para las que se ha encontrado una contrapartida oficial en el catálogo Herschel ATLAS.

Cuando no se dispone de medidas de alguna magnitud o de identificación en algún estudio, se indica con un $" - 99"$. Puesto que se han utilizado los datos de la banda de $250\ \mu m$ para obtener las ascensiones rectas y declinaciones de las fuentes astrónomicas, el valor de la anchura a media altura para este catálogo es $FWHM = 17.98"$, como se puede comprobar en [7].

4.2. Catálogo GAMA

4.2.1. El Proyecto GAMA

GAMA (Galaxy And Mass Assembly) es un proyecto internacional que aprovecha la última generación de instalaciones astronómicas tanto terrestres como espaciales para estudiar la cosmología del Universo: formación y evolución de galaxias, y la evolución de la masa, energía y estructura del Universo en los últimos mil millones de años, desarrollando en el proceso importantes avances en el área de observación, la eficiencia espectroscópica, la resolución espacial y la cobertura de longitudes de onda y convirtiéndola en la única base de datos capaz de avanzar en los estudios de galaxias de corrimientos al rojo bajos e intermedios [18].

En el corazón de este Proyecto se encuentra el estudio espectroscópico GAMA de ~ 300.000 galaxias a $r < 19.9$ mag sobre una región del cielo de ~ 286 grados cuadrados llevado a cabo usando el espectrógrafo multi-objetos AAOmega del Anglo-Australia Telescope (AAT) por el personal de GAMA durante 210 noches a lo largo de 7 años $(2008 - 2014)$. Este estudio se ve aumentado y complementado por estudios espectroscópicos anteriores como el Sloan Digital Sky Survey (SDSS), el UKIRT Infrared Deep Sky Survey, el 2dF Galaxy Redshift Survey (2dFGRS) y el Millenium Galaxy Catalogue (MGC).

Pero además, GAMA recopila los datos obtenidos por el VLT Survey Telescope (VST), el Visible and Infrared Survey Telescope for Astronomy (VISTA), el Australian Square Kilometre Array Pathfinder (ASKAP), el Observatorio Espacial Herschel, el Galaxy Evolution Explorer (GALEX), el Sloan United Kingdom Infrared Telescope (UKIRT), el Giant Metrewave Radio Telescope (GMRT), el Canada France Hawaii Telescope (CFHT), el X-ray Multi-Mirror Mission (XMM-Newton) y el Wide field Infrared Survey Explorer (WISE) con el fin de construir una base de datos estado del arte multi-longitud de onda fotométrica y espectroscópica de ~ 375.000 galaxias en el Universo local sobre una región del cielo de $360°$ cuadrados.

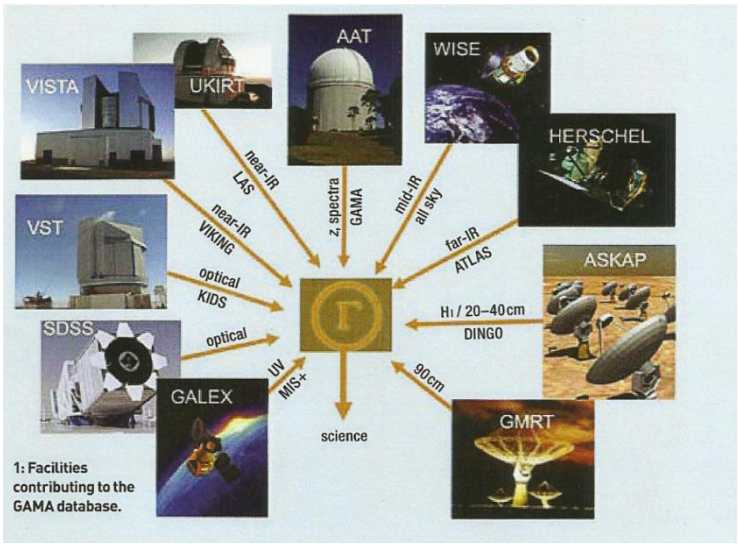

Figura 13. Esquema de las instalaciones que contribuyen a la base de datos GAMA [18].

El principal objetivo de GAMA es estudiar la estructura del Universo en escalas de 1 kpc a 1 Mpc, en las que los bariones juegan un papel fundamental en la formación y posterior evolución de las galaxias y donde nuestra comprensión de la estructura del universo termina. Así, la meta de GAMA es poner a prueba el modelo Lambda-CDM (Lambda-Cold Dark Matter) del Universo, el cual concuerda la teoría del Big Bang con las observaciones de la radiación cósmica de fondo de microondas, la estructura a gran escala del Universo y las observaciones de supernovas, arrojando luz sobre la aceleración de la expansión del Universo. En particular, los objetivos científicos principales de GAMA son:

- Comprobar las teorías de la gravedad modificada mediante la medición de la tasa de crecimiento de la estructura, el modelo CMD mediante la medición de la función de masa de la materia oscura del halo de grupos y cúmulos utilizando medidas de la dispersión de la velocidad de grupo y los modelos de formación de galaxias mediante la medición de la eficiencia de formación de estrellas en grupos.

- Medir la conexión entre el abastecimiento de combustible en la formación estelar, la acumulación de la masa estelar y los procesos de retroalimentación en las estrellas así como descubrir los mecanismos detallados que rigen la acumulación del contenido estelar de las galaxias.

- Medir directamente las tasas de fusión de galaxias recientes en función de la masa, la relación de masas, el medio ambiente local y el tipo de galaxia.

4.2.2. Publicación de Datos

En este trabajo se ha utilizado como catálogo astronómico la primera publicación de datos de GAMA, conocida originalmente como *Data Release 1* (GamaCoreDR1). Esta publicación, la cual fue divulgada el 25 de Junio de 2010, contiene información fotométrica y espectroscópica de 114441 galaxias de SDSS dentro de las regiones GAMA. Concretamente, el catálogo utilizado en este trabajo puede ser descargado en [15]. La información detallada de las columnas del catálogo puede encontrarse en [16]. A continuación se procede a describir las columnas del catálogo que se han utilizado:

- La columna **GAMA_IAU_ID** es el identificador asignado por la Unión Astronómica Internacional a la correspondiente fuente astronómica del catálogo GAMA.

- La columna **RA_J2000** es la ascensión recta en grados de la correspondiente fuente astronómica.

- La columna **DEC_J2000** es la declinación en grados de la correspondiente fuente astronómica.

- La columna **Z_HELIO** es la estimación del corrimiento al rojo espectroscópico de la correspondiente fuente astronómica. Si el corrimiento al rojo es desconocido se indica con un "9999", si se divulgará en una publicación posterior se indica con un " − 2" y si el corrimiento al rojo se ha observado pero es una medida pobre se indica con " − 0.9".

- La columna **Z_QUALITY** es una etiqueta de la calidad de la medida del corrimiento al rojo espectroscópico de la correspondiente fuente astronómica. Un "5" es el máximo nivel de confiabilidad, un "4" indica una confianza superior al 95%, un "3" indica por lo general una buena confianza $(90\% - 95\%)$, un "2" indica una medida pobre $(\sim 10\% - 90\%)$ y un "1" va asignado a las medidas del corrimiento al rojo pobres identificadas con " − 0.9".

El valor de la anchura a media altura para este catálogo es $\mathrm{FWHM} = 0.7"$, como se puede comprobar en [18].

5. Procedimiento de Implementación

5.1. Descripción de los Programas

Para implementar la herramienta de cross-identificación de los objetos extragalácticos de los catálogos Herschel ATLAS y GAMA, atendiendo a sus distancias angulares y a los valores de sus corrimientos al rojo, se ha hecho uso de dos programas de tratamiento de tablas: STILTS y TOPCAT, ambos basados en STIL (Starlink Tables Infrastructure Library), una librería pura de Java para el input, output y procesamiento genérico de datos tabulares.

