



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS, UNAM

FUNCIONES DE CORRELACIÓN DE DOS
PUNTOS CON ALGORITMOS DE
AGRUPAMIENTO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
LICENCIATURA EN FÍSICA

PRESENTA:
SAMANTHA RIZO FRANCO

TUTOR:

DR. JOSÉ ALBERTO VÁZQUEZ GONZÁLEZ

Ciudad Universitaria, CD. MX. 2023





**CONSTANCIA DE
PRESENTACIÓN DE
EXAMEN PROFESIONAL**

En la Universidad Nacional Autónoma de México, en la SALA DE EXÁMENES PROFESIONALES de la FACULTAD DE CIENCIAS siendo las 11:00 horas del día 03 de octubre de 2023, la alumna:

SAMANTHA RIZO FRANCO

de nacionalidad MEXICANA con número de cuenta 315122046 se presentó con el fin de sustentar el examen oral para obtener el título de:

FÍSICA

en su modalidad de titulación por TESIS con el trabajo titulado: Funciones de correlación de dos puntos con algoritmos de agrupamiento. La alumna cursó sus estudios en el periodo 2018-1 a 2023-2 habiendo obtenido un promedio de 9.35 y cumpliendo con los requisitos académicos señalados en el plan de estudios 1081 aprobado por el H. Consejo Universitario.

El Jurado designado por el Comité Académico integrado por:

PRESIDENTE: DR. SEBASTIEN MICKAEL MARC FROMENTEAU
SECRETARIO: DR. JOSE ALBERTO VAZQUEZ GONZALEZ
VOCAL: M. EN C. RICARDO MARTIN HERNANDEZ FLORES
SUPLENTE: DR. TONATIUH MATOS CHASSIN
SUPLENTE: DR. JUAN CARLOS HIDALGO CUELLAR

Tras el interrogatorio y deliberación resolvió otorgarle la calificación de:

Aprobada con Mención Honorífica

Procediendo a informarle el resultado, tomarle la Protesta Universitaria y dar por concluido el acto.



PRESIDENTE DEL JURADO



SECRETARIO DEL JURADO



VOCAL DEL JURADO

El Titular de la entidad académica hace constar que las firmas electrónicas que anteceden son válidas y corresponden a los miembros del jurado.


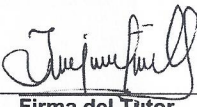



**"POR MI RAZA HABLARÁ EL ESPÍRITU"
DR. VÍCTOR MANUEL VELÁZQUEZ AGUILAR
DIRECTOR DE LA ENTIDAD ACADÉMICA**



FORMATO PARA REGISTRO DE OPCIÓN DE TITULACIÓN (FÍSICA)

Este documento puede ser llenado en computadora utilizando Acrobat Reader

OPCIÓN SOLICITADA			
<input checked="" type="checkbox"/> TESIS <input type="checkbox"/> ACTIVIDAD DE APOYO A LA INVESTIGACIÓN <input type="checkbox"/> SEMINARIO DE TITULACIÓN <input type="checkbox"/> EXAMEN GENERAL DE CONOCIMIENTOS <input type="checkbox"/> ALTO RENDIMIENTO ACADÉMICO	<input type="checkbox"/> ACTIVIDAD DE APOYO A LA DOCENCIA <input type="checkbox"/> TRABAJO PROFESIONAL <input type="checkbox"/> ESTUDIOS DE POSGRADO <input type="checkbox"/> AMPLIACIÓN Y PROFUNDIZACIÓN DE CONOCIMIENTOS <input type="checkbox"/> SERVICIO SOCIAL		
DATOS DEL ESTUDIANTE			
Rizo	Franco	Samantha	
Apellido Paterno	Apellido Materno	Nombre(s)	
FÍSICA			315122046
Carrera			Número de cuenta
(55) 5604 4780			(55) 2569 6125
Teléfono domicilio	Teléfono oficina	Teléfono celular	
samantharizo@ciencias.unam.mx		sammyrizorm@gmail.com	
Correo electrónico 1		Correo electrónico 2	
DATOS DEL TUTOR			
Dr.	José Alberto	Vázquez	González
Grado	Nombre(s)	Apellido Paterno	Apellido Materno
Instituto de Ciencias Físicas			
Dependencia UNAM o empresa en la que trabaja			
-	(777) 329-17-77 Ext 38343		-
Teléfono domicilio	Teléfono oficina	Teléfono celular	
javazquez@icf.unam.mx		-	
Correo electrónico 1		Correo electrónico 2	
DATOS DE LA INSTITUCIÓN EN LA QUE LLEVÓ A CABO EL TRABAJO PROFESIONAL O EL SERVICIO SOCIAL			
Instituto de Ciencias Físicas		Abril 2022 - Noviembre 2022	
Institución		Periodo en el que realizó la actividad	
TÍTULO TENTATIVO DEL TRABAJO ESCRITO			
Funciones de correlación de dos puntos con algoritmos de agrupamiento			
FIRMAS			
 Firma del Alumno		 Firma del Tutor	
DICTAMEN			
APROBADO (X)	 Dra. Gerardo Jorge Vázquez Fonseca Coordinador del Comité Académico de la Licenciatura en Física 04 DE MAYO DE 2022		
CONDICIONADO ()			
RECHAZADO ()			

FACULTAD DE CIENCIAS



DEPARTAMENTO DE FÍSICA

LIC. FÍSICA

Para todas las opciones de titulación, este registro debe entregarse junto con el proyecto y los documentos probatorios indicados en la sección "Anexos"

1. Datos del alumno

Rizo
Franco
Samantha
55 2569 6125
Universidad Nacional Autónoma de
México
Facultad de Ciencias
Física
315122046

2. Datos del tutor

Dr
José Alberto
Vázquez
González

3. Datos del sinodal 1

Dr
Tonatiuh
Matos
Chassin

4. Datos del sinodal 2

Dr
Juan Carlos
Hidalgo
Cuéllar

5. Datos del sinodal 3

Dr
Sébastien Mickaël
Marc
Fromenteau

6. Datos del sinodal 4

M en C
Ricardo Martí
Hernández
Flores

7. Datos del trabajo escrito

Funciones de correlación de dos puntos con algoritmos de agrupamiento
138p
2023

*A Fanny, Pascual, Graciela,
Ivanna, Valentina y Miguel Angel.
Por su guía y amor incondicional.*

Agradecimientos

Antes que nada, deseo expresar mi profunda admiración y gratitud al Dr. José Alberto Vázquez, mi asesor, por compartir conmigo su invaluable conocimiento, por su inagotable paciencia y por recordarme lo mucho que me apasiona la ciencia. Agradezco su apoyo incondicional al permitirme participar en diversos eventos y por ser una fuente constante de inspiración en mi desarrollo profesional. Además, valoro enormemente su guía y consejo en mi vida personal, enseñándome que todos cometemos errores y que algún día podremos reírnos de ellos. Sin lugar a dudas, es una persona excepcional y un destacado científico. Cada vez que reflexiono sobre su trayectoria, vuelvo a asombrarme, estoy verdaderamente honrada y agradecida por contar con usted como mi asesor, colega y amigo.

A Jazhiel, te agradezco por brindarme tu apoyo incondicional para culminar este proyecto. Valoraré siempre tus conocimientos compartidos, tu disposición para escuchar mis ideas y tu infinita paciencia. Hoy te considero no solo un colega, sino también un gran amigo.

Expreso mi gratitud a CONACYT, ICF y PAPIIT por el apoyo y las facilidades proporcionadas a través de los proyectos financiados: CONACYT (A1-S-21925) y PAPIIT (IA104221). Estos proyectos fueron la base para la realización de mi tesis, y les estoy sinceramente agradecida por ello.

A mis amigos, especialmente a Celeste, Jenn, Nelly y Martha les agradezco de corazón por acompañarme en los momentos buenos y malos. Gracias por enriquecer mi experiencia universitaria y convertir la facultad en una etapa llena de hermosos recuerdos.

A mi familia, les debo un amor infinito por el constante apoyo que siempre me han brindado. Agradezco profundamente por proporcionarme las condiciones necesarias para dedicarme por completo a mis estudios, por su comprensión en mis momentos de crisis y por haber compartido las mejores experiencias a su lado. Desde mi infancia, me motivaron en mis intereses científicos y por eso siempre estaré agradecida.

A mi padre, agradezco sinceramente por enseñarme que las matemáticas no son difíciles, por dedicar tu tiempo y energía en ayudarme con mis tareas y por compartir noches de desvelo mientras juntos trabajábamos en mis proyectos de física y por nunca pensar dos veces en apoyarme cuando se trata de salud o educación. Eres quien me brinda tranquilidad y confianza en cada paso que doy.

A mi madre, te agradezco desde lo más profundo de mi ser por ser la fuerza que me impulsa a soñar en grande. Tu inquebrantable fe en mí y tu constante compañía en cada

paso de mi camino han sido el motor que me impulsa a alcanzar mis metas. Gracias por enseñarme que no hay límites para lo que uno se propone y por demostrarme que los sueños pueden convertirse en realidad. Eres mi confidente y mi mayor inspiración para superarme constantemente.

A mi abu, gracias por consentirme tanto, por ser una de mis más grandes confidentes, por demostrarme que yo solita puedo lograr lo que me proponga, por estar a mi lado en cada desafío académico y por llenar mi vida con tu amor y apoyo incondicional. Tu influencia en mi vida es invaluable, te quiero con toda mi alma y espero algún día poder regresarte al menos un poco de todo lo que tú me has dado.

A mis hermanas, ustedes son las razones por las cuales deseo dejar la mejor versión de mí en este mundo. Las amo con todo mi corazón y les agradezco por tantas alegrías compartidas y por su constante comprensión. Ivannita gracias por aguantar cada locura, por todas las risas y por siempre estar cuando te necesito.

Finalmente quiero expresar mi más profundo agradecimiento a mi querido esposo, Miguel Angel, el amor de mi vida y mi mejor amigo. Te agradezco por ser mi principal confidente y mi soporte, por ser mi compañero en sueños y viajes, por permanecer a mi lado cada desvelo motivándome para culminar este y todos los proyectos de la carrera, por escuchar mis ideas y aportar tu punto de vista, por tus curiosas preguntas que me motivan a investigar para poder responderte, por apoyarme en todos los aspectos de mi vida tanto personal como profesional, por motivarme a seguir adelante, por formar una familia conmigo y por ser mi mayor inspiración para culminar este proyecto. Junto a ti todo ha sido más sencillo y divertido. Tu amor y apoyo incondicional han sido fundamentales en mi éxito y estoy segura que juntos cosecharemos todos los frutos de este esfuerzo.

Sin lugar a dudas, a todos los mencionados anteriormente les debo mucho, este logro también es en parte suyo. Su influencia y apoyo han sido invaluableles en mi trayectoria académica y personal, y siempre los recordaré con amor, gratitud y cariño.

*“The cosmos is within us. We are made of star-stuff.
We are a way for the universe to know itself.”
- Carl Sagan*

Índice general

Agradecimientos	1
Resumen	11
1. Marco teórico	15
1.1. Breve historia de la cosmología	15
1.1.1. Principio cosmológico	17
1.2. Modelo estándar (Λ CDM)	18
1.2.1. Ecuaciones de Einstein	18
1.2.2. Métrica de Friedman-Lemaître-Robertson-Walker (FLRW)	19
1.2.3. Expansión del Universo	21
1.3. Composición del Universo	23
1.3.1. Materia bariónica y radiación	24
1.3.2. Energía oscura	24
1.3.3. Materia oscura	24
1.4. Fondo cósmico de microondas (CMB)	26
1.4.1. Oscilaciones Acústicas de Bariónes	28
2. Distancias	33
2.1. Distancias en cosmología	33
2.1.1. Paralaje trigonométrico	33
2.1.2. Distancia lumínica	33
2.1.3. Distancia comóvil	34
2.1.4. Distancia angular	35
2.2. Mediciones	36
2.2.1. Relojes estándar	36
2.2.2. Candelas estándar	36
2.2.3. Reglas estándar	37
2.3. Simulaciones cosmológicas	37
2.3.1. Millenium	38
2.3.2. Bolshoi	39
2.3.3. Uchuu	39

3. Función de Correlación	41
3.0.1. Transformada de Fourier	42
3.0.2. Aplicaciones	42
3.0.3. Estimadores de la Función de Correlación	42
3.0.4. Versión angular de la función de correlación	44
4. Aprendizaje automático	46
4.1. Métodos de agrupamiento	47
4.1.1. DBSCAN	51
4.1.2. OPTICS	54
4.1.3. HDBSCAN	61
4.1.4. DRL-DBSCAN	66
5. Cálculo de funciones de correlación de dos puntos	68
5.1. Distribución de galaxias	68
5.1.1. Datos aleatorios	70
5.1.2. Datos de galaxias	77
5.1.3. 2pcf en función de la cantidad de galaxias y anillos	79
5.1.4. 2pcf en función de la perturbación	81
5.1.5. 2pcf con distinto porcentaje de galaxias en el centro respecto a las galaxias que hay sobre el anillo de BAO	82
5.1.6. 2pcf con distinta cantidad de galaxias sobre el anillo de BAO	83
5.2. Distintas formas de obtener la distribución aleatoria	84
5.3. Estimadores de la función de correlación	89
5.3.1. 2pcf en función de la cantidad de galaxias y anillos	89
5.3.2. 2pcf en función de la perturbación	89
5.4. Reconstrucción del pico de BAO original	90
6. Algoritmos de agrupamiento con funciones de correlación	94
6.1. Búsqueda de centros de BAO con DBSCAN	94
6.2. Reconstrucción de la posición original de las galaxias con 2pcf y DBSCAN	95
6.3. Análisis de DBSCAN con distintas distribuciones de galaxias	96
6.3.1. Modificando la cantidad de BAOs	97
6.3.2. Modificando la cantidad de galaxias sobre la circunferencia	102
6.3.3. Modificando la cantidad de galaxias y BAOs	102
6.3.4. Modificando la perturbación de galaxias sobre la circunferencia	104
6.4. Reconstrucción del BAO original con algoritmos de agrupamiento	107
6.5. Búsqueda de centros de BAO comparando DBSCAN, OPTICS y HDBSCAN	115
6.5.1. OPTICS	116
6.5.2. HDBSCAN	116
6.5.3. Análisis con distintas distribuciones de galaxias	119
7. Conclusiones y Perspectivas	123

Índice de figuras

1.	Diagrama de resumen esquemático de lo que se presenta en esta tesis. . .	14
1.1.	Modelo de la distribución de estrellas en la Vía Láctea de W. Herschel.	17
1.2.	Ejemplos de homogeneidad e isotropía.	18
1.3.	Ilustración de la geometría del espacio-tiempo.	20
1.4.	Ejemplo esquemático de como cambia la frecuencia del espectro visible de una galaxia dependiendo de si se está alejando o acercando a nosotros [Garcia, 2019].	21
1.5.	Composición del Universo.	23
1.6.	Diagrama del funcionamiento de una lente gravitacional	25
1.7.	Mapa de las anisotropías del CMB obtenidas por COBE, WMAP y Planck	27
1.8.	Ejemplos de mapas del CMB dependiendo de la curvatura del espacio-tiempo. Fuente: [Coble et al., 2018].	27
1.9.	Esquema del la historia del Universo. Fuente: Particle data group. . . .	28
1.10.	Esquema de región de sobredensidad donde hay interacción gravitacional e interacción fotón-materia [Hu, 2022].	29
1.11.	Ilustración esquemática de la medida tomada por BOSS y oscilaciones acústicas de bariones como regla estándar	30
1.12.	Función de correlación medida por BOSS.	31
1.13.	Datos recabados por distintos sondeos del Universo observable.	32
2.1.	Ejemplo de paralaje trigonométrico	34
2.2.	Ley del Inverso de los cuadrados	34
2.3.	Representación de coordenadas comóvil.	35
2.4.	Representación esquemática de la distancia de diámetro angular.	36
2.5.	Corte 2D de la simulación cosmológica de galaxias por Millennium XXL.	38
2.6.	Corte 2D de la simulación cosmológica de N-cuerpos Bolshoi.	40
2.7.	Corte 2D de la simulación cosmológica de N-cuerpos Uchuu.	40
3.1.	Campos utilizados para la estimación de la función de correlación. . . .	43
4.1.	Clasificación de distintas distribuciones de datos por distintos algoritmos de agrupamiento.	48
4.2.	Ejemplo de como algoritmos jerárquicos aglomerativo y divisivo trabajan los datos.	49
4.3.	Diferencia entre agrupamiento difuso y no difuso.	50

4.4. Ejemplos de como el algoritmo DBSCAN clasifica los datos.	52
4.5. Resultados correctos de datos de cúmulos de galaxias analizados en [Zhang, 2019].	53
4.6. 2000 datos aleatorios de la figura (4.5) analizados con DBSCAN. Fuente: [Zhang, 2019].	53
4.7. Agrupamientos con respecto a distintos parámetros de densidad.	54
4.8. Diferencia entre alcanzable por densidad (density-reachable) y densamente conectado (density-connected).	56
4.9. Diferencia entre <i>core-distance</i> y <i>Reachability-distance</i>	57
4.10. Gráfica de alcance (reachability plot).	60
4.11. Ejemplo de gráfica de alcanzabilidad mutua, árbol de expansión y árbol de expansión mínima.	62
4.12. Ejemplo de como se puede determinar los agrupamientos utilizando la función de densidad de probabilidad dependiendo del parámetro global que se utilice. Fuente: [Berba, 2020].	64
4.13. Ejemplo de cómo se ve una función de densidad de probabilidad (pdf) y cómo se vería su dendograma.	65
5.1. Diagrama de flujo del proceso para crear datos sintéticos de galaxias.	70
5.2. Dos distribuciones de 1000 datos aleatorios dentro de un Universo cuadrado de 200 unidades de longitud de lado.	71
5.3. Histograma de 1000 galaxias con distribución aleatoria contra 1000 puntos con distribución aleatoria, mostrados en la figura (5.2).	72
5.4. Función de correlación con estimador Peebles-Hauser de 1000 galaxias con distribución aleatorio contra 1000 puntos con distribución aleatoria, mostrados en la figura (5.2).	73
5.5. Funciones de correlación de una distribución aleatoria de 1000 galaxias y una distribución aleatoria de 1000 datos obtenida con 5 estimadores distintos.	73
5.6. Funciones de correlación (trasladadas) de una distribución aleatoria de 1000 galaxias y una distribución aleatoria de 1000 datos obtenida con 5 estimadores distintos.	74
5.7. Histogramas de galaxias con distribución aleatoria contra puntos con distribución aleatoria.	74
5.8. Función de correlación con estimador Peebles-Hauser de distribuciones aleatoria contra puntos con distribución aleatoria.	75
5.9. Función de correlación con distintos estimadores de distribuciones aleatorias.	76
5.10. Función de correlación con distintos estimadores de dos distribuciones aleatorias con 3000 datos.	77
5.11. Distribución esquemática de 3290 galaxias / 3290 puntos con distribución uniforme.	78
5.12. Histogramas de distintas distribuciones de galaxias con distintos radios ($R = 15, 25, 35, 45$).	79

5.13. Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para distribuciones de distintos radios ($R = 15, 25, 35, 45$).	80
5.14. Cuatro distribuciones de galaxias.	81
5.15. Cada histograma de esta figura corresponde al histograma de la distribución del respectivo cuadrante de la figura (5.14).	82
5.16. Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para distintas distribuciones de galaxias respectivas a cada cuadrante de la figura (5.14).	83
5.17. Distribuciones de galaxias con perturbaciones del 0 %, 5 % y 10 % respectivamente.	83
5.18. Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para las distribuciones de la figura (5.17) que presentan distintas perturbaciones (0 %, 5 % y 10 %).	84
5.19. Histograma de distribuciones de galaxias con distintos porcentajes (10 %, 15 %, 20 %, 25 % y 30 %) de galaxias en el centro respecto a los anillos.	84
5.20. Función de correlación de dos puntos obtenida con el estimador Peebles-Hauser para distribuciones de galaxias con distintos porcentajes (10 %, 15 %, 20 %, 25 % y 30 %) de galaxias en el centro respecto a los anillos.	85
5.21. Histograma de distribuciones de galaxias con 5 % de perturbación y distinta cantidad de galaxias por anillo (25, 35, 45, 55).	85
5.22. Función de correlación de dos puntos obtenida con el estimador Peebles-Hauser para distribuciones de galaxias con 1 % de perturbación, 10 % de galaxias en el centro respecto a las galaxias sobre el anillo y distinta cantidad de galaxias por anillo (25, 35, 45, 55).	86
5.23. Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica.	86
5.24. Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica modificando los radios.	87
5.25. Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica modificando la cantidad de galaxias en los centros.	88
5.26. Función de correlación con distintos estimadores de distintas distribuciones modificando la cantidad de anillos y galaxias por anillo.	90
5.27. Función de correlación con distintos estimadores de distintas distribuciones modificando la perturbación.	91
5.28. Detección del radio del BAO con función de correlación de dos puntos.	92
6.1. 7013 galaxias sintéticas analizadas con DBSCAN vs 7013 datos aleatorios analizadas con DBSCAN.	95
6.2. 7013 galaxias sintéticas analizadas con DBSCAN con reconstrucción de BAO.	97