STILTS (STIL Tool Set) es un conjunto de herramientas por línea de comandos para el procesamiento de datos tabulares. Ha sido diseñado para el uso de tablas astronómicas, tales como catálogos de fuentes astronómicas, siendo muy eficaz en el tratamiento de grandes conjuntos de datos, pero puede tratar también otra gran variedad de datos. Dispone tanto de herramientas genéricas que pueden trabajar con multitud de formatos como FITS o ASCII y otras específicas, lo que le dota de una alta portabilidad. STILS está disponible bajo General Public License (GNU) y puede descargarse en [27].

Entre las facilidades que ofrece STILTS se incluyen: conversión de formatos, comandos de crosmatching, representación gráfica, cálculo y reordenamiento de columnas, selección de filas, manipulación y visualización de datos y metadatos, clasificación, cálculo estadístico, cálculo de histogramas, validación de datos y acceso a servicios de datos remotos incluyendo Observatorio Virtual, las cuales se puede juntar de formas muy flexibles y eficientes. Toda la documentación concerniente al uso y comandos de STILTS se encuentra en [26].

TOPCAT (Tool for Operations on Catalogues And Tables) es una interfaz gráfica interactiva que permite examinar, analizar, combinar y editar datos tabulares. Al igual que STILTS, ha sido diseñado para su uso con tablas astronómicas como catálogos de objetos, pero no se limita a las aplicaciones astronómicas y es capaz de manejar grandes cantidades de datos en gran multitud de formatos. Comprende un gran número de diferentes formatos astronómicamente importantes y ofrece una gran variedad de formas de visualizar y analizar los datos, incluyendo un navegador para las celdas de los datos, visores y editores para la información de las tablas y los metadatos de las columnas, herramientas para unir tablas usando algoritmos de emparejamiento flexibles y extensas comodidades de visualización en 2D y 3D. TOPCAT está disponible bajo General Public License (GNU) y puede descargarse en [30]. Toda la documentación concerniente al uso y funciones de TOPCAT se encuentra en [29].

En cierto modo, la interfaz gráfica de usuario TOPCAT es la contrapartida de la herramienta por línea de comandos STILTS puesto que ambas comparten gran parte de sus funcionalidades. De este modo, nuestra herramienta de cross-identificación sobre los catálogos Herschel ATLAS y GAMA se ha implementado en un script utilizando los comandos de STILTS y se ha ejecutado desde la línea de comandos o terminal de UNIX. Posteriormente, se ha visualizado y analizado la tabla resultante utilizando TOPCAT. A continuación se procede detallar el script realizado.

5.2. Descripción del Script

Este es un script para calcular el factor de Bayes a partir de las posiciones angulares (ascensión recta y declinación) en radianes y las medidas de los redshifts de los objetos extragalácticos de los catálogos H-ATLAS y GAMA. El catálogo GAMA sólo proporciona medidas espectroscópicas del redshift mientras que el catálogo H-ATLAS proporciona medidas tanto espectroscópicas como fotométricas. Únicamente las medidas fotométricas del redshift poseen un error numérico, las medidas espectroscópicas tienen asignados unos valores enteros especificando la calidad de la medida realizada.

En primer lugar se han creado dos tablas "CatalogoGama.fits" y "CatalogoHATLAS.fits" que contienen únicamente los identificadores de los objetos extragalácticos, sus posiciones angulares (ascensión recta y declinación) y las medidas de sus redshifts, espectroscópicos y/o fotométricos. Con el comando "colmeta" se ha cambiado el nombre a las columnas de la ascensión recta, la declinación y los redshifts en ambos catálogos, para poder identificarlas cuando los juntemos. Con el comando "delcols" se ha borrado el resto de columnas que no se van necesitar para este cálculo:

```
./stilts tpipe ifmt=fits ofmt=fits in=HATLAS_SDP_catalogue.fits
out=CatalogoHATLAS.fits \
               cmd='delcols "5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 42 43 44 47 48 49 50 51 52
53 54 55 56 57 58 59 60 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
80 81 82 83 84 85 86 87 88 89 90 91"' \
               cmd='colmeta -name 'RA_HATLAS' '3'' \
               cmd='colmeta -name 'DEC_HATLAS' '4'' \
               cmd='colmeta -name 'z_photo_HATLAS' '5'' \
               cmd='colmeta -name 'z_photo_error_HATLAS' '6'' \
               cmd='colmeta -name 'z_spec_HATLAS' '7'' \
               cmd='colmeta -name 'z_spec_quality_HATLAS' '8'' \
               cmd='colmeta -name 'GAMA_IAU_ID_2' '9'' \
```

```
./stilts tpipe ifmt=fits ofmt=fits in=GamaCoreDR1_v1.fits
out=CatalogoGama.fits \
               cmd='delcols "6 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32 33"' \
               cmd='colmeta -name 'RA_GAMA' '4'' \
               cmd='colmeta -name 'DEC_GAMA' '5'' \
               cmd='colmeta -name 'z_spec_GAMA' '6'' \
               cmd='colmeta -name 'z_spec_quality_GAMA' '7'' \
```

En segundo lugar se ha utilizado el comando "tskymatch2" para realizar un emparejamiento cruzado de los objetos extragalácticos, de las dos tablas reducidas que se acaban de crear, en base a la proximidad de sus posiciones en el cielo. Las columnas de las ascensiones rectas y declinaciones de ambos catálogos se pasan en "grados" y se introduce un límite en la separación angular de las fuentes, en "segundos de arco", para que la asociación de dos objetos extragalácticos se dé por válida. En nuestro caso se han elegido "54 arcosegundos", aproximadamente el triple de la precisión del catálogo menos preciso (H-ATLAS). Con "find=all" se ha indicado que en la tabla final se incluyan todas las coincidencias o emparejamientos encontrados, por lo que habrá objetos de un catálogo que tengan múltiples contrapartidas en el otro, y viceversa. Con "join=1and2" se ha indicado que cada fila de la tabla resultante represente una coincidencia entre dos filas de las tablas iniciales:

```
./stilts tskymatch2 ifmt1=fits ifmt2=fits ofmt=fits in1=CatalogoHATLAS.fits
in2=CatalogoGama.fits out=CrossMatchingGamaHATLAS.fits \
                    ra1=RA_HATLAS dec1=DEC_HATLAS ra2=RA_GAMA dec2=DEC_GAMA
error=54 find=all join=1and2 \
```

En la tabla resultante aparece una columna "Separation" que muestra en "segundos de arco" las distancias o separaciones angulares entre los pares de objetos extragalácticos que se han emparejado, todas ellas inferiores al límite de 54" impuesto. Primero, se ha creado una columna "Separacion_Angular" en la que se ha calculado la separación angular entre los objetos extragalácticos en radianes. Se han añadido dos columnas "Precision_GAMA" y "Precision_HATLAS" que contienen las incertidumbres de ambos catálogos, en radianes, obtenidas a partir de los valores de FWHM encontrados en [7] y [18]. Por último, se ha añadido una última columna en la que, utilizando las separaciones angulares de la columna "Separacion_Angular", las incertidumbres de ambos catálogos y la expresión simplificada del factor de Bayes posicional en aproximación para distancias pequeñas entre fuentes mostrada en la ecuación (43), se ha calculado la contribución estrictamente posicional al factor de Bayes:

```
./stilts tpipe ifmt=fits ofmt=fits in=CrossMatchingGamaHATLAS.fits
out=BayesPosicionalGamaHATLASrad.fits \
               cmd='addcol -units 'radians' 'Separacion_Angular'
'degreesToRadians\('Separation'/3600\)'' \
               cmd='addcol -after 'Separacion_Angular' -units 'radians'
'Precision_HATLAS'
'degreesToRadians\(17.98/\(3600*\(2*sqrt\(2*ln\(2\)\)\)\)\)'' \
               cmd='addcol -after 'Precision_HATLAS' -units 'radians'
'Precision_GAMA' 'degreesToRadians\(0.7/\(3600*\(2*sqrt\(2*ln\(2\)\)\)\)\)'' \
               cmd='addcol -after 'Precision_GAMA' 'Factor_Bayes_Posicional'
'\(2/\('Precision_HATLAS'*'Precision_HATLAS'+'Precision_GAMA'*'Precision_GAMA'
\)\)*
exp\(-\('Separacion_Angular'*'Separacion_Angular'\)/\(2*\('Precision_HATLAS'*
'Precision_HATLAS'+'Precision_GAMA'*'Precision_GAMA'\)\)\)'' \
```

A continuación, se ha procedido a calcular la contribución al factor de Bayes debida estrictamente al corrimiento al rojo. Para ello se ha utilizado un enfoque de probabilidad gaussiana cuya aproximación es buena sólo cuando el error del redshift fotométrico es pequeño en comparación con la estimación del valor del redshift. Este suele siempre ser el caso de los redshifts espectróscopicos, pero no tiene por qué darse con los redshifts fotométricos. De este modo, se ha procedido diferenciando 3 posibles casos:

En el primer caso se crea una tabla "BayesRedshiftCaso1.fits" que contiene únicamente los objetos extragalácticos para los que la estimación del redshift fotométrico del catálogo H-ATLAS es mayor que su error, descartando aquellos objetos en los que el redshift espectroscópico del catálogo GAMA sea desconocido ($9999.$) o aún no se haya medido ($-2.$). No es preciso descartar ni hacer ninguna especificación sobre los objetos cuyas columnas "z_photo_HATLAS" y "z_photo_error_HATLAS" tengan valores desconocidos o no medidos ($-99.$) puesto que siempre que una de las dos columnas tiene el valor $-99.$, la otra tiene el mismo valor. Y estos objetos son inmediatamente descartados por la condición inicial ('z_photo_HATLAS' > 'z_photo_error_HATLAS'). La ecuación (54) utilizada para calcular la contribución del redshift al factor de Bayes en la columna "Factor_Bayes_Redshift" utiliza el redshift fotométrico del catálogo H-ATLAS (junto con su error) y el redshift espectroscópico del catálogo GAMA (sin ningún error puesto que las medidas espectroscópicas del redshift carecen de él):

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCaso1.fits \
               cmd='select " 'z_photo_HATLAS' > 'z_photo_error_HATLAS' &&
'z_spec_GAMA' != -2. && 'z_spec_GAMA' != 9999." ' \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift'
'\(2/sqrt\(2*PI\)\)*\(1/'z_photo_error_HATLAS'\)*exp\(-\('z_photo_HATLAS'-
'z_spec_GAMA'\)*\('z_photo_HATLAS'-
'z_spec_GAMA'\)/\(2*'z_photo_error_HATLAS'*'z_photo_error_HATLAS'\)\)'' \
```

En el segundo caso se crea una tabla "BayesRedshiftCaso2.fits" que contiene únicamente los objetos extragalácticos para los que la estimación del redshift fotométrico del catálogo H-ATLAS es menor o igual que su error, descartando aquellos objetos en los que el redshift espectroscópico del catálogo H-ATLAS no se haya podido medir ($-99.$) y el redshift espectroscópico del catálogo GAMA sea desconocido ($9999.$) o aún no se haya medido ($-2.$). La expresión para calcular la contribución del redshift al factor de Bayes en la columna "Factor_Bayes_Redshift" utilizaría el redshift espectroscópico del catálogo H-ATLAS y el redshift espectroscópico del catálogo GAMA, pero como el catálogo H-ATLAS utiliza el redshift espectroscópico del catálogo GAMA no estaríamos comparando dos mediciones de distinto origen sino la misma. Por este motivo se ha asignado un valor de 1. en la columna "Factor_Bayes_Redshift", de modo que al final, en estos objetos, sólo tenga peso la contribución posicional:

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCaso2.fits \
               cmd='select " 'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' != -99. && 'z_spec_GAMA' != -2. && 'z_spec_GAMA' != 9999." '
\
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En el tercer caso se han incluido todos aquellos objetos extragalácticos de los que no se ha podido obtener una contribución clara del redshift al factor de Bayes por la ausencia de medidas o desconocimiento de los redshifts en alguno de los catálogos. Por ello, se ha procedido por partes, estudiando cada grupo de estos objetos por separado y juntándolos todos al final en una única tabla "BayesRedshiftCaso3.fits". Para todos ellos se ha asignado un valor de 1. en la columna "Factor_Bayes_Redshift", de modo que al final, en estos objetos, sólo tenga peso la contribución posicional. Se tiene que poner "1." (double) en lugar de simplemente "1" (short) para que al unir las columnas "Factor_Bayes_Redshift" de todas las tablas no haya un problema de compatibilidad entre valores tipo double calculados usando la ecuación (54) y la columna "Factor_Bayes_Redshift" del segundo y tercer cas en los que todos los valores son 1:

En la parte A se han considerado los objetos para los que se usa el redshift fotométrico del catálogo H-ATLAS ('z_photo_HATLAS' > 'z_photo_error_HATLAS') pero cuyo redshift espectroscópico en el catálogo GAMA no se ha podido medir $(-2.)$:

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoA.fits \
               cmd='select "'z_photo_HATLAS' > 'z_photo_error_HATLAS' &&
'z_spec_GAMA' == -2."' \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte B se han considerado los objetos para los que se usa el redshift fotométrico del catálogo H-ATLAS ('z_photo_HATLAS' > 'z_photo_error_HATLAS') pero cuyo redshift espectroscópico en el catálogo GAMA es desconocido $(9999.)$:

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoB.fits \
               cmd='select "'z_photo_HATLAS' > 'z_photo_error_HATLAS' &&
'z_spec_GAMA' == 9999."' \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte C se han considerado los objetos para los que se usa el redshift espectroscópico del catálogo H-ATLAS ('z_photo_HATLAS' <= 'z_photo_error_HATLAS'), pero éste no se ha podido medir $(-99.)$ y no ha habido ningún problema en las medidas del redshift espectroscópico del catálogo GAMA:

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoC.fits \
               cmd='select "'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' == -99. && 'z_spec_GAMA' != -2. && 'z_spec_GAMA' != 9999."'  \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte D se han considerado los objetos para los que se usa el redshift espectroscópico del catálogo H-ATLAS ('z_photo_HATLAS' <= 'z_photo_error_HATLAS'), habiéndose medido correctamente, pero cuyo redshift espectroscópico en el catálogo GAMA no se ha podido medir $(-2.)$:

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoD.fits \
               cmd='select "'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' != -99. && 'z_spec_GAMA' == -2."'  \
```

```
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte E se han considerado los objetos para los que se usa el redshift espectroscópico del catálogo H-ATLAS ('z_photo_HATLAS' <= 'z_photo_error_HATLAS'), habiéndose medido correctamente, pero cuyo redshift espectroscópico en el catálogo GAMA es desconocido (9999.):

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoE.fits \
               cmd='select "'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' != -99. && 'z_spec_GAMA' == 9999."'  \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte F se han considerado los objetos para los que se usa el redshift espectroscópico del catálogo H-ATLAS ('z_photo_HATLAS' <= 'z_photo_error_HATLAS'), pero éste no se ha podido medir (−99.), y además el redshift espectroscópico en el catálogo GAMA no se ha podido medir (−2.):

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoF.fits \
               cmd='select "'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' == -99. && 'z_spec_GAMA' == -2."'  \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

En la parte G se han considerado los objetos para los que se usa el redshift espectroscópico del catálogo H-ATLAS ('z_photo_HATLAS' <= 'z_photo_error_HATLAS'), pero éste no se ha podido medir (−99.), y además el redshift espectroscópico en el catálogo GAMA es desconocido (9999.):

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesPosicionalGamaHATLASrad.fits
out=BayesRedshiftCasoG.fits \
               cmd='select "'z_photo_HATLAS' <= 'z_photo_error_HATLAS' &&
'z_spec_HATLAS' == -99. && 'z_spec_GAMA' == 9999."'  \
               cmd='addcol -after 'Factor_Bayes_Posicional'
'Factor_Bayes_Redshift' '1.'' \
```

Se procede a juntar todos estos objetos extragalácticos para los que la contribución del redshift al factor de Bayes es "1." en una única tabla "BayesRedshiftCaso3.fits":

```
./stilts tcat ifmt=fits in="BayesRedshiftCasoA.fits BayesRedshiftCasoB.fits
BayesRedshiftCasoC.fits BayesRedshiftCasoD.fits BayesRedshiftCasoE.fits
BayesRedshiftCasoF.fits BayesRedshiftCasoG.fits" out=BayesRedshiftCaso3.fits \
```