6.3.	Análisis con DBSCAN de distintas distribuciones de galaxias sintéticas con 50 % de galaxias en el centro respecto a la circunferencia.	98
6.4.	Trazo de BAOs a partir de la detección de centros en la figura (6.3). . .	99
6.5.	Detección de centros de BAO con DBSCAN en distintas distribuciones de galaxias sintéticas con 30 % de galaxias en el centro respecto a la circunferencia.	100
6.6.	Detección de centros de BAO con DBSCAN en distintas distribuciones de galaxias sintéticas con 30 % de galaxias en el centro respecto a la circunferencia, con reconstrucción del patrón de circunferencias	101
6.7.	Distribuciones de distintas cantidades de galaxias sintéticas con 100 circunferencias BAO las cuales tienen 30 % de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud y perturbación de 5 %, analizados con DBSCAN.	103
6.8.	Análisis de los centroides detectados mediante DBSCAN de las distribuciones de galaxias sintéticas representadas en la figura (6.7) con reconstrucción del patrón de circunferencias de BAO.	104
6.9.	Distribuciones de distintas cantidades de galaxias sintéticas con 30 % de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud, perturbación de 1 %, analizados con DBSCAN. . .	105
6.10.	Análisis de los centroides detectados mediante DBSCAN de las distribuciones de galaxias sintéticas representadas en la figura (6.9) con reconstrucción del patrón de circunferencias de BAO.	106
6.11.	Distribuciones de distintas cantidades de galaxias sintéticas con máximo 40 galaxias sobre la circunferencia del BAO y 30 % de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud, analizados con DBSCAN.	107
6.12.	Análisis de los centroides detectados mediante DBSCAN de las distribuciones de galaxias sintéticas representadas en la figura (6.11) con reconstrucción del patrón de circunferencias de BAO.	108
6.13.	Reconstrucción del BAO original con algoritmos de agrupamiento para 30 BAOs sin perturbación.	110
6.14.	Reconstrucción del BAO original con algoritmos de agrupamiento para 30 BAOs con 10 % de perturbación.	112
6.15.	Reconstrucción del BAO original con algoritmos de agrupamiento para 100 BAOs con 10 % de perturbación.	114
6.16.	Galaxias sintéticas distribuidas en 10 circunferencias BAO con máximo 5 galaxias sintéticas cada una con perturbación del 5 % y 800 % de galaxias sintéticas en el centro del BAO respecto a la circunferencia.	115
6.17.	Gráfica de alcanzabilidad de la distribución de la figura (6.16), así como la detección de agrupamientos con los algoritmos OPTICS y DBSCAN con distintos radios eps.	117
6.21.	Dendrograma simplificado de la distribución de la figura (6.16).	117

6.18. Análisis de la detección de agrupamientos con el algoritmo OPTICS y la gráfica de alcanzabilidad de la distribución de la figura (6.16).	118
6.19. Gráfica de alcanzabilidad de la distribución de la figura (6.16), así como la detección de agrupamientos con los algoritmos OPTICS (agregando el parámetro max_eps) y DBSCAN.	119
6.20. Árbol de expansión mínima de la distribución y dendrograma de la distribución.	120
6.22. Análisis de los centroides detectados mediante DBSCAN de las distribuciones de galaxias sintéticas representadas en la figura (6.9).	121
6.23. Gráfica de alcanzabilidad de la distribución de la figura (6.16), así como la detección de agrupamientos con los algoritmos OPTICS y DBSCAN con distintos radios y su reconstrucción del BAO original.	122

Resumen

La distribución de galaxias en el Universo es una fuente valiosa de información para la cosmología moderna, ya que a partir de ella podemos extraer diversas propiedades como la formación de estructura a gran escala, la evolución y contenido de materia-energía oscura.

Las oscilaciones acústicas de bariones (BAO, por su acrónimo en inglés: Baryon Acoustic Oscillations) son patrones en la densidad generados por las interacciones fotón-materia, poco después del Bing Bang, los cuales se pueden ver como una lucha entre la materia que por gravedad atrae hacia las regiones más densas y los fotones que ejercían presión al exterior. Esta lucha entre la atracción gravitatoria y la presión de radiación generó ondas esféricas. Al momento en que la materia y los fotones se desacoplaron, éstas ondas esféricas se congelaron y ahora su expansión sólo depende de la expansión del Universo. Es por eso que la sobredensidad de materia queda distribuida alrededor de "cascarones esféricos", es decir, la mayoría de galaxias se agrupan en promedio alrededor de estas esferas, que además contienen ligeras perturbaciones debido a atracciones gravitatorias, velocidades peculiares, entre otros. Por otro lado, también existe una sobredensidad o pico de materia en los centros de los cascarones debido a que la materia oscura no interacciona con la luz y por su fuerte atracción gravitatoria se formaron también galaxias en los centros. Como consecuencia, existe una distancia o escala característica de BAO la cual nos indica que partiendo de una galaxia es más probable encontrar otra galaxia con una separación dada por esta distancia específica.

El presente trabajo tiene como principal objetivo estudiar esta escala característica mediante el uso de diferentes estimadores de la función de correlación de dos puntos. Para este análisis, se empleó una simulación de galaxias simplificada en 2 dimensiones representando un corte del Universo, suponiendo que tiene curvatura cero, se consideró que las galaxias son puntuales y se agrupan en promedio alrededor de

circunferencias/anillos, cuyo radio esta dado por la escala característica. Se analizaron los estimadores: **Davis-Peebles**, **Hamilton**, **Hewett**, **Landy-Szalay** y **Peebles-Hauser**; entre los cuales se observó que **Peebles-Hauser** y **Hewett** demostraron un desempeño superior. A partir de estos resultados se exploró la posibilidad de que dichos patrones puedan ser identificados con técnicas de machine learning, en particular con el algoritmo de agrupamiento DBSCAN, y con ayuda de la escala característica reconstruir la posición original de las galaxias que se han desplazado de la onda BAO congelada.

Analizando los datos de galaxias sintéticas obtenidos con funciones de correlación se observó que entre mayor cantidad de galaxias por anillo de BAO se consideren mayor amplitud presentaran los picos correspondientes al radio y diámetro; mientras que entre mayor cantidad de anillos de BAO haya, más tenue será el pico de BAO. A medida que aumenta la magnitud de la perturbación, se observa una disminución en la amplitud del pico de BAO, lo que implica una disminución en la prominencia del pico (es decir, un achatamiento del mismo). Conforme disminuye el porcentaje de galaxias en el centro, se observa una reducción tanto en el ancho como en la altura del pico del radio, mientras que se aprecia una disminución en el ancho y un aumento en la altura del pico del diámetro. Se estudió la posibilidad de utilizar una función geométrica para el conteo de pares aleatorios en lugar de hacerlo por el método de integración Monte Carlo, concluimos que, para escalas distantes de cero, existe una similitud considerable entre ambos métodos. Por consiguiente, utilizar la función geométrica se revela como una opción altamente eficiente, sin embargo sólo nos sirve cuando conocemos la forma de nuestro espacio. Finalmente, se propone como mejor estimador a **Hewett** debido a que en diversas circunstancias ha sido el segundo que más marcado muestra el pico BAO, aunado a que es junto con **Hamilton** de los que menos ruido presenta para grandes distancias.

Por otro lado, en todas las distribuciones consideradas, a pesar de haber incluido una mayor cantidad de datos aleatorios, se logró una distinción efectiva entre los datos aleatorios y los datos de galaxias utilizando DBSCAN. Se detectó la mayoría de los centros, lo que permitió una reconstrucción exitosa del BAO original con la ayuda de la función de correlación de dos puntos. Con DBSCAN obtuvimos los siguientes resultados. A medida que se incrementa la cantidad de galaxias sintéticas consideradas, se incrementará la probabilidad de que el algoritmo detecte como agrupamientos el traslape de galaxias de diferentes circunferencias de BAO. Incluso, para un número

elevado de galaxias el algoritmo comienza a considerar múltiples centros dentro de un mismo agrupamiento. Aunado a que cuanto mayor sea la cantidad de galaxias presentes sobre el BAO, más sencillo será para DBSCAN detectar correctamente los centroides, debido a que aumentara la densidad de galaxias en los centroides. Por lo tanto, llegamos a la conclusión de que es posible reconstruir el BAO utilizando la función de correlación de dos puntos y modelos de agrupamiento basados en densidad, especialmente el algoritmo DBSCAN.

A continuación se presenta lo que se abordará en cada capítulo se sugiere observar la figura 1 para facilitar la comprensión del trabajo.

En el capítulo 1, se presenta una breve introducción de la cosmología, se introduce al Fondo Cósmico de Microondas y se explica el origen de las oscilaciones acústicas de bariones, por ejemplo, cómo las BAO originan una sobredensidad de materia formando la estructura a gran escala o filamentos y por lo tanto galaxias. También, se detalla como las oscilaciones acústicas de bariones se usan para medir distancias y su relación con la materia oscura. En el capítulo 2, se abordan los diversos tipos de distancias que se utilizan en cosmología junto con las reglas estándar o medidas de referencia en el Universo. De manera similar, se presentan algunas simulaciones como *Millenium*, *Bolshoi*, *Uchuu* y su relación con los sondeos. En el capítulo 3, se muestra qué es una función de correlación de dos puntos y diferentes estimadores para calcularla. Posteriormente, en el capítulo 4, se presenta el aprendizaje automático, los algoritmos de agrupamiento y específicamente nos enfocamos en el funcionamiento de DBSCAN. Más adelante, capítulo 5, se simulan diferentes patrones (número de galaxias y ruido) para ver cómo cambia el pico de BAO en la función de correlación. En el capítulo 6, se simulan diferentes patrones nuevamente, se analiza la detección de centros de BAO con DBSCAN y cómo se podría lograr una reconstrucción de las posiciones iniciales. Así mismo se realiza un pequeño análisis con OPTICS y HDBSCAN. Finalmente, en el capítulo 7, se otorgan las conclusiones del trabajo.

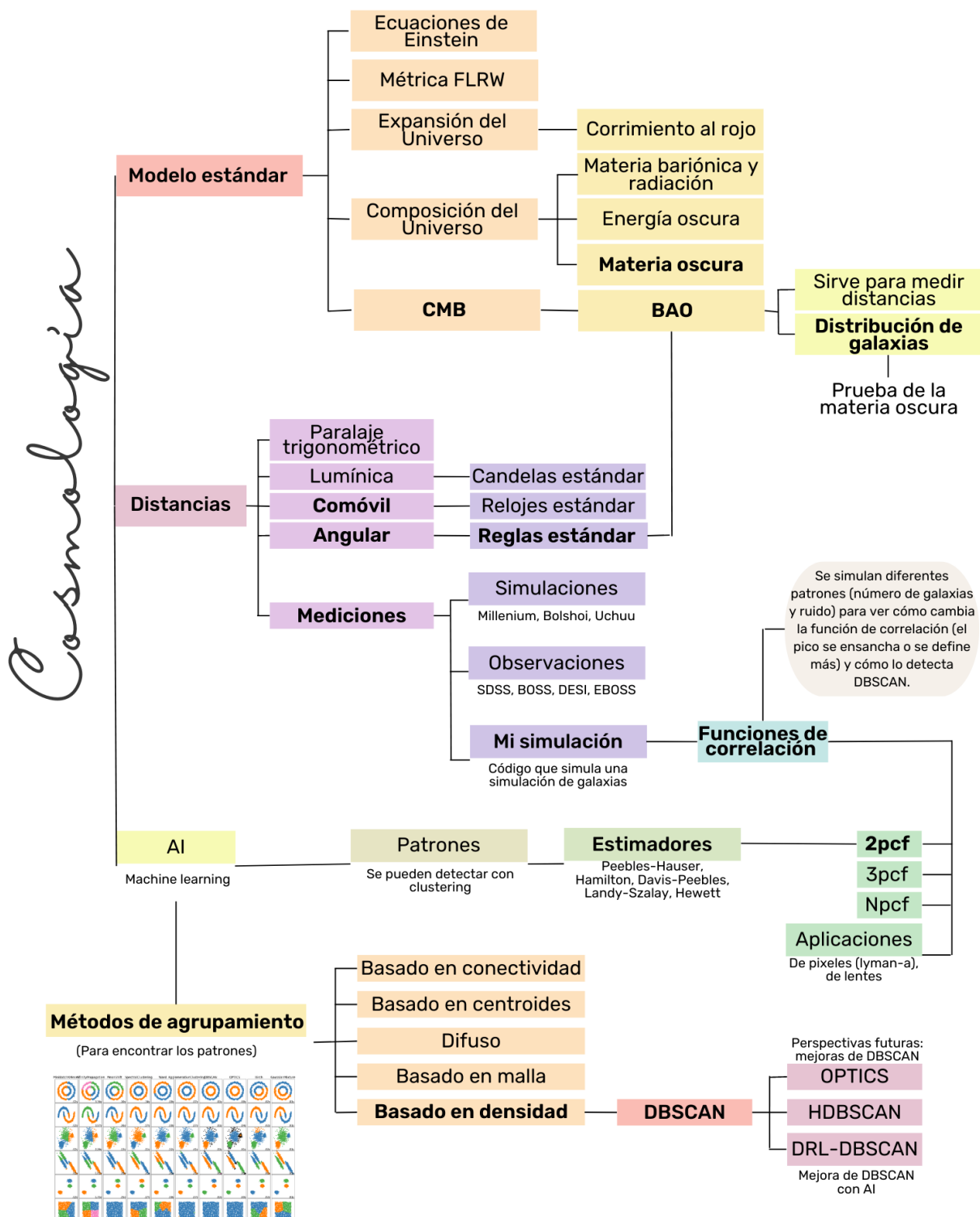


Figura 1: Diagrama de resumen esquemático de lo que se presenta en esta tesis.

Capítulo 1

Marco teórico

1.1. BREVE HISTORIA DE LA COSMOLOGÍA

Comprendiendo la importancia histórica y conceptual de la cosmología, en el presente capítulo nos basaremos en las siguientes citas para respaldar y enriquecer nuestro estudio: [Liddle, 2003, Téllez Tovar, 2018, Chacón Lavanderos, 2018].

Comenzaremos entendiendo las raíces etimológicas de la palabra “cosmología”. Ésta viene del griego ‘*kosmos*’ que significa Universo, orden, cosmos; la palabra ‘*logos*’ que significa tratado, estudio y el sufijo ‘*ía*’ que significa cualidad. Por lo que cosmología es la rama de la física que estudia al Universo como un conjunto, a diferencia de la astronomía que se concentra en objetos celestes individualmente. Desde la antigüedad, el Universo ha sido objeto de estudio de la humanidad, en sus inicios fue estudiado gracias a la luz visible que emiten ciertos cuerpos, en particular las estrellas. La agrupación de estrellas y cuerpos celestes en una determinada región se conoce como galaxia, mientras que las agrupaciones de galaxias se conocen como cúmulos.

Buscando comprender y explicar el comportamiento del Universo, Ptolomeo propuso el modelo geocéntrico, pero anteriormente Platón y Aristoteles, entre otros, ya habían formulado sus propias teorías geocéntricas. En 1543, Copérnico, siguiendo los estudios de Aristarco de Samos, formuló la teoría heliocéntrica del sistema solar, él pensaba que el Sol era el centro del Universo. Más tarde, observaciones astronómicas de Galileo concordaron con la teoría heliocéntrica. Posteriormente, Kepler, basándose en las observaciones y datos de Tycho Brahe, propuso que los planetas se movían alrededor del Sol en órbitas elípticas, considerando al Sol como un foco de la elipse y formuló las tres leyes que llevan su nombre para describir matemáticamente el movimiento de los planetas. En 1687, Newton publica sus leyes que explican el movimiento de los cuerpos [Newton, 1687]. Posteriormente, observaciones astronómicas permitieron comprender que las estrellas no están distribuidas uniformemente y que las más cercanas estaban dentro de nuestra galaxia, la Vía Láctea. Ésta contiene aproximadamente 10^{11} estrellas y está conformada por un bulbo central y un disco cuyo radio es de $12.5kpc$ y espesor de $0.3kpc$. En la definición 1.1.1 presentaremos

diferentes unidades de longitud cosmológicas que nos permiten comprender y medir las vastas dimensiones del Universo. Estas unidades son fundamentales para contextualizar la estructura y escala de objetos astronómicos, así como desempeñar un papel fundamental en el campo de la cosmología.

Def 1.1.1: Unidades de longitud cosmológicas

- **Unidades astronómicas (ua)**

Es la distancia promedio de la órbita de la Tierra al Sol.

- **Años luz (ly)**

Es la distancia que recorre un fotón en un año (velocidad de la luz: $c = 299,792,458 \frac{m}{s}$).

- **Parsecs (pc)**

Es la distancia a la que una unidad astronómica subtende un ángulo de un segundo de arco, ésta se basa en el método de la paralaje trigonométrico que se analizará más adelante. Se utiliza cuando las distancias son de más de miles de años luz. Para distancias entre galaxias se utilizan Megaparsecs (Mpc).

$$1pc = 3.2616ly = 206265ua = 3.0857 \times 10^{16}m. \quad (1.1)$$

Alrededor de 1785 el astrónomo Herschel, dividió el cielo en más de 600 zonas y con ayuda de su telescopio contó las estrellas que veía en cada zona, entre más tenue se veían, más lejos consideraba que estaban (lo cual no es correcto debido a que el espacio está lleno de polvo y esto afecta como percibimos el brillo de las estrellas) y de esta forma calculó como debía ser la estructura de la galaxia. Debido a que obtuvo aproximadamente el mismo número de estrellas en cualquier dirección, supuso que debíamos estar en el centro del Universo, como se muestra en la figura (1.1) para el modelo de distribución de estrellas de la Vía Láctea que obtuvo.

En 1915, Einstein publica la Teoría de la Relatividad General, una teoría de gravitación diferente a la propuesta por Newton donde relaciona la materia, el espacio y el tiempo. En 1917, el astrónomo Shapley estudió los cúmulos globulares, es decir, agrupaciones circulares de estrellas. También asumió que entre más lejos, más tenues serían. Al graficarlos observó que formaban una esfera y estaban en una parte específica del cielo, concluyendo que el centro de esa esfera debía ser el centro de la galaxia, el cuál estaba muy lejos de nuestro sistema, sin embargo, siguió creyendo que la Vía Láctea era el centro del Universo. Luego, en 1924 Hubble mide la distancia entre la Tierra y la galaxia Andrómeda, demostrando que debe estar fuera de la Vía Láctea y que hay más galaxias [Hubble, 1926]. Esto nos dejó claro que no somos el centro del Universo y que no estamos en un lugar privilegiado, lo cual nos guía al principio Copernicano y al principio cosmológico.

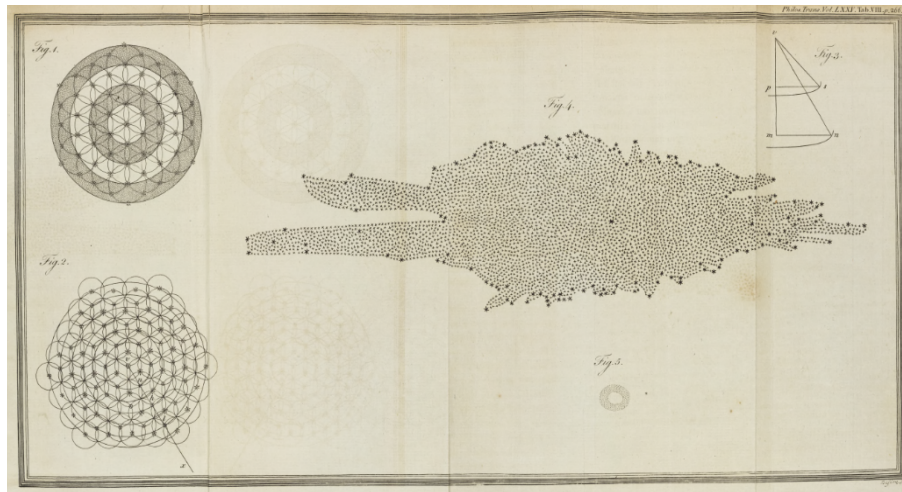


Figura 1.1: Modelo de la distribución de estrellas en la Vía Láctea de William Herschel [Herschel, 1785].

1.1.1. Principio cosmológico

El Principio Copernicano establece que *no ocupamos un lugar privilegiado en el Universo*. Por otra parte, al observar el Fondo Cósmico de Microondas (figura 1.7) o la distribución de galaxias a gran escala (figura 2.5) podemos apreciar que el Universo es isotrópico a gran escala. Ahora si recordamos el principio copernicano y la isotropía podemos asumir que el Universo es homogéneo. En el presente trabajo se considerará como cierto el principio cosmológico, el cual establece que *a gran escala, es decir $\approx 100\text{Mpc}$, el Universo es homogéneo e isotrópico*. Es homogéneo porque la materia está uniformemente distribuida, todos sus puntos se ven igual y es isotrópico debido a que sin importar la dirección en la que se observe se verán las mismas propiedades. Es muy importante comprender que es estadísticamente y a gran escala debido a que sabemos perfectamente que el Universo no es igual, por ejemplo, si estás parado en la Facultad de Ciencias en Ciudad Universitaria a si estas parado en la torre Eiffel, tampoco es lo mismo observar el interior del Sol que estar parado en el planeta Vinland ubicado en la galaxia Andrómeda. Sin embargo a grandes escalas que son del orden de 100Mpc , sí se cumple la homogeneidad e isotropía. Finalmente para dimensionar 100Mpc , recordemos que el diámetro de la Vía Láctea es de alrededor de 0.025Mpc .

En la figura (1.2) se muestran ejemplos de homogeneidad e isotropía para un mejor entendimiento. El primer círculo de la figura (1.2) es homogéneo porque no importa en que punto te coloques, la materia se ve uniformemente distribuida a lo largo de líneas diagonales pero no es isotrópico porque dependiendo de la dirección ves propiedades distintas (ya sea líneas paralelas, perpendiculares o diagonales). El segundo círculo es isotrópico si te colocas en el centro, ya que sin importar la dirección, se observan las mismas propiedades, sin embargo no es homogéneo porque la materia no se ve uniformemente distribuida en los puntos exteriores. Finalmente, el último círculo es homogéneo e isotrópico porque sin importar la posición o la dirección, se observaran las misma propiedades y la materia uniformemente distribuida.

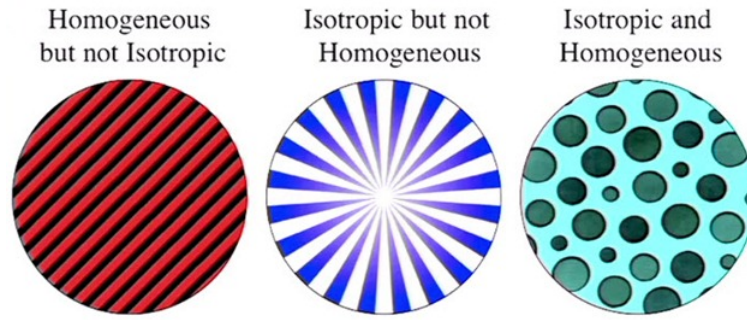


Figura 1.2: Ejemplos de homogeneidad e isotropía [Téllez Tovar, 2018].

1.2. MODELO ESTÁNDAR (Λ CDM)

Actualmente existen diferentes modelos que buscan explicar el comportamiento del Universo. En este trabajo sólo consideraremos el modelo estándar (Λ CDM, por sus siglas en inglés "Lambda-Cold Dark Matter") que describe la estructura y evolución a gran escala del Universo y asume que la teoría de Relatividad General es correcta a escalas cosmológicas. Es considerado el modelo estándar de la cosmología porque es el que proporciona una explicación relativamente sencilla para observaciones astronómicas actuales como el fondo cósmico de microondas, la distribución de galaxias y la expansión acelerada del Universo.

Por tanto, los pilares del modelo estándar de la cosmología, son: el principio cosmológico y la relatividad general. Éste modelo considera la inflación como una fase primordial de expansión acelerada, la predominante existencia de energía oscura (Λ) y materia oscura fría en el Universo; como lo veremos más adelante.

1.2.1. Ecuaciones de Einstein

A finales de 1915, Einstein presentó su ecuación de campo, la cual relaciona la materia con la geometría o curvatura del espacio-tiempo, de tal forma que la gravedad es un resultado de como la materia curva el espacio-tiempo y al mismo tiempo la geometría de éste determina cómo se mueve la materia. La teoría de la relatividad utiliza un espacio-tiempo curvo pseudoriemanniano con la siguiente métrica (que define las coordenadas del espacio-tiempo):

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \quad (1.2)$$

donde ds^2 es el elemento de línea que nos indica la distancia entre dos puntos en el espacio-tiempo y $g_{\mu\nu}$ es el tensor métrico, simétrico, que contiene toda la información del espacio-tiempo. El término ds^2 definirá los símbolos de Christoffel $\Gamma_{\mu\nu}^\alpha$, estos guardan la relación de cómo los ejes del espacio-tiempo se modifican en cada dirección y lo definiremos a partir del tensor métrico de la siguiente forma:

$$\Gamma_{\mu\nu}^{\alpha} = \frac{1}{2}g^{\alpha\beta} (g_{\nu\beta,\mu} + g_{\beta\mu,\nu} - g_{\nu\mu,\beta}). \quad (1.3)$$

Las ecuaciones de Einstein en presencia de materia y energía son:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{8\pi G}{c^4}T_{\mu\nu}, \quad (1.4)$$

donde $T_{\mu\nu}$ es el tensor de energía-momento que contiene la densidad y flujo de la energía y el momento de la materia, c es la velocidad de la luz, G es la constante de la gravitación universal y $G_{\mu\nu}$ es el tensor de curvatura de Einstein:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu}, \quad (1.5)$$

el cual es simétrico y tiene cero divergencia, esto es, satisface las identidades de Bianchi; donde $R_{\mu\nu}$ es el tensor de Ricci $R_{\mu\nu} \equiv R_{\mu\alpha\nu}^{\alpha}$, definido como la traza del tensor de Riemann $R_{\beta\mu\nu}^{\alpha}$, el cual describe como la curvatura cambia y está definido por los símbolos de Christoffel y sus derivadas:

$$R_{\beta\mu\nu}^{\alpha} = \Gamma_{\beta\nu,\mu}^{\alpha} - \Gamma_{\beta\mu,\nu}^{\alpha} + \Gamma_{\beta\nu}^{\sigma}\Gamma_{\sigma\mu}^{\alpha} - \Gamma_{\beta\mu}^{\sigma}\Gamma_{\sigma\nu}^{\alpha}. \quad (1.6)$$

R es el escalar de curvatura de Ricci, que es la contracción del tensor de Ricci $R \equiv R_{\mu}^{\mu}$, Λ es la constante cosmológica³. $G_{\mu\nu}$ es un tensor simétrico 4×4 , por lo que tiene diez componentes independientes, dando como resultado un conjunto de 10 ecuaciones diferenciales no lineales y acopladas.

1.2.2. Métrica de Friedman-Lemaître-Robertson-Walker (FLRW)

La métrica FLRW es una solución exacta de las ecuaciones de campo de Einstein. Ésta supone la homogeneidad e isotropía (es decir que es invariante ante traslaciones y rotaciones), que la parte espacial es independiente de la temporal, considera la relatividad especial, los efectos gravitacionales y describe la geometría y expansión del Universo a través del tiempo. Se define de la siguiente forma:

$$ds^2 = -c^2 dt^2 + a(t)^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right), \quad (1.7)$$

¹La coma representa la derivada respecto a las coordenadas, por ejemplo, $g_{\nu\beta,\mu} = \frac{\partial g_{\nu\beta}}{\partial x^{\mu}}$.

²En [Hirvonen, 2023] se presentan dos formas de derivar las ecuaciones de campo de Einstein.

³Propuesta en 1917 por Einstein para modelar un Universo homogéneo estático con simetría esférica y así contrarrestar el colapso que se produciría debido a la atracción gravitatoria. Luego en 1929 se midió el corrimiento al rojo de las galaxias y descubrió que el Universo se está expandiendo, lo cual llevó a Einstein a pensar que era un error haber agregado el término de la constante cosmológica. Después en 1998, las observaciones de supernovas demostraron que el Universo no sólo se estaba expandiendo, sino que también lo hacía aceleradamente debido a un fenómeno desconocido al que se le llamó "energía oscura" y para representarlo se retomó la constante cosmológica.

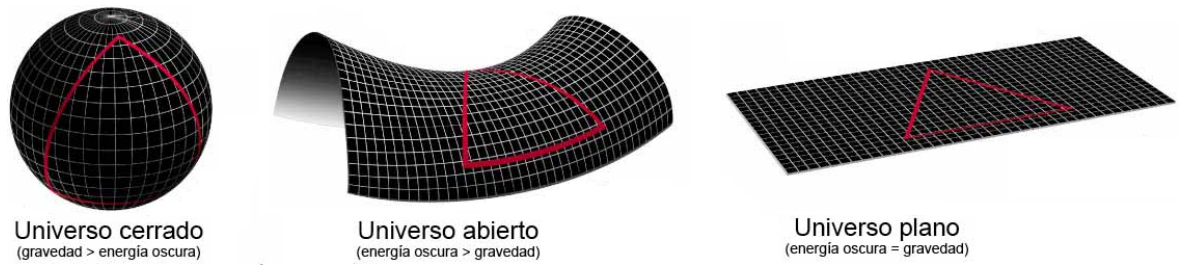


Figura 1.3: Ilustración de la geometría del espacio-tiempo tomada de [NASA, 2014].

donde $a(t)$ es el **factor de escala**, y k es la constante en el tiempo que describe la curvatura de la siguiente forma,

$$k = \begin{cases} -1 \longrightarrow \text{Universo cerrado,} \\ 0 \longrightarrow \text{Universo plano,} \\ 1 \longrightarrow \text{Universo abierto.} \end{cases} \quad (1.8)$$

En el caso de un Universo plano se tiene un espacio Euclídeo. A continuación, la figura (1.3) es una ilustración de cómo se vería en espacio-tiempo dependiendo de la curvatura que presente, es importante recordar que el espacio tiempo ocurre en cuatro dimensiones por lo que esta representación no es exacta.

Algo importante a mencionar es que, para resolver las ecuaciones de campo de Einstein, suponiendo la homogeneidad e isotropía es necesario que el tensor de energía–momento también lo sea, un ejemplo general es el de un **fluido perfecto**, el cual está dado en términos de su densidad, presión/tensión y cuadrivelocidad de la siguiente forma [Schutz, 2009]:

$$T_{\mu\nu} = \left(\rho + \frac{p}{c^2} \right) u_\mu u_\nu + p g_{\mu\nu}, \quad (1.9)$$

donde si $p > 0$ se conoce como presión y si $p < 0$ como la tensión. La densidad y presión son independientes de la posición (no del tiempo) debido a la homogeneidad. Para la métrica FLRW (1.7), y un observador comovil $u_\mu = (1, 0, 0, 0)$, el tensor métrico es diagonal,

$$T_{\mu\nu} = \text{diag}, (-\rho, p, p, p). \quad (1.10)$$

Por otro lado, las partículas se mueven a lo largo del espacio-tiempo sobre trayectorias **geodésicas** que son las curvas que unen dos puntos con mínima longitud en cierta superficie. La ecuación que describe estas geodésicas está dada por:

$$\frac{d^2 x^\sigma}{ds^2} + \Gamma_{\alpha\mu}^\sigma \frac{dx^\alpha}{ds} \frac{dx^\mu}{ds} = 0, \quad (1.11)$$

donde x^σ son las coordenadas y ds es el intervalo. Por ejemplo, las geodésicas en el espacio Euclídeo son las líneas rectas, sin embargo al ver en un mapa plano la

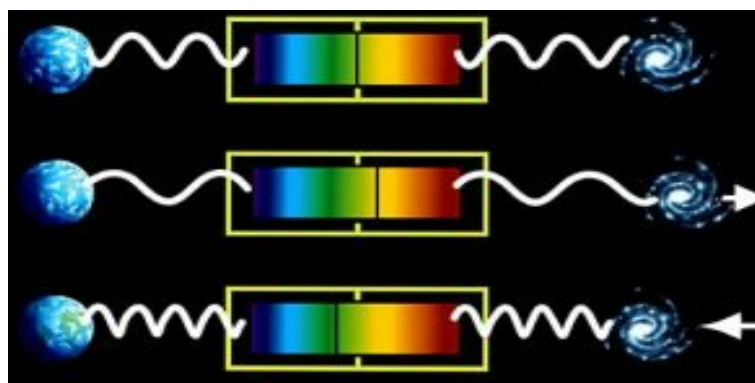


Figura 1.4: Ejemplo esquemático de como cambia la frecuencia del espectro visible de una galaxia dependiendo de si se está alejando o acercando a nosotros [García, 2019].

trayectoria geodésica que siguen los aviones podemos ver líneas curvas.

1.2.3. Expansión del Universo

La métrica FLRW describe un Universo dinámico (que puede expandirse o contraerse) y para entender cómo se descubrió que el Universo se expande debemos comprender el fenómeno de **corrimiento al rojo**.

Comencemos recordando el **efecto Doppler** para ondas sonoras, el cual es el cambio aparente en la frecuencia de la onda debido a un movimiento relativo de la fuente o del receptor, tomando en cuenta tanto la velocidad del receptor y de la fuente respecto al aire que es el medio por el que se propagan [Tippens and Ruiz, 2006]. Ahora, el corrimiento al rojo es el efecto Doppler aplicado a las ondas de luz, sabemos que la luz a diferencia del sonido se puede propagar en el vacío, por lo que sólo toma en cuenta la velocidad relativa entre la fuente y el receptor. Cada átomo tiene un espectro de emisión único con frecuencias características que conocemos. Los astrónomos observaron que, al obtener los espectros de emisión y absorción de galaxias, estrellas, entre otros cuerpos celestes y comparar los datos con los espectros de átomos conocidos estos se presentaban desplazados, esto quiere decir que el objeto se está alejando o acercando respecto a nosotros, ya que si estuviera en reposo respecto a nosotros se observaría una frecuencia correcta, en la figura (1.4) se muestra un ejemplo esquemático de como se modifica la frecuencia del espectro visible de una galaxia en función de si se aleja o se acerca a nosotros. Es importante reconocer que hablamos de movimientos radiales respecto a nosotros [Halliday, 2009].

Si se mueve hacia nosotros, las ondas de luz se juntan, elevando la frecuencia y como el azul está en el extremo de alta frecuencia del espectro visible, llamamos a este efecto como **desplazamiento al azul**. Ahora, si se aleja, su frecuencia disminuye y como el rojo está en el extremo de baja frecuencia del espectro visible, el efecto se conoce como **corrimiento al rojo**. Los astrónomos notaron que casi todas las galaxias y

estrellas presentan corrimiento al rojo, es decir se están alejando de nosotros.

Sean λ_{obs} y λ_{emit} las longitudes de onda de la luz de los puntos de observación y emisión respectivamente, el corrimiento al rojo z está definido por [Liddle, 2015]:

$$1 + z = \frac{\lambda_{obs}}{\lambda_{emit}} = \frac{a_0}{a(t)} = \frac{1}{a(t)}. \quad (1.12)$$

donde $a_0 = a(t_0)$ se iguala a 1 en el presente, “ $(1 + z)$ indica cuánto se ha expandido el Universo desde que la luz fue emitida” [Chacón Lavanderos, 2018]. Para galaxias relativamente cercanas, se puede obtener una aproximación de la velocidad a la que va una galaxia con la siguiente relación:

$$z \approx \frac{\bar{v}}{c}. \quad (1.13)$$

En 1929, comparando las velocidades de recesión⁴ de las galaxias obtenidas con la ecuación (1.13) y las distancias, Edwin Hubble observó que entre más lejos están la galaxias de la Tierra más rápido se alejan, esto se conoce como **Ley de Hubble** y está definida por [Liddle, 2015]:

$$\bar{v} = H_0 \bar{r}, \quad (1.14)$$

donde \bar{v} es la velocidad aparente, H_0 es la constante de Hubble, y \bar{r} es la distancia propia, que sería la distancia medida entre la Tierra y la galaxia a un cierto tiempo t . Esta ley se cumple a escalas cercanas, casi local, para grandes escalas las teorías predicen que $H = H(t)$ cambia respecto al tiempo, y es por esto que al parámetro de Hubble en el tiempo actual t_0 se le llama constante de Hubble. El parámetro de Hubble se define como:

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)}, \quad (1.15)$$

donde punto significa derivada respecto al tiempo propio. Distintas estimaciones de la constante de Hubble se han realizado. El Hubble Space Telescope Key Project obtuvo la distancia a 31 galaxias relacionando el periodo y luminosidad de estrellas Cefeidas con un error estadístico y un error sistemático [Freedman et al., 2001]

$$H_0 = 72 \pm 3 \pm 7 \text{ Km s}^{-1} \text{ Mpc}^{-1}. \quad (1.16)$$

En 2013, se publicó el siguiente valor de la constante de Hubble estimado por el Telescopio Espacial Hubble observando más de 600 Cefeidas [Riess et al., 2016]:

$$H_0 = 73.24 \pm 1.74 \text{ Km s}^{-1} \text{ Mpc}^{-1}. \quad (1.17)$$

En 2018, la colaboración Planck realizó una estimación de la constante de Hubble [Ade

⁴Velocidad con que un objeto se aleja de nosotros, independientemente de si se mueve en el mismo plano o no.

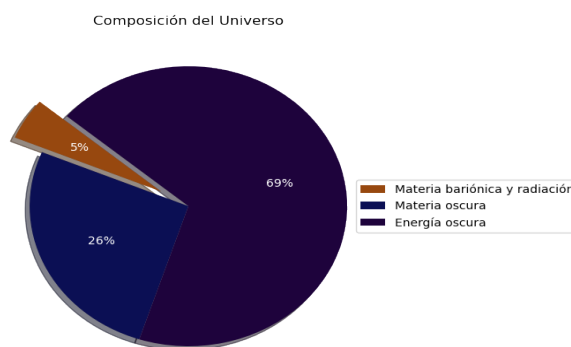


Figura 1.5: *Composición del Universo.*

et al., 2014]:

$$H_0 = 67.4 \pm 0.5 \text{ Km s}^{-1} \text{ Mpc}^{-1}. \quad (1.18)$$

En 2020 se publicó el valor de la constante de Hubble obtenido con eBOSS usando el modelo Λ CDM [Alam et al., 2021]:

$$H_0 = 67.35 \pm 0.97 \text{ Km s}^{-1} \text{ Mpc}^{-1}. \quad (1.19)$$

Es importante señalar que la medición exacta de la constante de Hubble sigue siendo objeto de investigación. Los valores cercanos a 73 se obtienen analizando el Universo local, mientras que los valores cercanos a a 67 se obtienen analizando el Universo primitivo. Esta discrepancia se conoce como tensión de Hubble debido a que algunos científicos argumentan que se necesitan más datos para analizar el Universo local a parte de efectos gravitacionales mientras que otros científicos argumentan que el modelo estándar podría estar mal por lo que no se está analizando adecuadamente el Universo primitivo.

1.3. COMPOSICIÓN DEL UNIVERSO

Sabemos que la constante de Hubble, H_0 , es un parámetro que describe la tasa a la que el Universo está expandiéndose. La composición del Universo, por otro lado, se refiere a la cantidad de materia y energía que compone el Universo. Ambas se relacionan debido a que la tasa de expansión está determinada por la cantidad de materia y energía en el Universo. Por ejemplo, si el Universo contiene una gran cantidad de materia oscura, la tasa de expansión será más lenta que si contuviera solo una cantidad mínima de materia oscura.

Como podemos observar en la figura (1.5) sólo el 4.9 % del Universo está compuesto de materia bariónica y radiación, el resto es desconocido y sólo podemos medirlo indirectamente resultando en aproximadamente 26.2 % de materia oscura y 68.9 % de energía oscura [Vázquez González, 2008, Aghanim et al., 2020].