Ahora se procede a unir las 3 tablas (3 casos) que se han tratado anteriormente de forma independiente. Las tres tablas deben tener un número y tipo de columnas compatible, lo cual está garantizado ya que proceden de una misma tabla original y tienen todas las mismas columnas con los mismos valores, excepto la última:

```
./stilts tcat ifmt=fits in="BayesRedshiftCaso1.fits BayesRedshiftCaso2.fits
BayesRedshiftCaso3.fits" out=BayesRedshiftGeneral.fits \
```

Por último, se procede a añadir una columna a la tabla resultante en la que se ha calculado el factor de Bayes conjunto como el producto de la contribución posicional y de la contribución del redshift al mismo, de acuerdo con la ecuación (55):

```
./stilts tpipe ifmt=fits ofmt=fits in=BayesRedshiftGeneral.fits
out=FactorBayes.fits \
               cmd='addcol 'Factor_Bayes'
''Factor_Bayes_Posicional'*'Factor_Bayes_Redshift' '' \
```

Para simplificar el análisis de los datos, se crea una tabla que contenga únicamente los objetos extragalácticos que tengan en el catálogo H-ATLAS su correspondiente identificación oficial del catálogo GAMA:

```
./stilts tpipe ifmt=fits ofmt=fits in=FactorBayes.fits
out=FactorBayesContrapartidas.fits \
               cmd='select "'z_spec_quality_HATLAS' != -99 &&
'z_spec_quality_HATLAS' != 0"' \
               cmd='select "'SDP_ID' != 61. && 'SDP_ID' != 5627. && 'SDP_ID'
!= 6264. && 'SDP_ID' != 1500. && 'SDP_ID' != 5631."' \
```

6. Resultados

Una vez ejecutado el script expuesto anteriormente, se procede a contrastar las identificaciones de objetos extragalacticos entre ambos catálogos logradas por nuestra herramienta con las contrapartidas oficiales de objetos del catálogo GAMA en el catálogo Herschel ATLAS. Como se ha explicado anteriormente, la magnitud utilizada para discriminar entre las múltiples contrapartidas posibles del catálogo GAMA para un objeto del catálogo Herschel ATLAS es el factor de Bayes. En caso de encontrarse más de una contrapartida, la que tenga un mayor factor de Bayes será nuestra identificación más probable.

En total, tras el cruce de 6876 fuentes astronómicas de Herschel ATLAS con 114441 objetos de GAMA, se han conseguido 3994 asociaciones de objetos de GAMA con objetos de Herschel ATLAS en una pequeña región del cielo ($134° < \mathrm{AR} < 139°, -1° < \mathrm{Dec} < 3°$). De todas estas asociaciones, sólo 1607 albergan identificaciones oficiales en GAMA, mientras que 2387 carecen de identificación oficial en GAMA.

Estas 1607 asociaciones con identificación oficial en GAMA se pueden clasificar en función del número de contrapartidas encontradas para un mismo objeto de Herschel ATLAS, como se muestra en la Tabla 3, obteniendo un total de 1024 objetos extragalácticos de Herschel ATLAS distintos con contrapartidas oficiales asignadas en GAMA.

N° de posibles contrapartidas	N° de identificaciones	Correctas		Incorrectas	
		B_{pos}	B	B_{pos}	B
1	613	596	596	17	17
2	295	282	278	13	17
3	79	78	74	1	5
4	25	25	24	0	1
5	8	8	7	0	1
6	1	0	0	1	1
7	3	3	3	0	0
	1024	96.9 %	95.9 %	3.1 %	4.1 %

Tabla 3. Registro de las 1607 asociaciones con identificación oficial obtenidas al cruzar los catálogos GAMA y Herschel ATLAS en función del número de contrapartidas encontradas en el catálogo GAMA para un mismo objeto de Herschel ATLAS. Se ha indicado el número de contrapartidas identificadas por nuestra herramienta que coinciden con las identificaciones oficiales de Herschel ATLAS para cada caso, considerando tanto el factor de Bayes conjunto (B) como el factor de Bayes posicional (B_{pos}) únicamente. Así mismo, se ha indicado el porcentaje final de contrapartidas identificadas correctamente para sendos factores de Bayes.

La comparación entre el número de identificaciones correctas encontradas usando el factor de Bayes conjunto (B) y las encontradas usando el factor de Bayes posicional (B_{pos}) tenía como propósito mostrar la influencia del factor de Bayes debido al corrimiento al rojo en nuestra herramienta de cross-identificación. En principio, se esperaba que la eficacia de este método mejorase al tener en cuenta el corrimiento al rojo. No obstante, tal y como se ve en la Tabla 3, la estadística de identificaciones correctas baja un 1 % respecto a considerar únicamente el factor de Bayes posicional. Esto es debido a que, en la gran mayoría de los casos en los que sucede esto, se carece de información sobre el corrimiento al rojo del objeto extragaláctico que sería la identificación correcta, lo que provoca que su factor de Bayes combinado sea discriminado en favor de otras contrapartidas erróneas (ver Tabla 4). También hay que considerar que de no haber incluido la información referente al corrimiento al rojo no se podría haber identificado correctamente el objeto extragaláctico mostrado en la Tabla 6.

Respecto a la eficacia de este método, cabe mencionar también que, tras realizar un estudio detallado de los 31 casos en los que no se ha logrado una identificación correcta de objetos extragalácticos de Herschel ATLAS ni con el factor de Bayes posicional ni con el factor de Bayes conjunto, sólo 1 de estas asociaciones incorrectas es debida a que el factor de Bayes mayor es asignado a una contrapartida que no se corresponde con la oficial. En los 30 casos restantes, la identificación oficial establecida por Herschel ATLAS ni si quiera aparece entre las contrapartidas posibles, no está incluida en la publicación de datos de GAMA utilizada en este trabajo (ver Tabla 5). Es por ello que, si se descartasen estos casos al no ser posible encontrar la contrapartida oficial para ellos en GAMA, la eficacia de nuestra herramienta aumentaría al 99.8 % usando el factor de Bayes posicional y al 98.8 % considerando el factor de Bayes conjunto.

Debido a su excesiva extensión (1607 filas), no se ha podido incluir en esta memoria de forma íntegra la tabla "FactorBayesContrapartidas.fits" que contiene todas estas asociaciones, la cual si va incluida junto a este trabajo y la tabla "FactorBayes.fits" en el CD. No obstante, se ha procedido a exponer algunos ejemplos de distintos casos de interés encontrados tras analizar todas estas asociaciones:

SDP_ID	GAMA_IAU_ID_2	GAMA_IAU_ID	$\Delta\theta$ / "	B_{pos}	B_z	B
793	J090146.36-001736.2	J090145.75-001748.3	14.71	$2.28 \cdot 10^8$	46.8	$1.07 \cdot 10^{10}$
		J090146.36-001736.2	1.26	$1.44 \cdot 10^9$	1	$1.44 \cdot 10^9$
1828	J090652.66-000752.3	J090653.60-000747.1	14.50	$2.41 \cdot 10^8$	21.3	$5.12 \cdot 10^9$
		J090652.66-000752.3	2.35	$1.39 \cdot 10^9$	1	$1.39 \cdot 10^9$
4594	J090212.51+002726.4	J090212.86+002728.9	9.31	$6.94 \cdot 10^8$	45.3	$3.14 \cdot 10^{10}$
		J090212.51+002726.4	6.01	$1.07 \cdot 10^9$	1	$1.07 \cdot 10^9$
565	J090131.66+001925.0	J090131.34+001925.5	3.11	$1.34 \cdot 10^9$	11.4	$1.53 \cdot 10^{10}$
		J090132.02+001927.3	7.38	$9.15 \cdot 10^8$	11.1	$1.01 \cdot 10^{10}$
		J090131.66+001925.0	1.84	$1.42 \cdot 10^9$	1	$1.42 \cdot 10^9$
5244	J091510.27-002108.6	J091509.81-002106.2	6.07	$1.06 \cdot 10^9$	13.1	$1.39 \cdot 10^{10}$
		J091510.27-002108.6	1.51	$1.43 \cdot 10^9$	1	$1.43 \cdot 10^9$
		J091511.30-002156.5	51.71	0.166	1	0.166
		J091508.93-002033.7	38.85	3558	1	3558