1.3.1. Materia bariónica y radiación

En cosmología nos referimos como materia bariónica a la materia compuesta de bariones (partículas subatómicas que se forman con 3 quarks, ejemplo de estas son los protones y neutrones) y leptones (como electrones y otras partículas no relativistas, salvo por algunos tipos de neutrinos), en otras palabras es la materia que compone los objetos visibles en el Universo, como las estrellas, planetas, galaxias, gas intergaláctico y nosotros mismos. Por otra parte, ejemplos de la radiación son las partículas ultra relativistas como fotones, neutrinos, entre otros.

1.3.2. Energía oscura

De la energía oscura se sabe muy poco, actúa a gran escala y se propuso su existencia cuando se descubrió que el Universo se expande aceleradamente, lo que indica que la energía oscura tiene un comportamiento gravitacionalmente repulsivo [Escamilla Torres, 2018]. Conocer la tasa de expansión del Universo nos ayuda a restringir lo que conocemos de la energía oscura.

1.3.3. Materia oscura

La primera vez que se planteó el término de materia oscura fue cuando el astrónomo Fritz Zwicky [Zwicky, 1937] estudiaba la dinámica interna del cúmulo de galaxias Coma Berenice, él calculó la masa teniendo un aproximado de cuántas estrellas tenía cada galaxia y su brillo [Vázquez González, 2008]. Posteriormente, con ayuda del teorema virial, relacionó la masa del conjunto de cuerpos unidos gravitacionalmente con sus velocidades. De esta manera obtuvo una masa mucho mayor a la que había calculado antes, debía haber algo más que ejercía una gran fuerza gravitacional impidiendo la expansión del cúmulo [Rodríguez, 2010]. A pesar de que otros astrónomos habían notado discrepancias entre las observaciones y sus cálculos, Zwicky fue el primero en plantear la existencia de una materia desconocida que no interactúa con la luz pero si interactúa gravitacionalmente con la materia bariónica.

Luego en 1970, la astrónoma Vera Rubin en el siguiente artículo [Rubin and Ford, 1970] estimó la velocidad de rotación de las estrellas en función de su distancia al centro de la galaxia y notó que ésta no disminuye conforme se alejan del centro de la galaxia, lo cual no era lo esperado (se esperaba que, por las leyes de Kepler, entre más lejos del centro gravitacional se encuentre, su velocidad debería disminuir). Este problema se puede justificar considerando la existencia de materia oscura [Vázquez González, 2008].

La dinámica y distribución de los cuerpos celestes es una forma de medir la materia oscura, otra forma es a través de lentes gravitacionales débiles. La teoría de la relatividad general de Einstein nos dice que las masas deforman el espacio-tiempo que las rodea. Por lo que entre más masivo sea un cuerpo, más deformará el espacio-tiempo a su alrededor y por lo tanto desviará la trayectoria de lo que pase cerca, incluso la de los fotones. Las lentes gravitacionales son el efecto de desviar la luz de un objeto

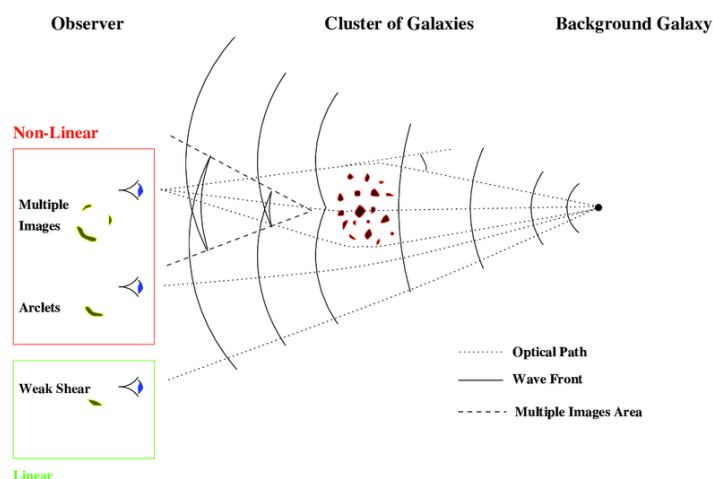


Figura 1.6: Diagrama del funcionamiento de lentes gravitacionales débiles y fuertes, donde un objeto masivo, en éste caso un agrupamiento de galaxias (rojo) desvía la luz emitida por un objeto fuente por lo que el observador ve imágenes virtuales; tomado de [Kneib and Natarajan, 2012].

detrás de esta, como se representa en la figura (1.6), pueden ocurrir como resultado de cualquier objeto masivo sin recurrir a la materia oscura. El resultado que se puede observar depende de las distancias entre la fuente de luz, el observador y el deflector, la alineación observador-deflector-fuente y las propiedades del deflector. Al observar las desviaciones por lentes gravitacionales débiles inferimos la existencia de la materia oscura [Vázquez González, 2008].

Actualmente se concluye que la materia oscura no interactúa con la materia bariónica ni la radiación, presenta un comportamiento gravitacionalmente atractivo, no tiene carga, estuvo en el Universo temprano y ayuda a explicar la estabilidad de las galaxias. Existen muchos modelos que buscan explicar las observaciones, en el artículo [Vázquez González, 2008] se mencionan los siguientes modelos:

- Materia oscura fría (CDM) [Armendariz-Picon and Neelakanta, 2014].
- Materia oscura auto-interactuante (SIDM) [Spergel and Steinhardt, 2000, Carlson et al., 1992].
- Materia oscura tibia (WDM) [Colin et al., 2000, Bode et al., 2001].
- Materia oscura repulsiva (RDM) [Goodman, 2000].
- Materia oscura difusa (FDM) [Hu et al., 2000].
- Materia oscura auto-aniquilante (SADM) [Kaplinghat et al., 2000].
- Materia oscura que decae (DDM) [Cen, 2001].

En este caso sólo mencionaremos el modelo de **materia oscura fría (CDM)**, acrónimo de Cold Dark Matter, en el que las partículas son frías (no se mueven a velocidades relativistas lo que permite que se agrupen gravitacionalmente), transparentes (no interactúan más que gravitacionalmente con materia bariónica ni radiación) y no colisionan. Las predicciones del modelo CDM concuerdan con los datos recabados a grandes escalas (\approx Mpc) sin embargo no concuerdan a menores escalas (\approx kpc) [Vázquez González, 2008].

1.4. FONDO CÓSMICO DE MICROONDAS (CMB)

El Fondo Cósmico de Microondas (CMB, por sus siglas en inglés “Cosmic Microwave Background”) es una herramienta valiosa para entender la composición del Universo, debido a que se encuentra en todas partes del cosmos, y proporciona información sobre la distribución y el movimiento de materia y energía en el universo primitivo. Además, las anisotropías en el fondo cósmico de microondas permiten investigar la distribución de la materia oscura y la materia visible, proporcionándonos información importante sobre la evolución y la composición de este.

En 1948, G. Gamow, R. Alpher, y R. Herman, para validar la teoría del Big Bang, propusieron la existencia de un Fondo Cósmico de Microondas que debía provenir de todas direcciones y ser homogéneo [Díaz, 2013]. Por otro lado, en 1964, los científicos Wilson y Penzias trabajaban con una antena de bocina en los laboratorios Bell en Holmdel, Nueva Jersey, sin embargo presentaba un ruido de fondo. Con el propósito de remover cualquier interferencia que pudiera afectar a la antena, ahuyentaron a las palomas, limpiaron sus desechos, enfriaron la antena para evitar el ruido térmico y removieron las ondas emitidas por estaciones de radio cercanas. A pesar de todos sus intentos de eliminar este ruido de fondo, no lo lograron, incluso notaron que este ruido constante venía de todas las direcciones. Por lo que concluyeron que el ruido no venía de la Tierra, ni si quiera de la Vía Láctea. Posteriormente, se descubrió que este ruido provenía de la radiación del CMB [Escamilla Torres, 2018]. El CMB es la foto más antigua que tenemos del Universo, de cuando tenía alrededor de 380,000 años. Es la radiación electromagnética que permea el cosmos, son los fotones emitidos al momento del desacoplamiento que vemos en forma de microondas. Por la energía de éstos fotones se puede determinar la temperatura del Universo, por lo que es una imagen térmica del plasma primitivo [Díaz, 2013]. Actualmente las mediciones indican que esta radiación es casi isotrópica y se encuentra a una temperatura casi uniforme de aproximadamente $2.725 \pm 0.001K$, se cree que sus variaciones de temperatura se debe a inhomogeneidades de la materia que fueron creciendo por las interacciones gravitacionales.

El satélite **COBE** (por sus siglas en inglés “Cosmic Background Explorer”) obtuvo el primer mapa detallado de las anisotropías del CMB, ésta se puede apreciar en la figura (1.7). También midió la distribución de la radiación, validando la teoría del Big

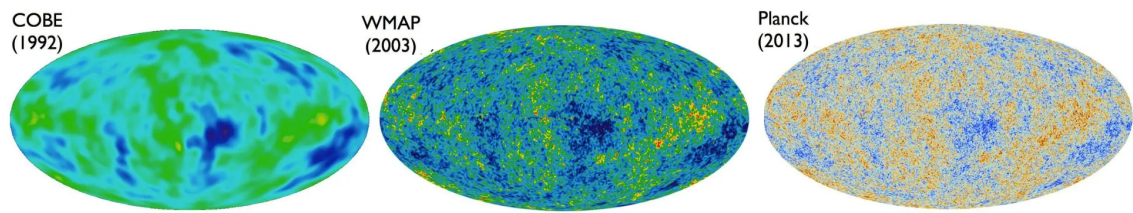


Figura 1.7: Mapa de las anisotropías del CMB obtenidas por COBE, WMAP y Planck respectivamente. Credits: NASA/COBE/DMR; NASA/WMAP science team; ESA and the Planck collaboration.

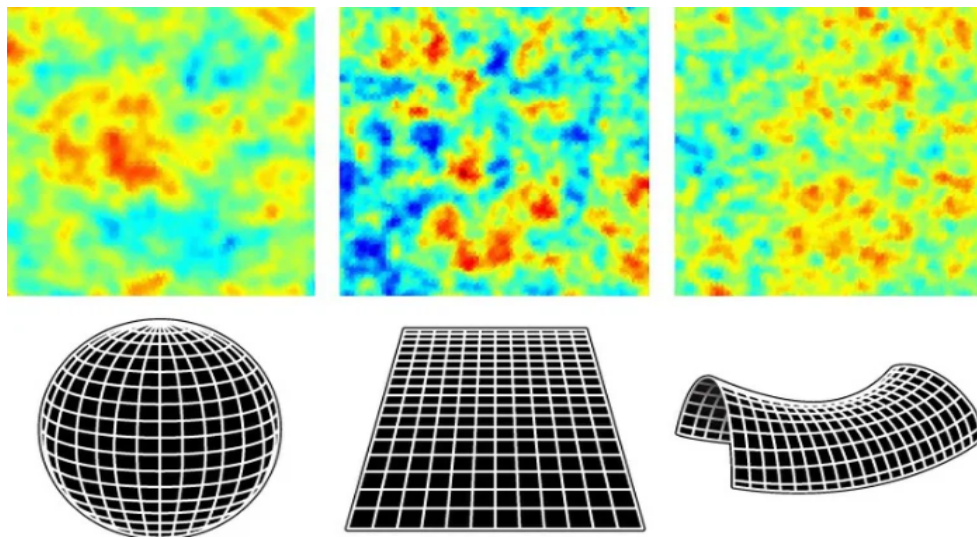


Figura 1.8: Ejemplos de mapas del CMB dependiendo de la curvatura del espacio-tiempo. Fuente: [Coble et al., 2018].

Bang y determinó la temperatura del Universo ($2.725K$) [Díaz, 2013]. Posteriormente, en 2001, se lanzó la sonda **WMAP** (por su acrónimo en inglés “Wilkinson Microwave Anisotropy Probe”). En 2003 se publicaron los datos obtenidos sobre las anisotropías de CMB y como se puede apreciar en la figura (1.7) tuvieron mayor precisión y detalle. De ésta manera se determinó la cantidad de materia oscura y energía oscura que hay en el Universo y que este tiene aproximadamente 13 mil millones de años [Díaz, 2013]. Luego, en 2009 se lanzó el satélite **Planck**. Los datos que obtuvo se comenzaron a publicar en 2013, igualmente se puede apreciar en la figura (1.7) que el mapeo fue más detallado que los anteriores. Confirmó resultados anteriores, determinó la constante de Hubble que mide la tasa de expansión del Universo y, combinada con otras observaciones, puede proporcionar una estimación de la edad del Universo, estimando 13.81 ± 0.05 miles de millones de años [Díaz, 2013].

Las manchitas que se observan en esta misma figura son las anisotropías del CMB. Éstas muestran que había más materia en determinadas regiones del Universo, lo cual fue determinante para la formación de las primeras estructuras, ya que las regiones más densas por gravedad atraen más materia llegando a formar estrellas, galaxias,

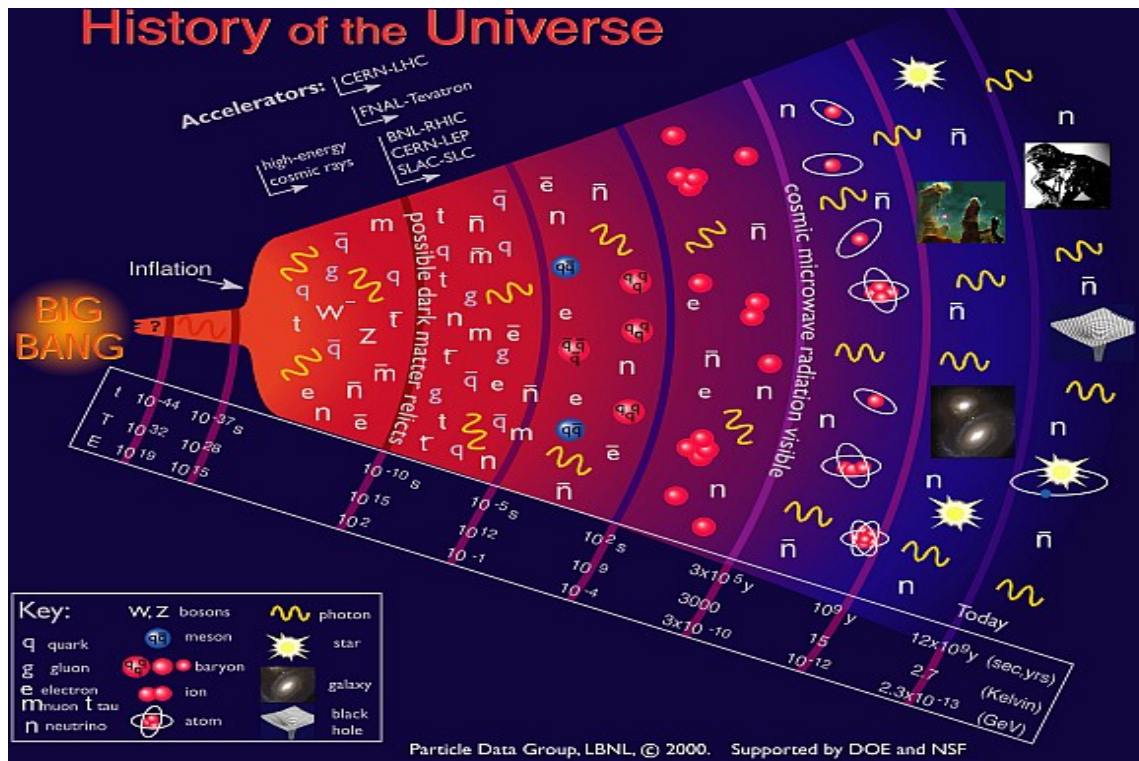


Figura 1.9: Esquema de la historia del Universo. Fuente: Particle data group.

cúmulos, etcétera. La forma de las anisotropías son un indicativo de la forma del Universo. Como se muestra en la figura (1.8): un Universo cerrado magnificará el tamaño de las anisotropías, mientras que uno abierto minimizará el tamaño de estas; mediciones de Planck han encontrado que nuestro Universo es casi plano, por lo que su densidad de materia y energía debe ser similar a la crítica [Escamilla Torres, 2018].

1.4.1. Oscilaciones Acústicas de Bariónes

Así como el CMB proporciona información sobre la estructura y la expansión del universo primitivo, las oscilaciones acústicas bariónicas (BAO del inglés: Baryonic Acoustic Oscillations) son patrones de densidad en la distribución de materia que permiten analizar la estructura y evolución del Universo. Para comprenderlos debemos regresar a lo que sabemos del origen del Universo.

Como se puede apreciar en la figura (1.9), después del origen del Universo, hace aproximadamente 13.8 miles de millones de años, el cosmos estaba lleno de un fluido caliente y denso llamado plasma. Pronto se logró expandir y enfriar lo suficiente para que los bariones se formaran (protones y neutrones). Estaba tan caliente y denso que las partículas al chocar rebotaban en lugar de unirse por lo que los átomos no se podían formar. Había millones de fotones por cada electrón, eran incapaces de viajar una distancia considerable antes de interactuar con los electrones a través de la dispersión de Thompson. En este estado, se dice que la luz estaba acoplada con

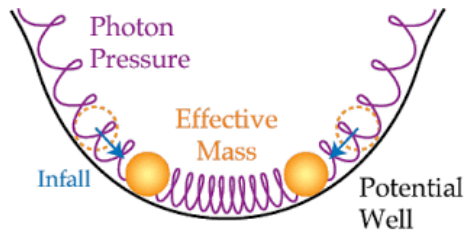


Figura 1.10: Esquema de región de sobredensidad donde hay interacción gravitacional e interacción fotón-materia [Hu, 2022].

la materia, los bariones y fotones formaban un plasma; el Universo era opaco, la luz era capaz de ejercer una enorme presión sobre este plasma lo que produjo ondas de sonido cuyo desplazamiento era cercano a la mitad de la velocidad de la luz [Ashley, 2020, Eisenstein et al., 2005, Matt, 2019].

Las regiones con mayor densidad atraen gravitacionalmente la materia en su entorno (materia oscura, materia bariónica), creando pozos de potencial gravitacional, como se ejemplifica en la figura (1.10). Particularmente, la materia oscura fluyó hacia este pico de densidad, donde los fotones atrapados ejercían una enorme presión en dirección opuesta. Las interacciones fotón-materia comienzan a ser dominantes y la radiación empuja al exterior, en el momento que la presión se aleja suficiente del material, las interacciones gravitacionales vuelven a dominar, provocando ondas sonoras esféricas. La materia oscura permanece en la sobre-densidad porque no interactúa con la luz, sólo interactúa a través de fuerzas gravitacionales, a diferencia de los bariones y fotones que salen con las ondas sonoras, resultando en ondas de presión llamadas BAO que se propagaron a través del plasma y se expandieron rápidamente [Ashley, 2020, Eisenstein et al., 2005, Matt, 2019].

A medida que el Universo se expandió, cuando apenas tenía 380,000 años, el plasma se enfrió por debajo de 3000 K (ver figura 1.9), a esta temperatura los electrones pudieron ser capturados por los núcleos y formar átomos ligeros como hidrógeno, helio y litio. A este evento lo llamamos recombinación, donde el cosmos se vuelve neutro y se elimina la presión sobre los bariones; los electrones atados a átomos están restringidos ahora a ciertas frecuencias correspondientes a los niveles de transición de energía de los átomos, por lo que la luz y la materia dejaron de estar acopladas y el Universo se volvió transparente a los fotones. La energía de los fotones disminuyó por lo que ya no ionizaban y pudieron propagarse libremente formando el Fondo Cósmico de Microondas. A partir de aquí, las ondas dejaron de propagarse, y “El radio característico de la capa esférica formada cuando la onda bariónica se detuvo, se imprime en la distribución de los bariones como un exceso de densidad. Los bariones y la materia oscura interactúan a través de la gravedad, por lo que la materia oscura también se agrupa preferentemente en esta escala” [Bassett and Hlozek, 2009]. Y como en estas ondas existía mayor cantidad de materia, varias galaxias se formaron a lo largo de estas, como se ilustra en la parte izquierda de la misma figura. Ahora, solo

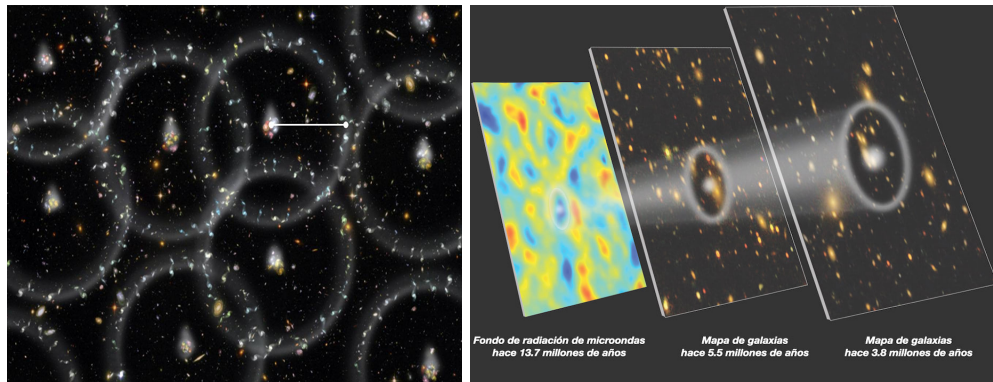


Figura 1.11: Izquierda: Ilustración esquemática de la medida tomada por BOSS [Anónimo, 2014]. Los círculos, muestran una representación gráfica del tamaño actual de las oscilaciones acústicas de bariones (BAO), que si lo propagamos en el tiempo, se asemejaría a un cono, como en la parte derecha. Derecha: "Visualización esquemática de las oscilaciones acústicas de bariones como regla estándar, establecida por la física del plasma primordial e impresa en las propiedades estadísticas del CMB en el Universo temprano (rebanada izquierda) y posteriormente en la distribución de galaxias (rebanadas central y derecha). Créditos: E. M. Huff, la colaboración SDSS-III, y la colaboración del South Pole Telescope. Gráficos: Zosia Rostomian)" [Verde, 2021].

depende de la expansión del Universo como se representa en la derecha de la figura (1.11); es por esto que actualmente es más probable encontrar una galaxia a 150Mpc de distancia de otra [Ashley, 2020, Eisenstein et al., 2005, Matt, 2019].

Por este motivo, los "cascarones esféricos" de radio 150Mpc (parte izquierda de la figura 1.11), quedaron con una sobre-densidad de materia, es decir, la mayoría de galaxias se agrupan en promedio alrededor de estos con cierta perturbación debida a atracciones gravitatorias, velocidades peculiares, entre otros. Por otro lado, las galaxias se conocen como trazadores de materia oscura, porque a pesar de que no se sabe mucho de ésta, sabemos que interactúa gravitacionalmente, por lo que las zonas donde se acumula la materia oscura también se acumula la materia visible, es decir, bariónica y por la atracción gravitacional es posible la formación de las galaxias. Las oscilaciones acústicas de bariones son una prueba indirecta de la materia oscura, dado que ésta permanece en el centro y actúa como un pozo de potencial gravitacional que atrae nuevamente a la materia, por esto también hay una sobredensidad de galaxias en los centros de las BAO. Como consecuencia, existe una escala característica de BAO a la que desde una galaxia es más probable encontrar otra galaxia.

La distribución de galaxias en el Universo es una fuente de información muy importante para la cosmología. El proyecto BOSS (acrónimo de Baryon Oscillation Spectroscopic Survey) realizó un mapeo de la distribución de galaxias rojas luminosas, cuásares y bosques de Lyman- α para detectar la escala característica impresa por las oscilaciones acústicas de bariones logrando una medición indirecta de la energía oscura, a través de su función de correlación de dos puntos (figura 1.12); como lo veremos más adelante. Posteriormente, como se muestra en la figura (1.13), se combinaron

fases anteriores del proyecto SDSS (acrónimo de Sloan Digital Sky Survey) con el proyecto eBOSS (acrónimo de Extended Baryonic Oscillation Spectroscopic Survey) obteniendo datos de la distribución de cuásares y galaxias desde que el Universo tenía entre 3 y 8 mil millones de años (momento en que la energía oscura comenzó a afectar la expansión del Universo) lo que permitió medir con precisión la historia de expansión del Universo (incluyendo parte de la época de desaceleración por efectos de la gravedad hasta la actual aceleración) y por tanto comprender mejor la energía oscura [SDSS_collaboration, 2022].

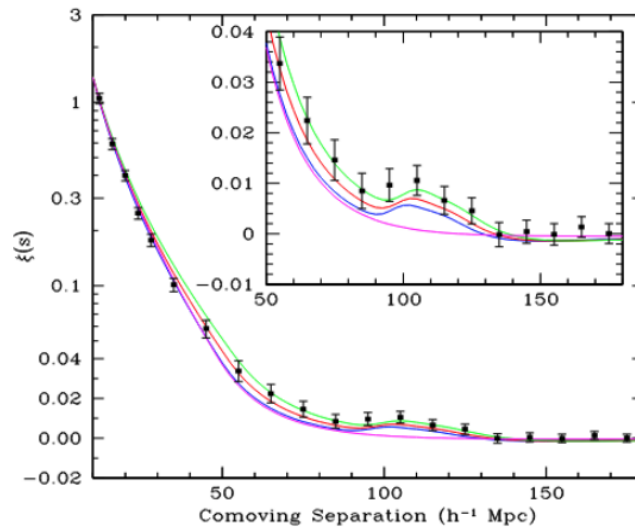


Figura 1.12: Función de correlación medida por BOSS, y el pico acústico bariónico (de sobredensidad) en el agrupamiento de la muestra de galaxias SDSS LRG a escalas de $100h^{-1}\text{Mpc}$ [Bassett and Hlozek, 2009].

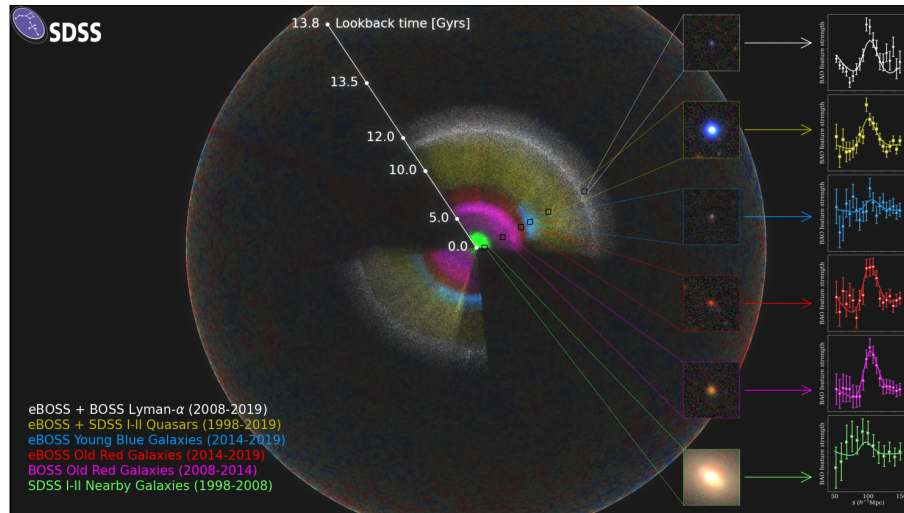


Figura 1.13: Este mapa muestra los datos recabados por distintos sondeos del Universo observable, considerándonos en el centro del mapa. “Los recuadros de cada sección muestran una imagen de galaxia o cuásar típico de la sección y la señal del patrón que el equipo de eBOSS mide allí. Crédito de la imagen: Anand Raichoor (EPFL), Ashley Ross (Universidad Estatal de Ohio) y la Colaboración SDSS” [Anónimo, 2020].

El proyecto DESI (acrónimo de Dark Energy Spectroscopic Instrument) comenzó a tomar datos en 2021, le tomará 5 años cartografiar aproximadamente diez veces más galaxias que sondeos anteriores, obtendrá el espectro óptico de 10 millones de galaxias y cuásares, aportara datos desde hace 11 mil millones de años hasta la actualidad lo cual nos permitirá estudiar la evolución del Universo con mayor precisión [Anónimo, 2021].

Capítulo 2

Distancias

Las distancias nos permiten medir la expansión y la edad del Universo, calibrar las escalas cosmológicas, determinar la distribución de galaxias y materia oscura, entre otras cosas. Para medir a que distancia se encuentran diferentes cuerpos celestes de la Tierra, se han utilizado diferentes métodos como los que se muestran a continuación.

2.1. DISTANCIAS EN COSMOLOGÍA

2.1.1. Paralaje trigonométrico

En el paralaje trigonométrico, análogamente a como medimos las profundidades con nuestros ojos de manera innata, se observa el objeto con un telescopio y luego se observa desde otra posición para detectar los desplazamientos ocurridos en la posición aparente, como se muestra en la figura (2.1). Sin embargo para objetos muy lejanos se deja de observar el cambio de posición aparente, por lo que esperamos a que la Tierra esté del otro lado del Sol para tener mayor distancia entre las observaciones y observar mejor los cambios aparentes, como se muestra en la parte derecha de la misma figura. Mientras más lejos se encuentra el cuerpo celeste, se necesitará observar desde dos puntos más distantes, por lo que sólo funciona para distancias relativamente cercanas. Una desventaja de este tipo de medición es que considera al Universo estático, es decir no considera el movimiento intrínseco de los cuerpos celestes ni la expansión del Universo.

2.1.2. Distancia lumínica

Esta distancia depende del flujo de luminosidad F y la luminosidad L (ecuación 2.1) como se observa en la figura (2.2). Proporciona una buena aproximación para objetos cercanos (dentro de la Vía Láctea). A mayores distancias la magnitud aparente se distorsiona debido a la curvatura del espacio-tiempo, el corrimiento al rojo y la dilatación del tiempo, por lo que se deben considerar estos factores

$$F = \frac{L}{4\pi D_L^2}. \quad (2.1)$$

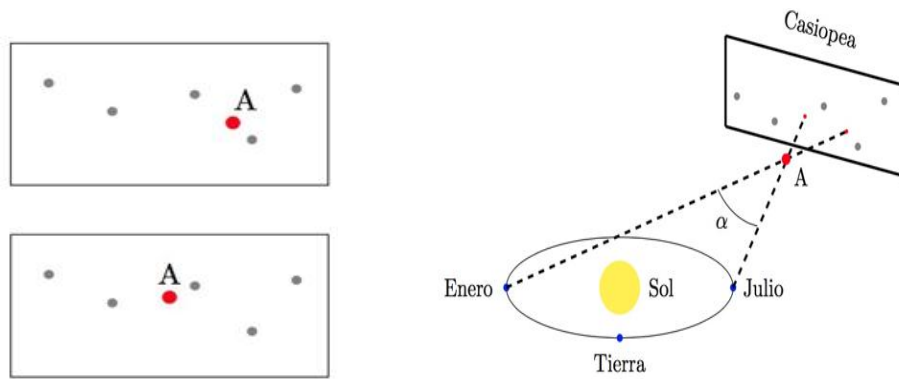


Figura 2.1: Izquierda: Al observar una estrella (en este caso Casiopea) desde la Tierra en diferentes posiciones respecto al Sol podemos observar dos posiciones aparentes. Derecha: Desplazamientos ocurridos en la posición aparente de un cuerpo celeste A al cambiar la posición del observador [M. Manero, 2020].

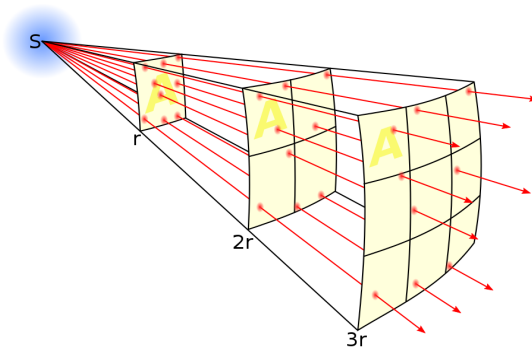


Figura 2.2: Ley del Inverso de los cuadrados la cual establece que: la intensidad de la luz disminuye en relación inversa al cuadrado de la distancia entre la fuente de luz y la superficie iluminada. Fuente: [Anónimo, 2022a].

2.1.3. Distancia comóvil

El sistema de referencia comóvil es el que se mueve con la partícula, por lo tanto para el sistema, la partícula siempre estará en reposo, esto se puede apreciar en la figura (2.3). Éste tipo de coordenadas son bastante útiles porque sirven para considerar la expansión del Universo. Sea $\bar{x} = 0$ nuestra posición y \bar{x} la coordenada comóvil de una galaxia, se tiene que su distancia propia \bar{r} está dada por [Chacón Lavanderos, 2018]:

$$\bar{r} = a(t)\bar{x}, \tag{2.2}$$

donde $a(t)$ es el factor de escala que no depende de las coordenadas, sólo depende del tiempo. Sustituyendo la ecuación (2.2) en la ecuación (1.14) recordando que $\bar{v}(t) = \dot{\bar{r}}(t)$, se define el parámetro de Hubble como [Chacón Lavanderos, 2018]:

$$\bar{v}(t) = \frac{d}{dt}a(t)\bar{x} = H(t)a(t)\bar{x}, \tag{2.3}$$

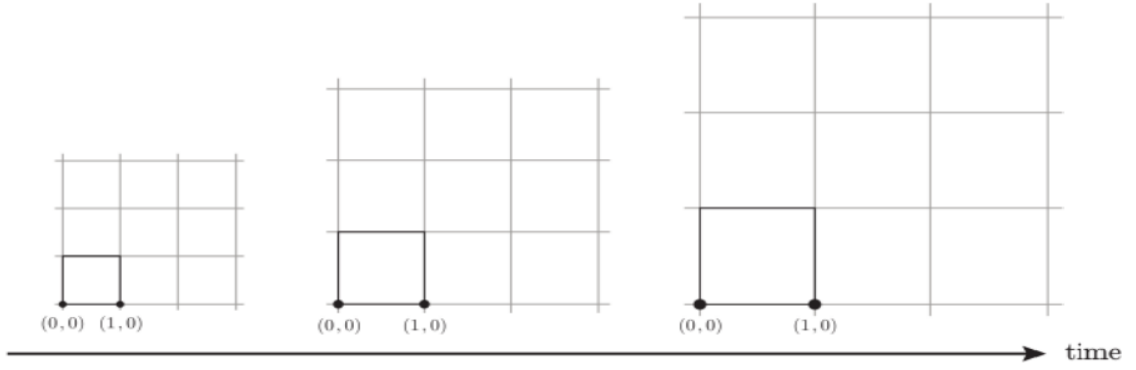


Figura 2.3: Representación de coordenadas comóvil donde a medida que el universo se expande, la separación entre puntos en la red de coordenadas imaginaria se mantiene constante, mientras que la distancia física aumenta en proporción al producto de la distancia comóvil y un factor de escala llamado $a(t)$ [Téllez Tovar, 2018, Baumann, 2015].

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)}, \quad (2.4)$$

mencionado anteriormente en (1.15).

Por otro lado, la distancia comóvil es la distancia instantánea a un corrimiento al rojo dado, se puede calcular con la fórmula (2.5) derivada de la métrica de Friedmann-Lemaître-Robertson-Walker para universos planos:

$$\chi(z) = c \int_0^z \frac{dz'}{H(z')}. \quad (2.5)$$

La distancia transversal comóvil $\chi(z)$ está relacionada con la distancia lumínica D_L , mediante

$$D_L = (1+z)\chi(z). \quad (2.6)$$

2.1.4. Distancia angular

Es la distancia que se genera debido al ángulo que existe entre el eje de visión de un extremo de su diámetro al otro. Se puede deducir la distancia de un objeto, de longitud conocida, por su distancia de diámetro angular, $D_A(z)$, como se muestra en la figura (2.4) con la siguiente fórmula:

$$D_A(z) = \frac{D}{\delta}, \quad (2.7)$$

donde D es el diámetro observable y δ el ángulo que subtiende el objeto. Y se relaciona con la distancia transversal comóvil de la siguiente forma:

$$D_A(z) = \frac{\chi(z)}{1+z}. \quad (2.8)$$

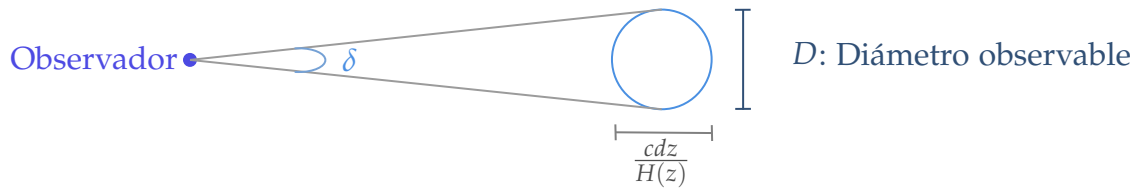


Figura 2.4: Representación esquemática de la distancia de diámetro angular, $D_A(z)$. La longitud radial del objeto está dada por $\frac{cdz}{H(z)}$, con dz la diferencia del corrimiento al rojo entre en frente del objeto y el fin del objeto [Bassett and Hlozek, 2009].

Es importante recordar que el espacio se curva debido a la gravedad, por lo que el mismo espacio actúa como una lente, así que esto complica el detectar si un objeto está cerca, lejos o distorsionado. Aunado a esto no conocemos el tamaño de los objetos extragalácticos en general, ni qué tan brillantes son intrínsecamente. Es por este motivo que en cosmología, se necesita una regla estándar para realizar mediciones [Bassett and Hlozek, 2009].

2.2. MEDICIONES

Ahora que tenemos las bases de cómo se miden las distancias en cosmología, para poder realizar mediciones del Universo observable es importante tener marcos de referencia estándares como: los relojes estándar, candelas estándar y reglas estándar.

2.2.1. Relojes estándar

Los relojes estándar también conocidos como "Cosmic Chronometers" fueron propuestos en el artículo [Jimenez and Loeb, 2002]. Son una técnica utilizada en cosmología para calcular la expansión del Universo, se basa en la selección de galaxias masivas (Large Red Galaxies) con poblaciones de estrellas antiguas y poco polvo estelar para obtener sus espectros fácilmente. Al comparar las distribuciones de edad de dos galaxias con diferentes redshifts, es posible obtener la diferencia de edades y redshifts [Padilla et al., 2021]. Esta técnica es importante porque proporciona una forma independiente de medir la expansión del Universo y ofrece una verificación de consistencia con otros métodos.

2.2.2. Candelas estándar

Se basan en la premisa de que si se conoce como brilla intrínsecamente un objeto astronómico, por la ley del inverso del cuadrado sabemos que el brillo decae como el cuadrado de la distancia de la fuente y sabemos como afecta el corrimiento al rojo. Sólo se debe medir su brillo aparente y sabremos a que distancia se encuentra. Por otro lado, si se desea conocer la velocidad de un objeto, se puede comparar la conocida luminosidad y el corrimiento al rojo. Existen diferentes candelas estándar, un ejemplo son las estrellas Cefeidas, que presentan una luminosidad absoluta proporcional a

su periodo de variabilidad. Otro ejemplo son las Supernovas tipo Ia porque se cree que todas pueden ser calibradas para presentar la misma luminosidad absoluta. Éste tipo de supernova ocurre en un sistema binario entre cualquier tipo de estrella y una enana blanca, la cual es muy densa por lo que atrae materia de la otra estrella y va aumentando su masa a tal punto que genera una reacción nuclear en cadena y explota. Como todas se comportan de forma muy parecida, emiten aproximadamente la misma cantidad de luz al explotar y por eso pueden servir como candelas estándar [Escamilla Torres, 2018]. Incluso fueron determinantes en el descubrimiento de la expansión acelerada del Universo [Riess et al., 1998, Perlmutter et al., 1999].

2.2.3. Reglas estándar

Debido a que sabemos como escalan los tamaños de objetos respecto a la longitud, podemos usarlos para medir distancias si se conoce el tamaño del objeto y se mide su tamaño aparente, es decir, su distancia angular. Se sabe que las galaxias se agrupan en promedio alrededor de cascarones de esferas (BAO) que actualmente tienen radio de 150Mpc, por lo que es la escala preferida a la que se encuentran las galaxias. Los BAO funcionan como regla estándar y al medirlos a diferentes corrimientos al rojo se puede ver cuál era su tamaño en cierto momento y con esto se mide la expansión del Universo, lo cual nos da información sobre el comportamiento de la energía oscura [Bassett and Hlozek, 2009, Alonso, 2013].