Tabla 4. Registro de diversos casos de asociaciones entre objetos de Herschel ATLAS y GAMA en los que, debido a la ausencia de información referente al corrimiento al rojo (B_z) de la contrapartida correcta en GAMA, sólo el factor de Bayes posicional (B_{pos}) permite identificar la contrapartida correcta mientras que el factor de Bayes conjunto (B) que corresponde a la identificación oficial de Herschel ATLAS es discriminado en favor de otras contrapartidas erróneas. Aquí y en las tablas siguientes, SDP_ID es el identificador del correspondiente objeto del catálogo Herschel ATLAS utilizado, GAMA_IAU_ID_2 es el identificador de la contrapartida oficial del objeto de Herschel en el catálogo GAMA, GAMA_IAU_ID es el identificador de las diversas contrapartidas obtenidas por nuestra herramienta tras realizar el cross-matching y $\Delta\theta$ es la separación angular entre el objeto de Herschel y sus posibles contrapartidas en GAMA.

SDP_ID	GAMA_IAU_ID_2	GAMA_IAU_ID	$\Delta\theta$ / "	B_{pos}	B_z	B
2310	J090335.37+002048.8	J090335.37+002048.8	4.91	$1.19 \cdot 10^9$	92.3	$1.09 \cdot 10^{11}$
		J090334.99+002046.3	2.09	$1.40 \cdot 10^9$	90.1	$1.27 \cdot 10^{11}$
148	J090926.84-002144.5	J090925.43-002157.6	24.85	$7.35 \cdot 10^6$	19.2	$1.41 \cdot 10^8$
247	J085908.08+004520.3	J085907.21+004457.7	24.61	$8.14 \cdot 10^6$	$1.02 \cdot 10^{-7}$	0.87
		J085907.77+004507.6	11.89	$4.34 \cdot 10^8$	1.82	$7.95 \cdot 10^8$
4846	J091358.13+000342.4	J091357.33+000420.3	40.34	1295	23.6	30579
		J091359.17+000326.2	24.68	$7.91 \cdot 10^6$	3.28	$2.59 \cdot 10^7$
		J091359.38+000331.7	24.96	$7.02 \cdot 10^6$	3.09	$2.17 \cdot 10^7$
		J091359.27+000350.0	23.47	$1.30 \cdot 10^7$	3.15	$4.10 \cdot 10^7$
		J091356.09+000334.5	26.87	$3.01 \cdot 10^6$	1	$3.01 \cdot 10^6$
		J091357.85+000417.7	37.14	10817	1	10817

Tabla 5. Registro de diversos casos de asociaciones entre objetos de Herschel ATLAS y GAMA en los que no se ha logrado una identificación correcta de la contrapartida oficial del objeto de Herschel ni con el factor de Bayes posicional ni con el conjunto, bien debido a que se identifica una contrapartida errónea en su lugar (2310) o bien porque la identificación oficial establecida por Herschel ATLAS ni si quiera aparece entre las contrapartidas posibles. En el caso del objeto 2310, aunque la contrapartida correcta no es la tiene una menor separación angular, el corrimiento al rojo sí señala hacia la contrapartida correcta por lo que si se añadiese más información a priori referente al color, la morfología o la luminosidad del objeto se podría identificar bien.

SDP_ID	GAMA_IAU_ID_2	GAMA_IAU_ID	$\Delta\theta$ / "	B_{pos}	B_z	B
2627	J085920.83+005142.1	J085920.83+005142.1	6.11	$1.06 \cdot 10^{9}$	29.1	$3.08 \cdot 10^{10}$
		J085920.30+005141.5	3.49	$1.31 \cdot 10^{9}$	$3.92 \cdot 10^{-43}$	$5.15 \cdot 10^{-34}$

Tabla 6. Registro del único caso encontrado, tras el análisis de todos los datos, en el que el factor de Bayes referente al corrimiento al rojo resulta determinante para identificar correctamente la contrapartida oficial en GAMA para un objeto de Herschel ATLAS, puesto que como se observa la contrapartida correcta no es aquella para la que la separación angular es menor y el factor de Bayes posicional por sí solo no es revelador.

SDP_ID	GAMA_IAU_ID_2	GAMA_IAU_ID	$\Delta\theta$ / "	B_{pos}	B_z	B
3817	J090444.54-002440.6	J090444.54-002440.6	0.39	$1.46 \cdot 10^{9}$	43.1	$6.28 \cdot 10^{10}$
		J090444.27-002441.7	4.21	$1.25 \cdot 10^{9}$	42.6	$5.33 \cdot 10^{10}$
2116	J090831.40-005434.2	J090831.06-005438.5	4.93	$1.18 \cdot 10^{9}$	6.9	$8.23 \cdot 10^{9}$
		J090831.40-005434.2	3.11	$1.34 \cdot 10^{9}$	6.4	$8.55 \cdot 10^{9}$
460	J091404.76-000124.7	J091404.32-000129.7	7.68	$8.80 \cdot 10^{8}$	57.5	$5.06 \cdot 10^{10}$
		J091404.76-000124.7	1.06	$1.44 \cdot 10^{9}$	55.0	$7.94 \cdot 10^{10}$
29	J090750.07+010141.0	J090750.07+010141.0	0.55	$1.45 \cdot 10^{9}$	9.7	$1.41 \cdot 10^{10}$
		J090749.91+010142.7	3.30	$1.33 \cdot 10^{9}$	9.7	$1.29 \cdot 10^{10}$
232	J091051.39+011207.5	J091051.50+011203.6	2.44	$1.39 \cdot 10^{9}$	5.2	$7.15 \cdot 10^{9}$
		J091051.39+011207.5	1.78	$1.42 \cdot 10^{9}$	5.2	$7.42 \cdot 10^{9}$
		J091051.89+011233.8	28.91	$1.14 \cdot 10^{6}$	2.9	$3.31 \cdot 10^{6}$
2784	J091055.22+001554.9	J091055.22+001554.9	5.27	$1.15 \cdot 10^{9}$	11.6	$1.34 \cdot 10^{10}$
		J091055.18+001606.2	6.06	$1.06 \cdot 10^{9}$	12.3	$1.31 \cdot 10^{10}$
		J091055.60+001551.6	10.16	$6.02 \cdot 10^{8}$	1	$6.02 \cdot 10^{8}$
6108	J090529.13+003455.7	J090529.28+003505.6	9.15	$7.11 \cdot 10^{8}$	$3.84 \cdot 10^{-5}$	27289
		J090529.13+003455.7	2.38	$1.39 \cdot 10^{9}$	17.8	$2.47 \cdot 10^{10}$
3195	J090118.51+000414.0	J090118.04+000413.5	4.40	$1.23 \cdot 10^{9}$	9.3	$1.15 \cdot 10^{10}$
		J090118.51+000414.0	2.71	$1.37 \cdot 10^{9}$	23.0	$3.15 \cdot 10^{10}$
3507	J090259.51+013406.6	J090259.51+013406.6	2.08	$1.40 \cdot 10^{9}$	21.4	$3.01 \cdot 10^{10}$
		J090259.99+013414.6	8.86	$7.44 \cdot 10^{8}$	21.4	$1.59 \cdot 10^{10}$
3042	J085813.09+012222.1	J085812.87+012202.2	18.16	$8.65 \cdot 10^{7}$	8.8	$7.57 \cdot 10^{8}$
		J085813.09+012222.1	3.13	$1.34 \cdot 10^{9}$	8.8	$1.17 \cdot 10^{10}$

Tabla 7. Registro de diversos casos de asociaciones entre objetos de Herschel ATLAS y GAMA en los que se ha logrado una identificación correcta de las contrapartidas oficiales de los distintos objetos de Herschel ATLAS tanto con el factor de Bayes posicional como con el factor de Bayes conjunto. Entre todos estos ejemplos, hay casos en los que tanto B_{pos} como B_z son muy similares entre las distintas contrapartidas posibles $(3817, 29, 232)$, casos en los que B_z favorece la identificación entre las contrapartidas $(6108, 3196)$, casos en los que B_{pos} es determinante al poseer las contrapartidas corrimientos al rojo similares $(3507, 3042)$ y casos en los que B_z es mayor en una contrapartida distinta de la identificada oficialmente $(2116, 460, 2784)$.