2.3. SIMULACIONES COSMOLÓGICAS

Las simulaciones cosmológicas nos ayudan a estudiar el comportamiento del Universo debido a que se modelan de acuerdo a observaciones de sondeos, como SDSSIII/-BOSS [SDSS-II_Collaboration, 2014], PanSTARRS [Kaiser, 2007] o DESI [DESI_Collaboration et al., 2016] que escanean el cielo a detalle, y en éstas podemos hacer que el tiempo pase de manera distinta, por lo que nos ayudan a predecir como se vería el Universo en distintos instantes (corrimientos al rojo) y escenarios, así como seguir la evolución de los halos. Las simulaciones son costosas computacionalmente debido a que requieren un gran número de partículas con suficiente resolución de masa para identificar los halos. El tiempo, cantidad de partículas y longitud del espacio simulado son clave para obtener resultados cercanos a la realidad.

Se ha trabajado con simulaciones cosmológicas computacionales cómo actualmente las conocemos desde los años setenta, como una herramienta para entender la formación de estructura a gran escala del Universo, sin embargo la primera simulación astrofísica se remonta a 1941, dónde Erik Holmberg, registró la primer simulación de N-cuerpos en el artículo [Holmberg, 1941]. Holmberg simuló la interacción en un plano de dos galaxias elípticas con una computadora analógica, donde los elementos de masa fueron representados por bombillas, la masa era proporcional a la intensidad lumínica y ésta se midió mediante una fotocélula. Cada nebulosa se representó por

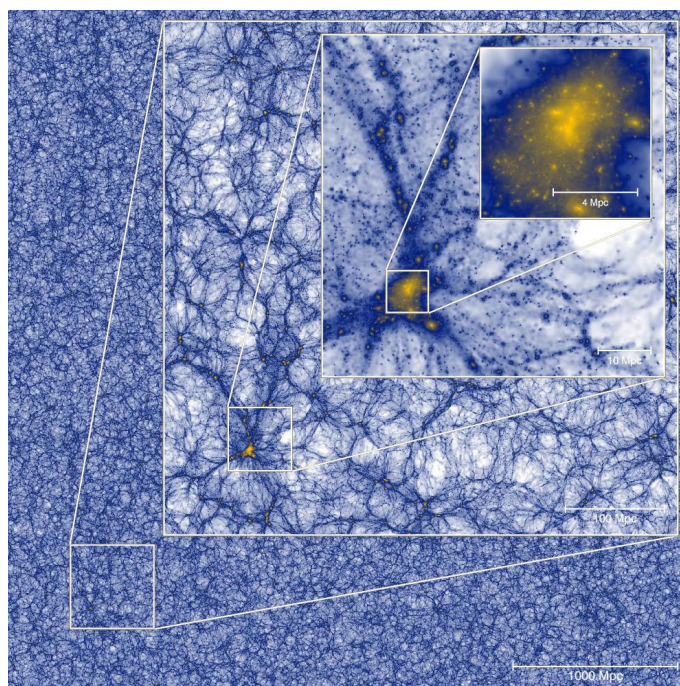


Figura 2.5: Corte 2D de la predicción de la distribución de galaxias por Millennium XXL. La escala está marcada con una línea blanca para cada cuadro, y tiene un valor de 1000Mpc, 100Mpc, 10Mpc y 1Mpc respectivamente. Fuente: [Angulo, 2011].

37 bombillas [Holmberg, 1941]. A diferencia de las simulaciones astrofísicas que se centran en escalas más pequeñas, como estrellas, galaxias y objetos celestes específicos; las simulaciones cosmológicas se enfocan en la evolución a gran escala del universo y la formación de grandes estructuras cósmicas. Con éstas podemos realizar aproximaciones estadísticas del comportamiento del Universo y los resultados de las predicciones de estas simulaciones se pueden comparar y/o contrastar con los datos observacionales, lo cual nos da una idea de que tan precisa es la predicción [Alonso, 2020].

Cada simulación cosmológica es única y sigue su propia metodología, sin embargo, comúnmente se establecen las condiciones iniciales que describen el estado del Universo después del Big Bang, se resuelven las ecuaciones diferenciales y se modela la interacción de partículas. En general las simulaciones de Materia Oscura (N-cuerpos) son simulaciones de la evolución de un fluido no colisional autogravitante, esto requiere resolver la ecuación de Boltzmann no colisional para un fluido perfecto y para las simulaciones de materia bariónica también se debe realizar Hidrodinámica de partículas suavizadas (SPH, por sus siglas en inglés "Smoothed Particle Hydrodynamics").

2.3.1. Millenium

Es una simulación cosmológica de N cuerpos, en 2005 fue publicado por un grupo internacional llamado "Virgo Consortium" **Millennium Run**, esta es una simulación

de Materia Oscura (N-cuerpos) que utiliza 10^{10} partículas en una región cúbica de $500\text{Mpc}/h$ por lado, con resolución espacial de $5\text{kpc}/h$ y $8.6 \times 10^8 M_\odot h$ masa de partículas. En su momento fue la simulación más grande y sigue el modelo estándar de cosmología [Lemson, 2006]. A pesar de su gran resolución de masa, su volumen es insuficiente para los datos estadísticos que posteriormente se desearon obtener. Es por eso que en 2008 se publicó **Millennium-II**, el cual utiliza la misma cantidad de partículas que Millenium sólo que su región cúbica disminuyó a $100\text{Mpc}/h$ por lado, mejorando la resolución espacial a $1\text{kpc}/h$, con una resolución de masa 125 veces mejor y $8.6 \times 10^8 M_\odot h$ masa de partículas [Boylan-Kolchin et al., 2009].

Posteriormente los resultados de la simulación dejaron de concordar con los nuevos datos observacionales, sin embargo se fueron ajustando ciertos parámetros para mejorar estas simulaciones. Luego, en 2010 se publicó **Millenium XXL**, ésta utiliza 6720^3 partículas en una región cúbica de $3000\text{Mpc}/h$, y $6.17 \times 10^9 M_\odot h$ masa de partículas. En la figura (2.5) se muestra como se ve la simulación Millenium XXL a diferentes escalas.

2.3.2. Bolshoi

En 2008 se desarrollo la simulación cosmológica de N-cuerpos **Bolshoi**, con $250\text{Mpc}/h$ de tamaño de caja y 2048^3 partículas. Los resultados del WMAP5 se utilizaron para determinar su cosmología, se desarrolló "en el supercomputador Pleiades del NASA Ames Research Center, empleando el código ART (Adaptive Refinement Tree) y OpenMP para el proceso de paralelizado" [Alonso, 2020]. La figura (2.6) presenta la simulación Bolshoi. Luego, con la información aportada por la misión Planck de la ESA, se realizó la simulación **BolshoiP** con las mismas características de tamaño y partículas, de tal manera que se estudiaron diferentes cosmologías [Alonso, 2020].

2.3.3. Uchuu

Ésta simulación de N-cuerpos, mostrada en la figura 2.7 sigue la cosmología de Planck y utiliza 12800^3 partículas de materia oscura en una región cúbica de $2000\text{Mpc}/h$, con masa de partícula $3.27 \times 10^8 M_\odot h$. Posteriormente se publicó **Shin-Uchuu**, ésta utiliza 6400^3 partículas de materia oscura en una región cúbica de $140\text{Mpc}/h$, con masa de partícula $8.97 \times 10^5 M_\odot h$ [Ishiyama et al., 2021].

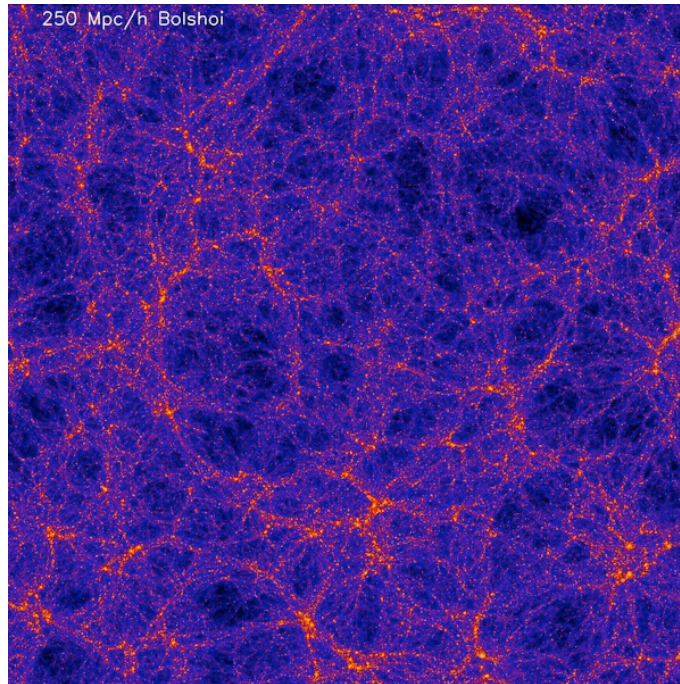


Figura 2.6: Corte 2D de la simulación cosmológica de N-cuerpos Bolshoi. Fuente: [Bol, 2015].

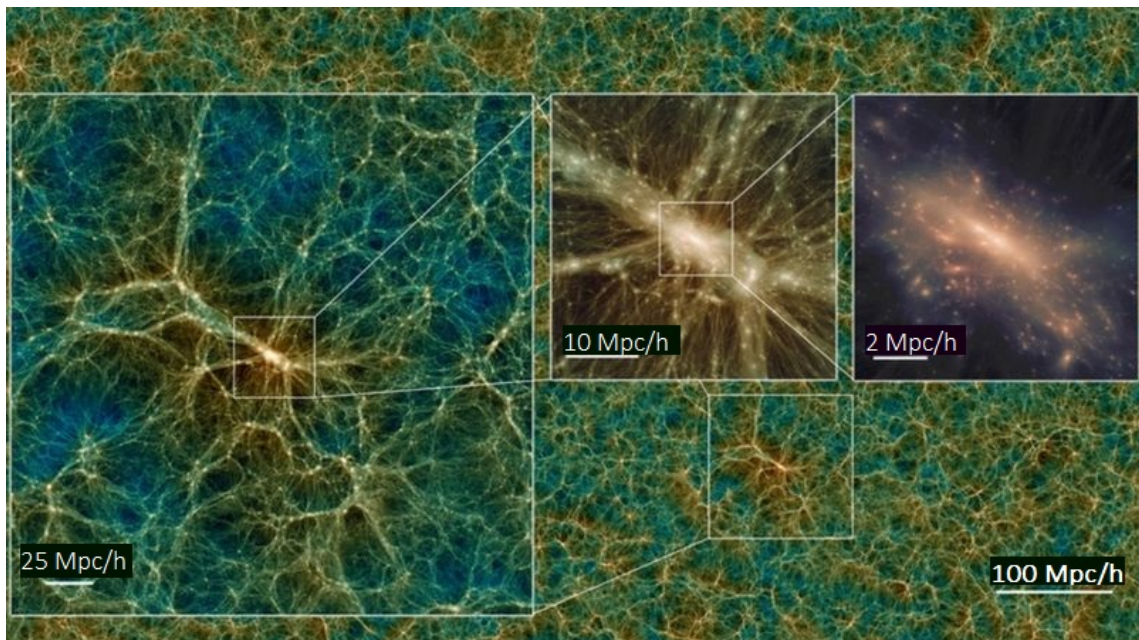


Figura 2.7: En el fondo se muestra un corte en 2D de la simulación cosmológica de N-cuerpos Uchuu. La escala está marcada con una línea blanca para cada cuadro, y tiene un valor de 100Mpc, 25Mpc, 10Mpc y 2Mpc respectivamente. Fuente: skiesanduniverses.org.

Capítulo 3

Función de Correlación

En los capítulos anteriores vimos que a grandes escalas el Universo es homogéneo. Para una distribución de este tipo se cumple que la probabilidad de encontrar un dato, en este caso lo llamaremos galaxia, en un elemento de volumen infinitesimal está dado por [Biermann, 2019],

$$dP_1 = \bar{n}dV. \quad (3.1)$$

Dónde \bar{n} es la densidad numérica y dV es el volumen infinitesimal. Al multiplicar el diferencial de volumen se multiplica la probabilidad de encontrar galaxias. Y si consideramos todas las diferenciales de probabilidad, integrando (3.1) obtenemos la cantidad media de galaxias en ese volumen lo cual queda [Biermann, 2019],

$$\langle N \rangle = \bar{n}V. \quad (3.2)$$

La función de correlación de dos puntos está definida para un campo continuo y es el exceso de probabilidad, respecto a una distribución aleatoria, de que dos partículas en distintos volúmenes dV_1 y dV_2 respectivamente estén separadas por una distancia r_{12} [Alonso, 2020],

$$dP = n_g^2 [1 + \xi(r_{12})] dV_1 dV_2. \quad (3.3)$$

Si se toman dos distribuciones homogéneas las probabilidades de encontrar galaxias en dos elementos de volumen son independientes, por lo que no hay correlación y $\xi(r) = 0$. Mientras que si existe una sobre-agrupación de datos, se tiene que $\xi(r) > 0$ y si hay una infra-agrupación $\xi(r) < 0$. De esta forma sabiendo que en dV_1 se encontró una galaxia, la probabilidad de encontrar una galaxia en dV_2 está dada por [Biermann, 2019]:

$$dP(2|1) = \bar{n}dV_2 [1 + \xi(r_{12})]. \quad (3.4)$$

Ahora, si nos colocamos sobre una galaxia en dV_1 , la probabilidad de encontrar otra en dV a una distancia r está dada por [Biermann, 2019]:

$$dP(r) = \bar{n}dV [1 + \xi(r)]. \quad (3.5)$$

3.0.1. Transformada de Fourier

La transformada de Fourier de $\xi(r)$ de una función de correlación nos proporciona información sobre el espectro de potencias. Esta se utiliza para describir la estructura del Universo ya que relaciona la cantidad de agrupación a ciertas escalas [He, 2021a]; sin embargo es un tema que dejaremos para futuros trabajos.

3.0.2. Aplicaciones

Además de funciones de correlación de galaxias también existen de pixeles (lyman- α) [Sjödahl, 2019], de cúmulos de galaxias [García-Lambas, 1984], de lentes gravitacionales [González Gonzáles, 2013], entre otras. Las funciones de correlación de dos puntos no son la única forma en que se puede correlacionar datos, también existen las correlaciones cruzadas [Nadathur et al., 2019], funciones de correlación de 3 puntos [Takada and Jain, 2003] y funciones de correlación de N puntos.

3.0.3. Estimadores de la Función de Correlación

Un método para extraer información estadística sobre la distribución de galaxias (como la escala característica) es a través de la función de correlación de dos puntos, $\xi(r)$, que cuantifica estadísticamente el exceso de agrupamiento de una determinada distribución de objetos en relación con una distribución uniforme, o análogamente la probabilidad de encontrar un par de galaxias separadas por la distancia espacial r o la distancia angular θ con respecto a la probabilidad de encontrar un par de galaxias separadas por la misma distancia o ángulo en una distribución aleatoria y uniforme [Ponce et al., 2012]. Las funciones de correlación están definidas en un espacio continuo, sin embargo para extraer información estadística, en la práctica definimos estimadores de la función de correlación en un espacio discreto, para esto contamos pares de datos separados un cierto rango de distancia (por ejemplo los datos separados por una distancia entre 0 y 1 unidades o entre 1 y 2 unidades). Es importante notar que el tiempo de cálculo aumentará con el cuadrado de datos que se trabajen [Alonso, 2013].

Una escala característica en el agrupamiento de galaxias se observará en la función de correlación como un pico o caída, dependiendo de si hay un exceso o una deficiencia de agrupamiento en esa escala [Bassett and Hlozek, 2009]. Existen diferentes estimadores, cada uno tiene sus ventajas y desventajas, sin embargo en el artículo [Pons-Borderia et al., 1999], se analizaron seis estimadores, y los autores no encontraron ningún ganador destacado entre estos; por otro lado "en [Kerscher, 1999, Kerscher et al., 2000a], se consideraron nueve estimadores, y los estimadores que presentaron mejores propiedades fueron los estimadores de Landy-Szalay y Hamilton" [Vargas-Magaña et al., 2013]. Por este motivo, en este trabajo se analizarán distintos estimadores.

Ahora, la forma más sencilla de estimar la función de correlación de dos puntos

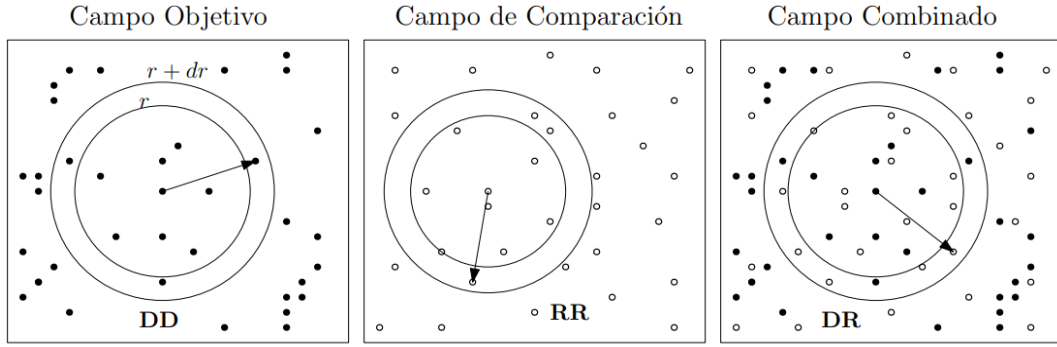


Figura 3.1: Campos utilizados para la estimación de la función de correlación [Alonso, 2020].

$\xi(r)$ es con el estimador natural:

$$\xi(r) = \frac{dd}{rr} - 1, \quad (3.6)$$

donde dd es la distribución de distancias de pares de galaxias dentro de un conjunto de datos, como se muestra en el primer cuadrante de la figura (3.1), donde se contabilizan todos los pares de datos separados por cada rango de distancia. Y rr es la distribución de distancias de un conjunto aleatorio, como se muestra en el segundo cuadrante de la figura (3.1). Si $\xi(r) = 0$ la distribución de galaxias en los datos también es aleatoria; $\xi(r) > 0$ existe una sobredensidad y, contrariamente, para $\xi(r) < 0$ una infradensidad.

No obstante, como la cantidad de puntos aleatorios puede ser diferente a la cantidad de galaxias debemos normalizarlo. Es por eso que se define a n_d como la cantidad de objetos en la muestra de datos y n_r la cantidad de objetos pseudoaleatorios. Normalizamos los pares de la muestra de datos de la siguiente forma [Vargas-Magaña et al., 2013]:

$$DD(s) = \frac{dd(s)}{n_d(n_d - 1)/2}, \quad (3.7)$$

análogamente normalizamos los pares de la muestra aleatoria de la siguiente forma:

$$RR(s) = \frac{rr(s)}{n_r(n_r - 1)/2}. \quad (3.8)$$

Por lo que el estimador natural normalizado queda:

$$\xi(r) = \frac{DD}{RR} - 1. \quad (3.9)$$

Ahora, también se han propuesto estimadores que involucran el conteo de separación de pares cruzados $\bar{d}r$ entre el conjunto de datos y el conjunto aleatorio, como se muestra en el tercer cuadrante de la figura (3.1), con el fin de reducir la varianza de estimación inducida por efectos de borde. También normalizamos estos pares y queda

de la siguiente forma [He, 2021a, Vargas-Magaña et al., 2013]:

$$DR(s) = \frac{dr(s)}{n_r n_d}. \quad (3.10)$$

Otros ejemplos de funciones de correlación son las obtenidas con diferentes estimadores presentados en el artículo [Vargas-Magaña et al., 2013]:

Peebles & Hauser (1974):

$$\zeta_{PH}(s) = \frac{DD}{RR} - 1, \quad (3.11)$$

Hewett (1982):

$$\zeta_{Hew}(s) = \frac{DD - DR}{RR}, \quad (3.12)$$

Davis & Peebles (1983)

$$\zeta_{DP}(s) = \frac{DD}{DR} - 1, \quad (3.13)$$

Hamilton (1993)

$$\zeta_H(s) = \frac{DD \times RR}{DR^2} - 1, \quad (3.14)$$

Landy-Szalay (1993)

$$\zeta_{LS}(s) = \frac{DD - 2DR + RR}{RR}. \quad (3.15)$$

En el artículo [Kerscher et al., 2000b] se evaluaron diferentes estimadores para la función de correlación de dos puntos en más de 500 submuestras de una muestra de 222,052 cúmulos de *Virgo Hubble volume simulation* y se llegó a las siguientes conclusiones: A pequeñas escalas no hubo un estimador destacable, sin embargo, para grandes escalas hubo dos estimadores que presentaron las menores fluctuaciones: **Landy Szalay** y **Hamilton**. Ambos estimadores presentan resultados similares, no obstante, el estimador **Hamilton** es más sensible a la cantidad de puntos aleatorios empleado, por lo que en ese sentido el estimador **Landy-Szalay** es más recomendado. De los demás estimadores se obtuvieron resultados con mayor varianza, los presentamos de mejor a peores resultados de la siguiente forma: **Davis-Peebles**, **Hewett** y por último **Peebles Hauser**. Es importante destacar que se aplicaron también estimadores geométricos pero no los abordaremos en este trabajo [Kerscher et al., 2000b].