Atendiendo al estudio de las identificaciones oficiales realizado anteriormente, se ha comprobado que para poder considerar razonable una contrapartida, esta debe tener un factor de Bayes posicional de al menos $B_{pos} = 9.82 \cdot 10^{8}$, lo que equivale a una separación angular máxima de $\Delta\theta = 6.79"$. De este modo, entre las 2387 asociaciones conseguidas que carecen de identificación oficial en GAMA se han identificado 124 identificaciones muy posibles, las cuales se recogen en el Apéndice II.

7. Conclusiones

Se ha conseguido implementar un método probabilístico bayesiano de cross-identificación de objetos extragalácticos, basado en el formalismo teórico propuesto por Tamás Budavári y Alexander S. Szalay [1], sobre dos catálogos astronómicos obtenidos con instrumentos de distintas resoluciones angulares, como son el Observatorio Espacial Herschel y el proyecto GAMA. Este método no sólo ha incorporado información tanto de las distintas resoluciones angulares y las distancias relativas entre los candidatos, sino también otra información adicional de carácter astrofísico conocida a priori como es el corrimiento al rojo.

Respecto a la eficacia del método desarrollado, el 95.9 % de las identificaciones conseguidas coincide con las identificaciones oficiales de Herschel ATLAS. Si se considera únicamente el factor de Bayes posicional éste número asciende al 96.9 %, debido a la falta de información sobre el corrimiento al rojo en objetos extragalácticos que constituirían la contrapartida correcta. No obstante, hay que considerar que de no incluirse información referente al corrimiento al rojo no se podría discernir correctamente entre dos posibles contrapartidas a un objeto muy cercanas posicionalmente pero con valores del corrimiento al rojo muy distintos, como el caso mostrado en la Tabla 6. En este trabajo sólo se ha encontrado un caso de este tipo, pero en los grandes cartografiados del Universo que albergan del orden de $10^7 - 10^8$ objetos astronómicos esta cuestión adquiere mucho más peso.

Si se ahonda más en las causas de no haber identificado correctamente algunos objetos extragalácticos, se descubre que en la gran mayoría de estos casos es debido a que la contrapartida oficial no estaba incluida en la publicación de datos de GAMA utilizada. Es por ello que, la eficacia del método aumentaría al 99.8 % usando el factor de Bayes posicional y al 98.8 % considerando el factor de Bayes conjunto. Una vez analizadas todas las asociaciones de objetos obtenidas, se ha llegado a la conclusión de que se podía haber impuesto una separación máxima entre objetos menor que los 54" utilizados, no repercutiendo en la eficacia de la herramienta pero facilitando el análisis de los datos.

Como se observa, el método desarrollado proporciona muy buenos resultados, teniendo en cuenta las simplificaciones y aproximaciones tomadas para las funciones densidad de probabilidad a priori y las funciones de verosimilitud, y la inclusión del corrimiento al rojo como única información adicional a priori, sirviendo como acercamiento al problema creciente en astrofísica de tratar una gran cantidad de datos en múltiples observaciones distintas.

Este trabajo puede servir como pie para la realización de un proyecto de cross-identificación de objetos extragalácticos más ambicioso que no sólo no se limitase a cotejar únicamente dos catálogos astronómicos sino que además utilizase modelos de distribuciones de probabilidad lo más realistas posibles e incluyese cualquier otra información conocida a priori de carácter astrofísico como son el color, la luminosidad o la morfología de los objetos extragalácticos, permitiendo una completa e inequívoca identificación de las contrapartidas estudiadas.

Referencias Bibliográficas

[1] Tamás Budavári & Alexander S. Szalay / Probabilistic Cross-Identification of Astronomical Sources. *The Astrophysical Journal*, 679:301-309, 2008 May 20.

[2] Tamás Budavári / Probabilistic Cross-Identification of Cosmic Events. *The Astrophysical Journal*, 736:155 (5pp), 2011 August 1.

[3] Gianfranco de Zotti, Marcella Massardi, Mattia Negrello & Jasper Wall / Radio and Millimeter Continuum Surveys and their Astrophysical Implications. *Astron Astrophys* Rev, 18:1-65, 2010.

[4] http://www.h-atlas.org/

[5] http://www.h-atlas.org/results/highlights/first-data-release-herschel-atlas

[6] http://www.h-atlas.org/public_data/HATLAS_SDP_catalogue.fits.gz

[7] http://www.h-atlas.org/public_data/HATLAS_SDP_catalogue.README

[8] http://www.h-atlas.org/public_data/HATLAS_SDP_catalogue.columns

[9] http://sci.esa.int/herschel/

[10] http://herschel.cf.ac.uk/

[11] http://www.cosmos.esa.int/web/herschel/home

[12] http://www.mpi-hd.mpg.de/personalhomes/rjt/index.shtml

[13] http://www.gama-survey.org/

[14] http://www.gama-survey.org/dr1/YR1public.php

[15] http://www.gama-survey.org/dr1/data/GamaCoreDR1_v1.fits

[16] http://www.gama-survey.org/dr1/data/GamaCoreDR1_v1.par

[17] http://www.gama-survey.org/dr1/data/GamaCoreDR1_v1.notes

[18] Driver et al / GAMA: Towards a physical understanding of galaxy formation. Astronomy & Geophysics, Volume 50, Issue 5, pp. 5.12-5.19. October 2009.

[19] https://en.wikipedia.org/wiki/Galaxy_And_Mass_Assembly_survey

[20] http://halweb.uc3m.es/esp/Personal/personas/causin/esp/2012-2013/SMB/Tema6.pdf

[21] http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Bayes/tema1bayes.pdf

[22] http://www.ugr.es/~fdeasis/Material/InferenciaLicenciatura/Inferencia_Tema4Clase.pdf

[23] Robert E. Kass & Adrian E. Raftery / Bayes Factors. Journal of the American Statistical Association, June 1995, Vol. 90, No. 430, Review Paper.

[24] https://en.wikipedia.org/wiki/Bayesian_statistics

[25] http://www.star.bris.ac.uk/~mbt/stilts/

[26] http://www.star.bris.ac.uk/~mbt/stilts/sun256.pdf

[27] www.star.bris.ac.uk/~mbt/stilts/stilts.jar

[28] http://www.star.bris.ac.uk/~mbt/topcat/

[29] http://www.star.bris.ac.uk/~mbt/topcat/sun253.pdf

[30] http://www.star.bris.ac.uk/~mbt/topcat/topcat-full.jar

Apéndice I

- Resolución numérica del Factor de Bayes considerando una Distribución Normal Esférica para caracterizar la precisión de las observaciones astronómicas

En este apéndice se trata la resolución matemática del factor de Bayes en el caso común, cuando se considera una distribución normal esférica para modelar la precisión astrométrica. Además, también se adopta una distribución de probabilidad a priori de todo el cielo en esta derivación.

En primer lugar, partiendo de la definición del factor de Bayes de la ecuación (34) y de las parametrizaciones de $p(D|H)$ y $p(D|K)$ de las ecuaciones (35) y (36), y asumiendo la distribución normal esférica de la ecuación (38) como función de verosimilitud para modelar la precisión astrométrica y adoptando la función densidad de probabilidad a priori trivial para todo el cielo de la ecuación (31), se tiene que:

$$p(D|\boldsymbol{m},H)=p(\{\boldsymbol{x_i}\}|\boldsymbol{m},H)=\prod_{i=1}^{n}p_i\,(\boldsymbol{x_i}|\boldsymbol{m},H)=\prod_{i=1}^{n}N\,(\boldsymbol{x_i}|\boldsymbol{m},w_i)=\prod_{i=1}^{n}\frac{w_i\delta(|\boldsymbol{x_i}|-1)}{4\pi\sinh w_i}e^{w_i\boldsymbol{m}\boldsymbol{x_i}}$$

$$p(D|\boldsymbol{m_i},K)=p(\{\boldsymbol{x_i}\}|\{\boldsymbol{m_i}\},K)=\prod_{i=1}^{n}p_i\,(\boldsymbol{x_i}|\boldsymbol{m_i},K)=\prod_{i=1}^{n}N\,(\boldsymbol{x_i}|\boldsymbol{m_i},w_i)=\prod_{i=1}^{n}\frac{w_i\delta(|\boldsymbol{x_i}|-1)}{4\pi\sinh w_i}e^{w_i\boldsymbol{m_i}\boldsymbol{x_i}}$$

donde se aprovecha el hecho de que el producto de distribuciones normales tiene la misma forma funcional.