3.0.4. Versión angular de la función de correlación

Ahora, utilizando la versión angular de la función de correlación $\omega(\theta)$. Landy-Szalay (1993), encontraron un estimador con varianza mínima que es el estándar utilizado en los análisis cosmológicos [Ponce et al., 2012]:

$$\omega(\theta) = 1 + \left(\frac{N_{random}}{N_{datos}} \right) 2 \cdot \frac{DD(\theta)}{RR(\theta)} - 2 \left(\frac{N_{random}}{N_{datos}} \right) 2 \cdot \frac{DR(\theta)}{RR(\theta)}, \quad (3.16)$$

donde N_{datos} es el número de galaxias, N_{random} es el número de datos aleatorios, $DD(\theta)$ es el número de pares separados por una distancia angular θ en el catálogo de galaxias, $RR(\theta)$ es el número de pares separados por una distancia angular θ en el catálogo aleatorio y $DR(\theta)$ es el número de pares separados por una distancia angular θ en el catálogo de datos con respecto al catálogo aleatorio [Ponce et al., 2012].

Capítulo 4

Aprendizaje automático

En cosmología al trabajar con sondeos o simulaciones, se necesita una gran cantidad de datos para obtener resultados asertivos, esto implica mucho computo y recientemente con el aumento de la tecnología y el poder de computo ya es posible realizar esta tarea fácilmente. Además de hacer cálculos, también es importante disminuir tiempo y hacerlos mas eficientes, por tanto recurrimos al nuevo campo llamado **aprendizaje automático**. Particularmente, en este trabajo, nos interesa detectar agrupamientos de galaxias por lo que nos enfocaremos en los **métodos de agrupamiento**.

La **inteligencia artificial** es el campo de estudio de la informática cuyo objetivo es hacer que las computadoras imiten y/o desarrollen comportamientos inteligentes. Esta definición es relativa ya que primero tendríamos que definir lo que es la "inteligencia". Actualmente hay muchos ejemplos de inteligencia artificial, como programar una máquina para que pueda mover cajas de un lugar a otro, la aspiradora inteligente que detecta distintas superficies, los carros que se manejan solos, entre otros. El **aprendizaje automático**, comúnmente llamado "machine learning", es una rama de la inteligencia artificial que tiene como objetivo programar computadoras para que aprendan de los datos, de esta forma adquieren la habilidad de resolver problemas sin ser programadas explícitamente. Por ejemplo, no sólo se programa la máquina para moverse de un lado a otro, sino que se le enseña a caminar por distintos lugares y se programa para que, basado en la experiencia que le dio caminar en distintos lugares, pueda hacerlo en nuevos caminos. Los algoritmos de aprendizaje automático se pueden clasificar de la siguiente forma: supervisado, no supervisado, semisupervisado, y aprendizaje por refuerzo [Géron, 2017].

Los algoritmos de **aprendizaje supervisado** reciben datos etiquetados con su respuesta deseada, esto requiere intervención humana para etiquetar los datos, luego establece la correspondencia o patrones que hay entre los datos y su etiqueta, a esta etapa se le llama entrenamiento. A continuación, se le entrega un nuevo conjunto de datos sin etiquetar y a partir de la experiencia que obtuvo con los datos etiquetados realiza las predicciones de las nuevas etiquetas [Géron, 2017, Garreta and Moncecchi, 2013]. Los algoritmos de **aprendizaje no supervisado**, como entrenamiento, reciben

datos que no están etiquetados por categorías, a partir de estos se buscan patrones para comprender el conjunto de datos. A continuación, se le entrega un nuevo conjunto de datos y de acuerdo a los patrones encontrados predice las etiquetas [Géron, 2017, Garreta and Moncecchi, 2013]. Los algoritmos de **aprendizaje semisupervisado** son una combinación de los dos tipos de algoritmos anteriores. Como entrada recibe datos etiquetados y sin etiquetar. Son muy útiles cuando se tienen pocos datos etiquetados [Garreta and Moncecchi, 2013]. Los algoritmos de **aprendizaje por refuerzo** como su nombre lo dice, tienen como objetivo enseñar a su agente (modelo) a tomar las decisiones más óptimas por medio de refuerzos negativos y/o positivos. El agente se entrena para tomar las mejores decisiones, éste interactúa con el ambiente que establece las limitaciones, así como las leyes de la física establecen nuestros límites. En ese ambiente el agente puede tomar diferentes acciones y a partir de la acción tomada el estado del ambiente cambia y el agente recibe su refuerzo negativo o positivo [Géron, 2017, Garreta and Moncecchi, 2013, Zhang et al., 2022a].

4.1. MÉTODOS DE AGRUPAMIENTO

En este trabajo para el estudio de la formación de estructura de galaxias nos enfocaremos en los **algoritmos de agrupamiento** e intentaremos determinar cuales algoritmos son mejores candidatos para esta tarea.

Los métodos de agrupamiento o clustering son un tipo de algoritmos de aprendizaje automático no supervisado. Hacen referencia a diversas técnicas usadas para encontrar similitudes, patrones, o grupos dentro de un conjunto de datos, los datos que se guardan en un grupo (también llamado *cluster*) deben ser similares entre si y distintos de otros grupos, de acuerdo a diversos criterios de comparación. Debido a que su aplicación abarca gran variedad de campos, se han desarrollado distintos métodos de agrupamiento que se pueden dividir en diferentes tipos, de acuerdo al enfoque que tienen para crear agrupaciones o los parámetros que utilizan para determinarlas. Para medir que tan similares son dos puntos se utilizan funciones de similitud y distancias. En el presente trabajo nos centraremos en las distancias. Sean $X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, $Y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ dos puntos, la distancia de Minkowski de orden p está definida de la siguiente forma [Torres Rudloff, 2017]:

$$L_p(X, Y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}, \quad (4.1)$$

donde $p \geq 1$ para que sea una métrica que resulta de la desigualdad de Minkowski. Cuando $p = 1$, se conoce como distancia Manhattan, esta es mayor que la euclidiana pero en ciertas ocasiones es más útil, un ejemplo es en la ciudad, donde aunque las distancias más cortas entre dos puntos son líneas rectas, no se puede tomar ese camino debido a que hay edificios debemos rodear las cuadras. Considerar este tipo

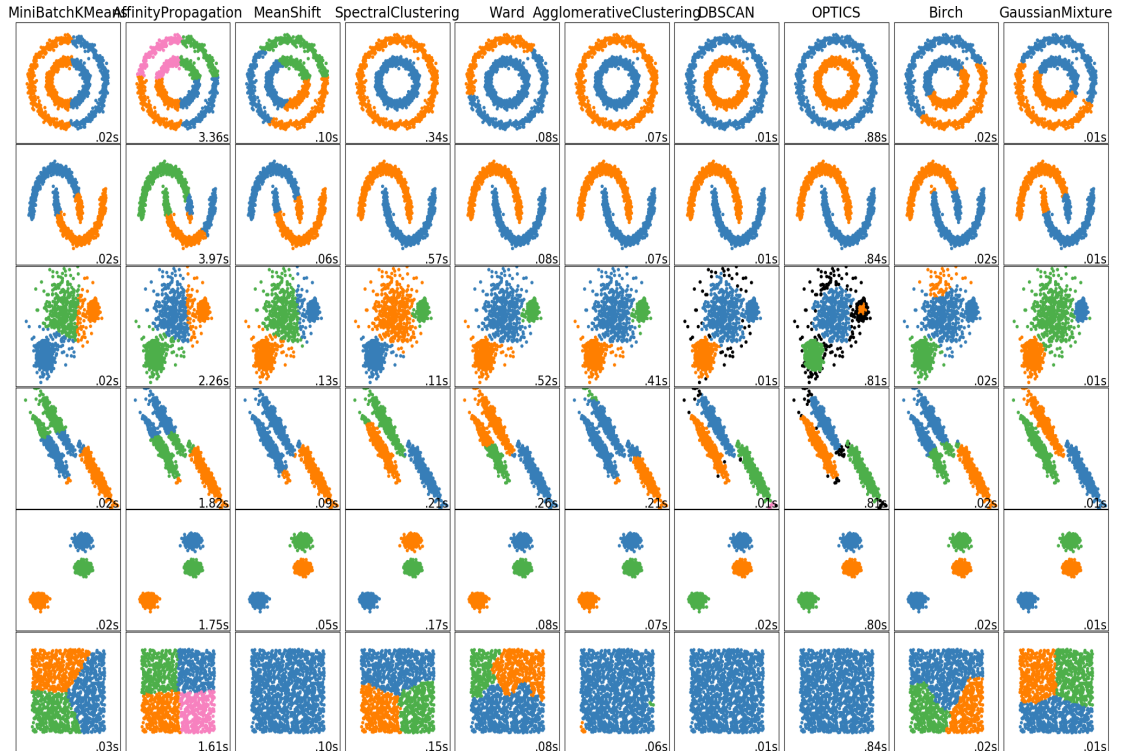


Figura 4.1: Clasificación de distintas distribuciones de datos por diez algoritmos de agrupamiento distintos. Foto obtenida de [Martinez Heras, 2020].

de distancia tiende a formar agrupamientos hiperrectangulares [Torres Rudloff, 2017]

$$L_1(X, Y) = \left| \sum_{i=1}^n |x_i - y_i| \right|. \quad (4.2)$$

Por otro lado, cuando $p = 2$ se conoce como distancia Euclidiana, ésta es la representación estadística de la varianza total en el agrupamiento y presenta una tendencia a formar agrupamientos hiperesféricos [Torres Rudloff, 2017]

$$L_2(X, Y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}. \quad (4.3)$$

Mientras que cuando $p \rightarrow \infty$ es conocida como distancia de Chebyshev, en ésta la distancia entre dos vectores es la mayor diferencia entre cualquiera de sus ejes.

$$L_\infty(X, Y) = \lim_{k \rightarrow \infty} \left(\sum_{i=1}^n |p_i - q_i|^k \right)^{1/k}. \quad (4.4)$$

En la figura (4.1) se muestra una comparación de como distintos algoritmos de agrupamiento clasifican distintas distribuciones de datos y se puede ver que para nuestros propósitos DBSCAN y OPTICS se asemejan más a nuestro problema.

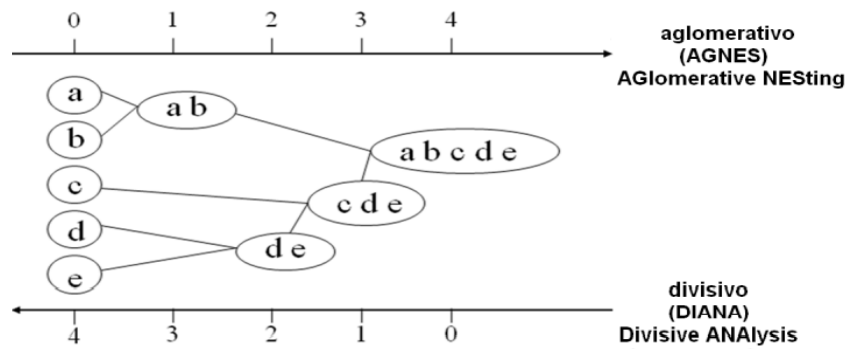


Figura 4.2: Ejemplo de como algoritmos jerárquicos aglomerativo y divisivo trabajan los datos. Fuente: [Lara Torralbo, 2010].

Con el presente proyecto se busca estudiar algoritmos de agrupamiento para localizar patrones dentro de las distribuciones de galaxias y hacer una comparación directa con mediciones recientes. Notese que este tipo de algoritmos de agrupamiento se ha usado ampliamente para la búsqueda de halos de materia oscura en simulaciones numéricas, por ejemplo friends-of-friends, o una adaptación jerárquica llamada ROCKSTAR [Behroozi et al., 2012]. A continuación se presentan los diferentes tipos de algoritmos de agrupación [Anónimo, 2022b]:

- **Agrupamientos jerárquicos (Hierarchical Clustering):** Algunos ejemplos de algoritmos jerárquicos son: BIRCH [Zhang et al., 1996], CURE [Guha et al., 1998], ROCK [Guha et al., 2000] y Chamaleon [Karypis et al., 1999]. Este tipo de algoritmos busca dividir los datos jerárquicamente y se puede realizar de manera divisiva o aglomerativa. Como se observa en la figura (4.2), en la forma divisiva todos los datos pertenecen al mismo agrupamiento y se van dividiendo en subagrupamientos. Por otro lado, en la forma aglomerativa cada objeto comienza en su agrupamiento y se van juntando de tal manera que sube en el nivel jerárquico [Torres Rudloff, 2017], esto también se ejemplifica en la figura (4.2). Son útiles para datos con formas arbitrarias y es relativamente fácil detectar relaciones jerárquicas. Sin embargo son complejos y se necesita conocer el número de agrupamientos antes [Torres Rudloff, 2017].
- **Agrupamientos basados en centroides (Centroid Based Clustering):** A diferencia de los algoritmos jerárquicos, éstos producen una partición plana (de un sólo nivel) [Ankerst et al., 1999]. En este tipo de algoritmos el usuario debe entregar el número de agrupamientos y el algoritmo determina el número de centroides, cada algoritmo sigue cierto método para elegirlos minimizando lo más posible la distancia de cada centroide y los datos que pertenecen a éste. Los algoritmos de este tipo en general requieren poco tiempo computacional, sin embargo son sensibles al ruido (outliers), no son útiles para conjuntos de datos no convexos¹ y

¹Se considera que un conjunto de datos es convexo si cada par de datos se puede unir por una línea recta sin salir del agrupamiento.

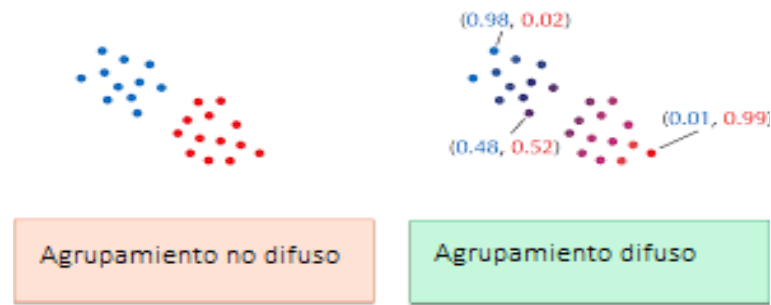


Figura 4.3: La figura muestra la diferencia entre agrupamiento no difuso o "hard clustering" y agrupamiento difuso o "soft clustering". El color rojo o azul representa el agrupamiento al que pertenecen. Fuente: [Wendy,]

necesitan que el usuario sepa la cantidad de agrupamientos que hay en los datos [Torres Rudloff, 2017]. Algunos ejemplos de algoritmos basados en centroides son: k-Means [Morissette and Chartier, 2013], k-Mediods [Schubert and Rousseeuw, 2021], CLARA [Schubert and Rousseeuw, 2021] y CLARANS [Schubert and Rousseeuw, 2021].

- **Agrupamientos difusos (Fuzzy Based Clustering):** Este tipo de algoritmo no agrupa los datos de forma discreta, como se aprecia en la figura (4.3), a diferencia de los otros tipos de algoritmos de agrupamiento donde cada dato pertenece a un agrupamiento, en este caso cada dato puede tener un grado de pertenencia a distintos agrupamientos. Estos algoritmos son más realistas, sin embargo suelen caer en óptimos locales y requiere que el usuario entregue el número de agrupamientos [Torres Rudloff, 2017]. Algunos ejemplos de algoritmos de clustering difuso son: Fuzzy C-Means [Bezdek et al., 1984].
- **Agrupamientos basados en malla (Grid Based Clustering):** Este tipo de algoritmos divide el espacio de datos en cuadrantes con tamaños iguales de una malla cuadrangular, descarta los cuadrantes sin datos, los cuadrantes de baja densidad son etiquetados como ruido y los de alta densidad representan agrupamientos, luego combina agrupamientos de cuadrantes contiguos. El éxito de este método depende del tamaño de las celdas, el cual es un parámetro de entrada para el usuario, si son muy pequeñas darán una estimación de densidad muy "ruidosa", mientras que si son muy grandes tienden a suavizar en exceso la estimación de densidad [Ankerst et al., 1999]. Una de las desventajas de este tipo de algoritmo es que no puede agrupar datos con densidades muy distintas [Torres Rudloff, 2017]. Algunos ejemplos de algoritmos basados en densidad son: CLIQUE [Forster and Murphy, 2009] y MAFIA [Burdick et al., 2005].
- **Agrupamientos basados en densidad (Density-Based Clustering):** Este tipo de algoritmo agrupa los datos de acuerdo a regiones con mucha concentración de datos, que se separa de regiones con pocos datos que se consideran ruido. Son

útiles para datos de distintas formas, no requieren que el usuario proporcione el número de agrupamientos, al considerar el ruido los valores atípicos afectan en menor medida a los agrupamientos [Torres Rudloff, 2017]. Es importante una buena elección de parámetros para obtener los agrupamientos adecuados y no puede agrupar datos con densidades muy distintas.

Algunos ejemplos de algoritmos basados en densidad son:

- DBSCAN [Ester et al., 1996]
- OPTICS [Ankerst et al., 1999]
- DenClue [Rehioui et al., 2016]
- Mean-shift [Comaniciu and Meer, 2002]

4.1.1. DBSCAN

Buscaremos implementar el algoritmo de aprendizaje no supervisado llamado DBSCAN (por su acrónimo en inglés Density-Based Spatial Clustering of Applications with Noise) en Python Scikit-Learn. Este algoritmo fue propuesto en 1996 y asume que los agrupamientos son regiones densas en el espacio que están separadas por regiones menos densas. No requiere el número de agrupamientos como entrada. Tiene dos parámetros importantes: eps , minPts . eps es la distancia mínima requerida para que dos puntos sean denominados vecinos, es decir, dos puntos se consideran vecinos cuando la distancia entre ellos es menor o igual a eps . Elegir un valor adecuado es necesario para evitar agrupaciones deficientes. Si su valor es demasiado pequeño, entonces la mayoría de los puntos no estarán en el vecindario y se tratarán como valores atípicos. En cambio, si su valor es demasiado alto, la mayoría de los puntos de datos permanecerán en el mismo agrupamiento. Para elegir un valor adecuado de eps , debe estar en función de la distancia del conjunto de datos. Por otro lado, minPts es el número mínimo de puntos de datos que deben estar en la región para definir el agrupamiento. Se puede elegir en función del conocimiento de su dominio. Si no se conoce, un buen punto de referencia es tener $\text{minPts} \geq D + 1$ donde D es la dimensión del conjunto de datos [Kumar, 2021].

Como se aprecia en Algorithm (1), DBSCAN recibe como entrada los datos y los parámetros eps , minPts . Después, el algoritmo toma dato por dato clasificándolos en tres tipos como se observa en la figura (4.4): central (el que tomado como centro, cuenta con al menos minPts en su región circundante dentro del radio eps), fronterizo (aquel a que está dentro del radio eps con centro en el punto central y tiene menos número de puntos minPts dentro de su región circundante) y ruido (no se encuentra en ninguna región circundante de radio eps de ningún centro) [Kumar, 2021].