Primero, se centra la atención en la función de verosimilitud de la hipótesis H, sustituyendo la función densidad de probabilidad a priori de la ecuación (31) y la expresión de $p(D|\boldsymbol{m},H)$ anterior:

$$p(D|H)=\int\frac{\delta(|\boldsymbol{m}|-1)}{4\pi}\prod_{i=1}^{n}\frac{w_i\delta(|\boldsymbol{x_i}|-1)}{4\pi\sinh w_i}e^{w_i\boldsymbol{m}\boldsymbol{x_i}}\,d^3m$$

A continuación, se sacan fuera de la integral los términos no dependientes de la variable de integración m, teniendo en cuenta la siguiente propiedad de las exponenciales:

$$\prod_{i=1}^{n}e^{w_i\boldsymbol{m}\boldsymbol{x_i}}=e^{\sum_{i=1}^{n}w_i\boldsymbol{m}\boldsymbol{x_i}}$$

De este modo, se tiene que:

$$p(D|H)=\left[\prod_{i=1}^{n}\frac{w_i\delta(|\boldsymbol{x_i}|-1)}{4\pi\sinh w_i}\right]\int\frac{\delta(|\boldsymbol{m}|-1)}{4\pi}\,e^{\sum_{i=1}^{n}w_i\boldsymbol{m}\boldsymbol{x_i}}\,d^3m$$

Seguidamente, introduciendo:

$$w\boldsymbol{x}=\sum_{i=1}^{n}w_i\boldsymbol{x_i}$$

Se llega a la expresión:

$$p(D|H) = \left[\prod_{i=1}^{n} \frac{w_i \delta(|\boldsymbol{x_i}| - 1)}{4\pi \sinh w_i}\right] \int \frac{\delta(|\boldsymbol{m}| - 1)}{4\pi} \, e^{w\boldsymbol{mx}} d^3 m$$

Ahora se multiplica y divide por $\sinh w \,/w$ para llegar a:

$$p(D|H) = \left[\frac{\sinh w}{w} \prod_{i=1}^{n} \frac{w_i \delta(|\boldsymbol{x_i}| - 1)}{4\pi \sinh w_i}\right] \int \frac{w \, \delta(|\boldsymbol{m}| - 1)}{4\pi \sinh w} \, e^{w\boldsymbol{mx}} d^3 m$$

De esta forma, dentro de la integral se puede identificar una distribución normal esférica $N(\boldsymbol{m}|\boldsymbol{x}, w)$ en su forma normalizada, análoga a la mostrada en la ecuación (38), por lo que se tiene que:

$$\int \frac{w \, \delta(|\boldsymbol{m}| - 1)}{4\pi \sinh w} \, e^{w\boldsymbol{mx}} d^3 m = \int N(\boldsymbol{m}|\boldsymbol{x}, w) \, d^3 m = 1$$

Simplificándose la expresión de $p(D|H)$ hasta:

$$p(D|H) = \frac{\sinh w}{w} \prod_{i=1}^{n} \frac{w_i}{\sinh w_i} \frac{\delta(|\boldsymbol{x_i}| - 1)}{4\pi}$$

La función de verosimilitud de la hipótesis alternativa K se calcula de forma similar, obteniendo que:

$$p(D|K) = \prod_{i=1}^{n} \int \frac{\delta(|\boldsymbol{m_i}| - 1)}{4\pi} \, \frac{w_i \delta(|\boldsymbol{x_i}| - 1)}{4\pi \sinh w_i} e^{w_i \boldsymbol{m_i x_i}} \, d^3 m_i = \prod_{i=1}^{n} \frac{\delta(|\boldsymbol{x_i}| - 1)}{4\pi}$$

Y por lo tanto, la resolución analítica de la expresión del factor de Bayes de la ecuación (34) queda tal y como se mostró en la ecuación (40):

$$B(H, K|D) = \frac{\sinh w}{w} \prod_{i=1}^{n} \frac{w_i}{\sinh w_i}$$

con

$$w = \left| \sum_{i=1}^{n} w_i \boldsymbol{x_i} \right|$$

Apéndice II

En este apéndice se exponen las 124 posibles identificaciones encontradas entre las 2387 asociaciones de objetos del catálogo GAMA y objetos del catálogo Herschel ATLAS que carecen de una identificación oficial, considerando que deben tener un factor de Bayes posicional de al menos $B_{pos} = 9.82 \cdot 10^8$, es decir a una separación angular máxima permitida de $\Delta\theta = 6.79"$. Nuevamente, en esta tabla SDP_ID es el identificador del correspondiente objeto del catálogo Herschel ATLAS y GAMA_IAU_ID es el identificador de las diversas contrapartidas de GAMA obtenidas por nuestra herramienta tras realizar el cross-matching:

SDP_ID	GAMA_IAU_ID	$\Delta\theta$ / "	B
65	J090913.28+012105.5	5.77	$1.10 \cdot 10^9$
133	J090215.16+012256.5	5.69	$1.11 \cdot 10^9$
	J090215.69+012246.6	7.33	$9.21 \cdot 10^8$
142	J090600.63+021234.2	4.27	$1.25 \cdot 10^9$
179	J091233.04-002729.2	4.57	$1.22 \cdot 10^9$
183	J090404.05+005623.9	4.08	$1.26 \cdot 10^9$
269	J090726.99+003157.6	6.40	$1.03 \cdot 10^9$
298	J090809.25+000024.2	6.24	$1.04 \cdot 10^9$
308	J090347.46+022010.0	4.20	$1.25 \cdot 10^9$
313	J091235.78-005218.8	4.04	$1.27 \cdot 10^9$
387	J090024.10+015703.6	4.21	$1.25 \cdot 10^9$
405	J090223.27-003327.6	4.72	$1.20 \cdot 10^9$
410	J090743.85+014236.1	5.81	$1.09 \cdot 10^9$
429	J090037.92-003525.5	3.73	$1.29 \cdot 10^9$
450	J090547.62+022243.0	4.99	$1.18 \cdot 10^9$
461	J091249.77-000359.7	5.41	$1.14 \cdot 10^9$
474	J085957.90+015634.1	2.41	$1.39 \cdot 10^9$
494	J091027.07+010934.8	1.95	$1.41 \cdot 10^9$
496	J090956.10+015226.7	5.30	$1.15 \cdot 10^9$
544	J091055.15+004457.9	4.33	$1.24 \cdot 10^9$
579	J085917.86+004559.9	5.69	$1.11 \cdot 10^9$
581	J090449.64+002822.8	5.09	$1.17 \cdot 10^9$
612	J091403.57-005035.2	5.04	$1.17 \cdot 10^9$
627	J090725.60+021912.5	5.05	$1.17 \cdot 10^9$
680	J091232.98-001303.8	4.49	$1.23 \cdot 10^9$
692	J090456.99+014346.4	6.72	$9.90 \cdot 10^8$
697	J090334.64+014321.6	4.51	$1.22 \cdot 10^9$
740	J091322.66+000016.5	5.26	$1.15 \cdot 10^9$
744	J090619.73+015440.6	4.69	$1.21 \cdot 10^9$

SDP_ID	GAMA_IAU_ID	$\Delta\theta$ / "	B
750	J090543.08+005813.9	2.96	$1.35 \cdot 10^9$
	J090542.89+005818.1	2.45	$1.39 \cdot 10^9$
799	J090252.47-005559.1	4.57	$1.22 \cdot 10^9$
816	J090115.26+013345.8	5.37	$1.14 \cdot 10^9$
818	J091138.38-004253.9	5.30	$1.15 \cdot 10^9$
827	J090843.67+003258.0	5.05	$1.17 \cdot 10^9$
916	J090144.56+005008.2	4.05	$1.27 \cdot 10^9$
919	J090108.77+012839.6	4.59	$1.22 \cdot 10^9$
995	J090813.78+022655.2	4.85	$1.19 \cdot 10^9$
1037	J090333.20-001659.4	5.26	$1.15 \cdot 10^9$
1067	J090622.09+014426.4	3.65	$1.30 \cdot 10^9$
1069	J090628.89+001815.1	4.76	$1.20 \cdot 10^9$
1094	J090758.22+023232.9	4.78	$1.20 \cdot 10^9$
1097	J085903.67+000350.6	5.93	$1.08 \cdot 10^9$
1106	J090410.98+000244.2	6.63	$1.00 \cdot 10^9$
1162	J090734.94+013118.8	3.95	$1.28 \cdot 10^9$
1218	J090224.17+002844.9	6.32	$1.04 \cdot 10^9$
1271	J090902.09-000523.0	4.55	$1.22 \cdot 10^9$
1364	J091159.95-002737.0	6.56	$1.01 \cdot 10^9$
1378	J090213.96+015348.0	5.72	$1.10 \cdot 10^9$
1382	J091320.29-001052.2	4.69	$1.21 \cdot 10^9$
1446	J090639.33+000336.4	6.00	$1.07 \cdot 10^9$
1475	J090130.13-000218.9	3.19	$1.34 \cdot 10^9$
	J090130.40-000214.4	3.28	$1.33 \cdot 10^9$
1486	J090105.21-001105.5	2.45	$1.39 \cdot 10^9$
1521	J085819.54+010349.6	6.46	$1.02 \cdot 10^9$
1618	J090759.85+020856.2	6.76	$9.85 \cdot 10^8$
1620	J091105.93+012942.8	6.02	$1.07 \cdot 10^9$
1640	J090421.19+021040.8	4.96	$1.18 \cdot 10^9$

SDP_ID	GAMA_IAU_ID	$\Delta\theta$ / "	B
1701	J090449.42+004036.4	5.70	$1.10 \cdot 10^9$
1711	J090110.99+003701.5	5.80	$1.09 \cdot 10^9$
1728	J090719.02+013715.7	6.76	$9.85 \cdot 10^8$
1860	J091310.41+002839.6	5.32	$1.14 \cdot 10^9$
1873	J091421.75-001656.5	5.03	$1.17 \cdot 10^9$
1945	J090324.39+010104.8	5.06	$1.17 \cdot 10^9$
1968	J091440.77-003629.1	6.49	$1.02 \cdot 10^9$
1984	J090916.82+025333.0	5.30	$1.15 \cdot 10^9$
2023	J091033.44+012816.8	5.48	$1.13 \cdot 10^9$
2081	J091333.38-000448.6	5.85	$1.09 \cdot 10^9$
2119	J085713.53+013054.2	5.11	$1.16 \cdot 10^9$
2162	J090539.74+004705.1	5.91	$1.08 \cdot 10^9$
2212	J090939.65-000357.4	6.64	$9.99 \cdot 10^8$
2282	J090859.74+000117.2	4.49	$1.23 \cdot 10^9$
2353	J090153.75-003354.2	3.61	$1.30 \cdot 10^9$
2390	J090614.17+015031.4	5.43	$1.13 \cdot 10^9$
2403	J090615.94+022712.6	6.38	$1.03 \cdot 10^9$
2414	J090830.31+001220.4	5.86	$1.09 \cdot 10^9$
2417	J085942.66+001532.0	4.70	$1.21 \cdot 10^9$
2457	J085924.82+001300.9	5.70	$1.10 \cdot 10^9$
2469	J090323.73+012040.8	5.49	$1.13 \cdot 10^9$
2514	J090325.72+020234.1	5.27	$1.15 \cdot 10^9$
2548	J091334.93-004534.1	6.67	$9.96 \cdot 10^8$
2549	J090803.52+002251.4	5.21	$1.16 \cdot 10^9$
2626	J091231.91-005018.6	6.43	$1.02 \cdot 10^9$
2689	J090657.45-002806.9	6.50	$1.02 \cdot 10^9$
2744	J090109.93+012243.1	6.10	$1.06 \cdot 10^9$
2767	J091027.68+001709.8	6.46	$1.02 \cdot 10^9$
2858	J090631.26+004603.7	2.22	$1.40 \cdot 10^9$
	090631.34+004607.8	2.74	$1.37 \cdot 10^9$
2995	J091030.43+020746.1	5.32	$1.14 \cdot 10^9$
3113	J090839.10+004107.0	5.73	$1.10 \cdot 10^9$
3193	J090507.81-000704.1	5.34	$1.14 \cdot 10^9$
3245	J090423.18+003406.6	5.75	$1.10 \cdot 10^9$
3270	J090500.87+005209.5	6.12	$1.06 \cdot 10^9$
3285	J091303.35-004858.8	6.02	$1.07 \cdot 10^9$

SDP_ID	GAMA_IAU_ID	$\Delta\theta$ / "	B
3327	J090247.41+014119.9	5.68	$1.11 \cdot 10^9$
3384	J090112.03+013916.6	5.82	$1.09 \cdot 10^9$
3393	J090334.04+022011.1	6.69	$9.93 \cdot 10^8$
3394	J090844.06+020816.3	6.07	$1.06 \cdot 10^9$
3512	J091123.62-004946.9	4.45	$1.23 \cdot 10^9$
3519	J090511.34-000846.1	5.64	$1.11 \cdot 10^9$
3594	J090240.82-002759.4	2.56	$1.38 \cdot 10^9$
3650	J090554.18+005127.1	3.95	$1.28 \cdot 10^9$
	J090554.17+005122.3	6.17	$1.05 \cdot 10^9$
3657	J090300.01-000852.0	6.70	$9.92 \cdot 10^8$
3719	J090904.46+020729.0	3.68	$1.30 \cdot 10^9$
3756	J090530.53+001442.5	5.26	$1.15 \cdot 10^9$
4025	J090714.56-001834.1	6.17	$1.05 \cdot 10^9$
4123	J091239.97+002645.9	5.52	$1.12 \cdot 10^9$
4139	J091233.75+002906.8	6.49	$1.02 \cdot 10^9$
4396	J090136.70+020433.8	4.58	$1.22 \cdot 10^9$
4411	J085914.00+000637.3	6.12	$1.06 \cdot 10^9$
4506	J090202.79-001238.3	3.92	$1.28 \cdot 10^9$
4555	J090646.94-001356.5	5.24	$1.15 \cdot 10^9$
4568	J090248.50+005129.4	6.22	$1.05 \cdot 10^9$
4677	J090715.36+024106.3	6.64	$9.99 \cdot 10^8$
4692	J090911.42+011730.0	6.35	$1.03 \cdot 10^9$
4720	J090021.50+011234.4	5.40	$1.14 \cdot 10^9$
4847	J090403.70+015911.6	6.15	$1.05 \cdot 10^9$
4878	J090950.86+013955.1	5.52	$1.12 \cdot 10^9$
4906	J090146.49-002059.2	5.81	$1.09 \cdot 10^9$
4947	J090432.37+012853.9	5.78	$1.10 \cdot 10^9$
4967	J091110.87-000502.2	5.63	$1.11 \cdot 10^9$
5034	J090038.68+013752.6	5.40	$1.14 \cdot 10^9$
5041	J091014.29+021321.6	6.49	$1.02 \cdot 10^9$
5079	J091309.41-003933.4	6.60	$1.00 \cdot 10^9$
5775	J090604.73+014944.7	6.26	$1.04 \cdot 10^9$
6979	J090148.16-003714.4	6.79	$9.83 \cdot 10^8$
8800	J091121.79-003736.1	2.84	$1.36 \cdot 10^9$
11962	J090450.16+015008.4	5.03	$1.17 \cdot 10^9$