La ventaja de DBSCAN es que determina el número de agrupamientos basándose en la forma de los datos (es compatible con prácticamente cualquier forma de agrupamiento) y es poco sensible al ruido. Es por estos motivos que “DBSCAN es el algoritmo

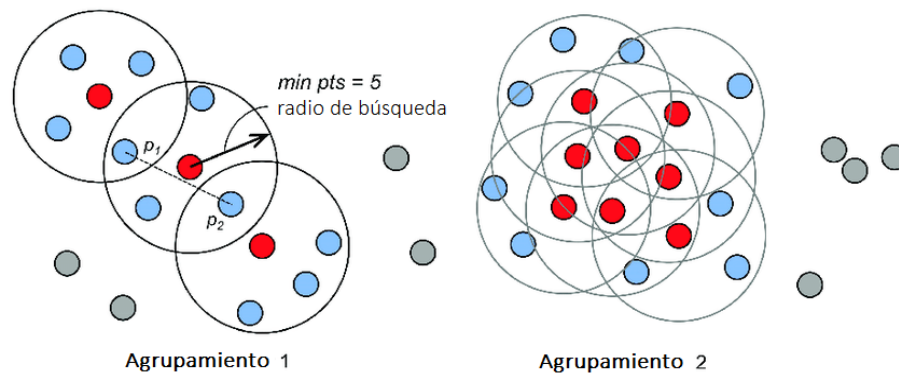


Figura 4.4: Ejemplos de como el algoritmo DBSCAN clasifica los datos. Fuente: [DiFrancesco et al., 2020].

Algorithm 1 DBSCAN

Data: Datos, eps y minPts

for x *in* Datos **do**

if x no ha sido seleccionado **then**

 Determine la vecindad de x con eps

if $x_{vecindario} \geq minPts$ **then**

 Se marca como un punto central y como visitado

else if $x_{vecindario} < minPts$ **then**

if x está junto a un punto central **then**

 Se marca como frontera

else

 Se marca como ruido

end

end

Result: Entrega la etiqueta sobre el agrupamiento que pertenece cada dato

preferido para muchos problemas de agrupamiento como: análisis financiero [Huang et al., 2019, Yang et al., 2014], estudios de mercado [Fan et al., 2021, Wei and Sun, 2019], planeación urbana [Li and Li, 2007, Pavlis et al., 2017], sismología [Fan and Xu, 2019], sistemas de recomendación [Guan et al., 2018, Kuzelewska and Wichowski, 2015], ingeniería genética [Francis et al., 2011, Mohammed et al., 2018], entre otros” [Zhang et al., 2022b].

Particularmente en cosmología el algoritmo de agrupamiento comúnmente utilizado para clasificar cúmulos es FoF (Friends-of-Friends), éste es un algoritmo de agrupamiento basado en densidad muy parecido a DBSCAN. A continuación se muestra un ejemplo en el que se usa DBSCAN para clasificar cúmulos de galaxias. En el artículo [Zhang, 2019] utilizan los resultados de la simulación numérica CoDECS, distancias euclidianas y el algoritmo DBSCAN para clasificar cúmulos de galaxias. En la figura (4.5) se pueden apreciar los datos de galaxias con los resultados correctos sobre a que agrupamiento pertenecen. A continuación eligen 2000/3000/4000 galaxias

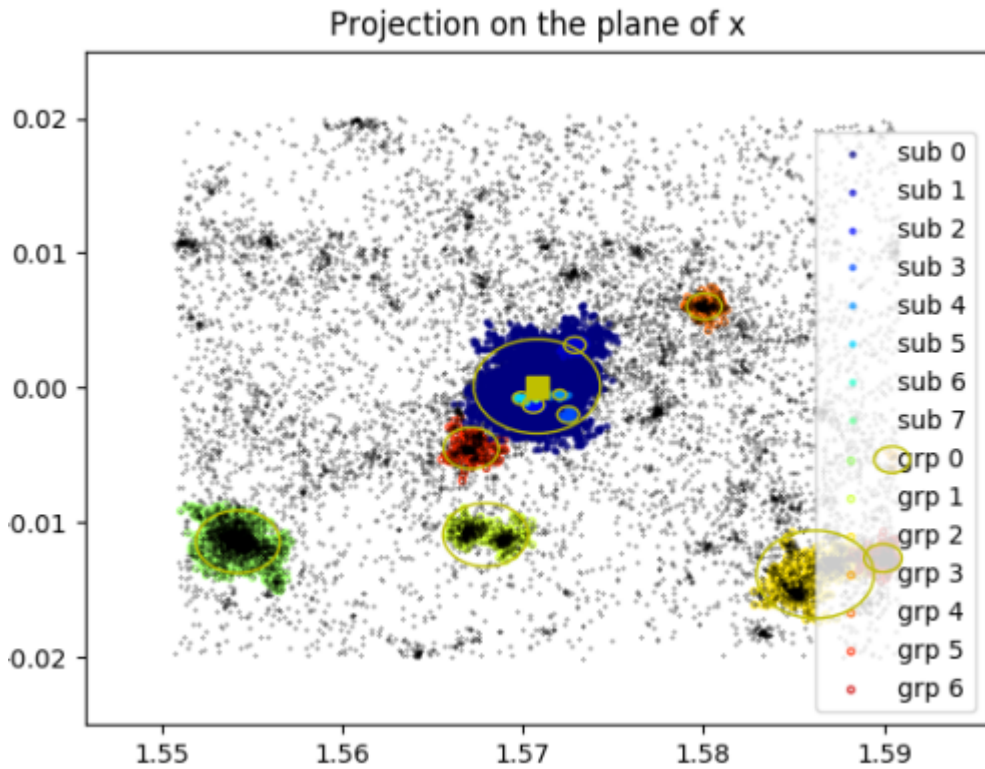


Figura 4.5: Resultados correctos de datos de cúmulos de galaxias analizados en [Zhang, 2019].

aleatorias de esos datos para ser analizados y clasificados con DBSCAN (un ejemplo es la figura (4.6) para posteriormente comparar los resultados correctos con los de DBSCAN. Finalmente obtiene precisiones del 60 % a más de 85 %, mostrando que DBSCAN funciona muy bien para áreas suficientemente densas, pero la tasa de reconocimiento tendrá grandes fluctuaciones para cúmulos con pocas galaxias [Zhang, 2019].

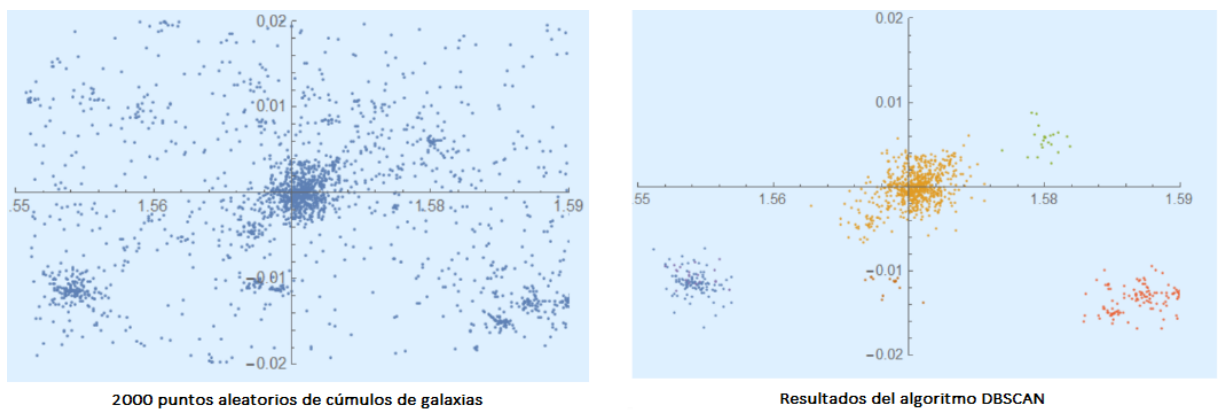


Figura 4.6: 2000 datos aleatorios de la figura (4.5) analizados con DBSCAN. Fuente: [Zhang, 2019].

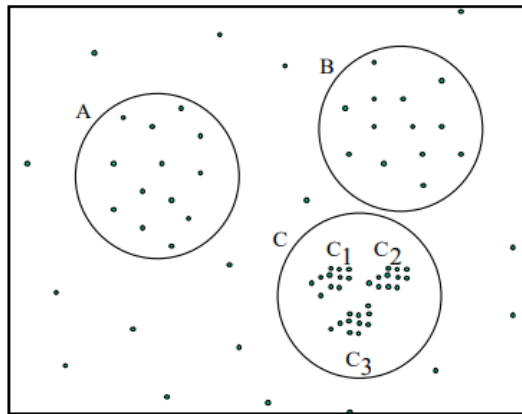


Figura 4.7: Agrupamientos con respecto a distintos parámetros de densidad. Figura presentada en [Ankerst et al., 1999].

4.1.2. OPTICS

La mayoría de los algoritmos requieren como entrada parámetros que son complicados de determinar e influyen bastante en los resultados finales, entregando resultados muy diferentes al modificar mínimamente el parámetro. Aunado a es esto, para gran parte de las bases de datos reales no existe un parámetro global que nos entregue todos los agrupamientos adecuados; un ejemplo esquemático se muestra en la figura (4.7) donde con un parámetro global de densidad sólo se pueden detectar los agrupamientos A, B y C (donde detecta a C_1 , C_2 y C_3 como un sólo agrupamiento), o los agrupamientos C_1 , C_2 y C_3 (para este caso los A y B serían considerados como ruido) [Ankerst et al., 1999].

Una forma de solucionar este problema es utilizar un algoritmo de agrupamiento jerárquico, como el método de enlace simple también conocido como "single-link". En este tipo de algoritmo cada dato se considera un agrupamiento y basándose en la cercanía a otros datos se van juntando, hasta que se tiene un solo agrupamiento o hasta que se cumple un criterio de parada predefinido. El problema con este método es el efecto de enlace simple, es decir que si dos agrupamientos están cerca o los unen pocos datos de ruido, se pueden fusionar incorrectamente. Por otro lado, los resultados que entregan este tipo de algoritmos (llamados dendrogramas) son difíciles de interpretar para grandes cantidades de datos. Otra forma en la que se puede resolver el problema es con algoritmos basados en densidad con diferentes configuraciones de parámetros sin embargo existe un número infinito de valores para estos y podríamos pasar por alto el adecuado. Es por esto que se propone un algoritmo jerárquico basado en densidad, el cual entregue información de cada nivel de agrupamiento hasta una cierta distancia ϵ [Ankerst et al., 1999].

En 1999 se propone el algoritmo OPTICS (por su acrónimo en inglés Ordering Points To Identify the Clustering Structure), el cual es importante resaltar que no entregue como resultado el agrupamiento al que pertenece cada dato, sino que crea una lista

ordenada de la base de datos que representa la estructura de los agrupamientos basada en densidad, con la distancia núcleo/objeto central y la distancia de alcanzabilidad de cada objeto (éstas se definirán más adelante) [Ankerst et al., 1999]. Análogamente a DBSCAN, OPTICS utiliza los parámetros ε y $MinPts \in \mathbb{N}$ con la diferencia de que ε (llamado "distancia generadora") será el radio máximo, es decir, buscará el menor radio entre $0 < \varepsilon_i \leq \varepsilon$ que cumpla con tener un objeto central, de manera que funciona como un DBSCAN extendido/mejorado debido a que es menos sensible a los parámetros iniciales [Torres Rudloff, 2017, Ankerst et al., 1999]. En OPTICS se considera un **objeto central** al objeto que cumple que $Card(N_\varepsilon(q)) \geq MinPts$ ($Card(N)$ denota la cardinalidad² del conjunto N), es decir, dentro de un radio ε con centro en el objeto dado, haya por lo menos $MinPts$ objetos [Ankerst et al., 1999].

Para explicar el funcionamiento de OPTICS, en [Ankerst et al., 1999] nos dan las siguientes definiciones, las cuales recomiendo ampliamente que se lean con detenimiento para una fácil comprensión del algoritmo.

Def 4.1.1: Directamente alcanzable por densidad (Directly density-reachable)

Un objeto p es directamente alcanzable por densidad desde un objeto q bajo los parámetros ε y $MinPts$ en un conjunto de datos D si:

1. $p \in N_\varepsilon(q)$ ($N_\varepsilon(q)$ es el subconjunto de D contenido en el vecindario ε de q)
2. $Card(N_\varepsilon(q)) \geq MinPts$ ^a

^aSólo desde objetos centrales, otros objetos pueden ser directamente alcanzables por densidad.

Def 4.1.2: Alcanzable por densidad (Density-reachable)

Un objeto p es alcanzable por densidad desde un objeto q bajo los parámetros ε y $MinPts$ en un conjunto de datos D si existe una cadena de objetos que pertenecen a D $p_1 = q, \dots, p_i, p_{i+1}, \dots, p_n = p$ tal que p_{i+1} es **directamente alcanzable por densidad** desde p_i .^a

^aComo se aprecia en la figura (4.9), esta relación no es simétrica, notemos que sólo dos objetos centrales pueden ser mutuamente alcanzables por densidad.

Def 4.1.3: Conectado por densidad (Density-connected)

Un objeto p está conectado por densidad desde un objeto q bajo los parámetros ε y $MinPts$ en un conjunto de datos D si existe un objeto $o \in D$ tal que p y q son **alcanzables por densidad** desde o .^a

^aComo se aprecia en la figura (4.9), esta relación es simétrica.

²La cardinalidad se refiere al número de elementos en un conjunto.

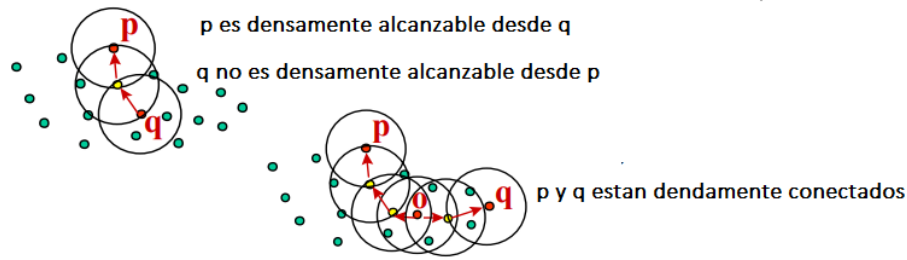


Figura 4.8: Diferencia entre *density-reachable* y *density-connected*. Figura presentada en [Ankerst et al., 1999].

En la figura (4.8) se muestran puntos alcanzables por densidad y conectados por densidad para apreciar mejor la diferencia.

Def 4.1.4: Agrupamiento y ruido

Dado un conjunto de datos D , bajo los parámetros ϵ y $MinPts$ decimos que C es un agrupamiento si cumple:

1. $\forall p, q \in D$: si $p \in C$ y q es alcanzable por densidad, $\Rightarrow q \in C$.
2. $\forall p, q \in C$: p está conectado por densidad a q .

Todo objeto que no pertenezca a un agrupamiento se considera ruido.

Def 4.1.5: Distancia central de un objeto p (Core distance of an object p)

Dado un conjunto de datos D , ϵ y $MinPts$. Definimos $N_\epsilon(p)$ como la ϵ -vecindad de p con $p \in D$ y $MinPts$ -distance(p) como la distancia radial mínima ϵ_i con centro en p hasta donde en su vecindario ($N_\epsilon(p)$) tenga contenido $MinPts$ objetos.

$$distancia-central_{\epsilon, MinPts}(p) = \begin{cases} \text{Indefinido,} & \text{si } Card(N_\epsilon(p)) < MinPts \\ MinPts-distance(p), & \text{otro caso} \end{cases} \quad (4.5)$$

Def 4.1.6: Distancia de alcanzabilidad (Reachability-distance)

Dado un conjunto de datos D , los objetos $p, o \in D$ y los parámetros ϵ y $MinPts$. La distancia de alcanzabilidad (dA) de p con respecto a o se define como:

$$dA_{\epsilon, MinPts}(p, o) = \begin{cases} \text{Indefinido,} & \text{si } |N_\epsilon(o)| < MinPts \\ \text{máx}(distancia-central(o), distancia(o, p)), & \text{otro caso} \end{cases}$$

En la figura (4.9) se muestra gráficamente ejemplos de *distancia central* (core-distance) y *alcanzable por densidad* (reachability-distance).

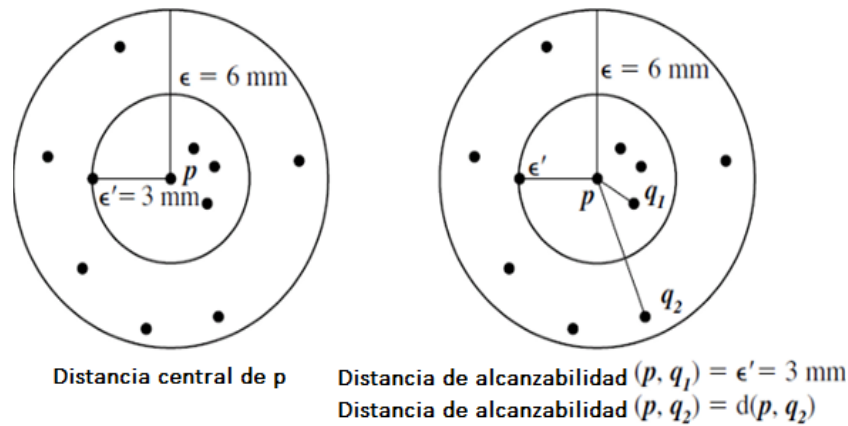


Figura 4.9: Diferencia entre distancia central (core-distance) y reachability-distance. Figura presentada en [Sinclair, 2019].

Ahora, procedemos a explicar el funcionamiento de OPTICS con ayuda de los pseudocódigos proporcionados por [Ankerst et al., 1999]. Se dividió el funcionamiento total de OPTICS en cuatro pseudocódigos para facilitar su comprensión. Comenzamos con el algoritmo 2 OPTICS central, éste pide como entrada: la base de datos de los objetos a analizar (SetOfObjects), la distancia generadora (ϵ), el mínimo de objetos (MinPts) y un archivo (OrderedFile). A continuación abre el archivo OrderedFile, luego para cada objeto en SetOfObjects se checa si ha sido procesado, en caso de no haber sido procesado aún, se analizará con el algoritmo 3 ExpandClusterOrder [Ankerst et al., 1999].

Algorithm 2 OPTICS [Ankerst et al., 1999]

Data: SetOfObjects, ϵ , MinPts, OrderedFile

OrderedFile.open()

for i from 1 to SetOfObjects.size **do**

 Object := SetOfObjects.get(i)

if NOT Object.Processed **then**

 ExpandClusterOrder(SetOfObjects, Object, ϵ , MinPts, OrderedFile)

else

end

OrderedFile.close()

Result: Entrega una lista ordenada de los objetos en la base de datos SetOfObjects con sus correspondientes distancias de alcanzabilidad y distancias al centro del agrupamiento

El algoritmo 3 ExpandClusterOrder, recibe como entrada los mismos datos que el algoritmo 2 añadiendo el dato Object que es el objeto seleccionado por el algoritmo 2. A continuación obtiene los vecinos de Object, marca a Object como procesado, establece su distancia central como indefinida y calcula las distancias centrales (core distance) de sus vecinos. Luego, se escribe el objeto Object en OrderedFile. A continuación si el objeto no tiene distancia central (core distance) indefinida quiere decir que es un objeto central para la distancia generada dada y se procesan los vecinos del objeto y

Algorithm 3 ExpandClusterOrder [Ankerst et al., 1999]

```
Data: SetOfObjects, Objectc,  $\epsilon$ , MinPts, OrderedFile
neighbors := SetOfObjects.neighbors(Object,  $\epsilon$ );
Object.Processed := TRUE
Object.reachability_distance := UNDEFINED
Object.setCoreDistance(neighbors,  $\epsilon$ , MinPts)
OrderedFile.write(Object)
if Object.core_distance != UNDEFINED then
  OrderSeeds.update(neighbors, Object)
  while NOT OrderSeeds.empty() do
    currentObject := OrderSeeds.next()
    neighbors:=SetOfObjects.neighbors(currentObject,  $\epsilon$ )
    currentObject.Processed := TRUE
    currentObject.setCoreDistance(neighbors,  $\epsilon$ , MinPts)
    OrderedFile.write(currentObject)
  end
  if currentObject.core_distance != UNDEFINED then
    | OrderSeeds.update(neighbors, currentObject)}
  else
else
end
```

el objeto `Object` con el algoritmo 4 `OrderSeeds::update`. Por el contrario si el objeto no es un objeto central se regresa al bucle de OPTICS y se selecciona el siguiente objeto no procesado. Más adelante, en el bucle WHILE mientras no este vacía la lista `OrderSeeds`, se selecciona el objeto `currentObject` con la menor distancia de alcanzabilidad en la lista, se determina su vecinario, su distancia central, se marca como procesado y se escribe en la lista `OrderedFile`, si el objeto no tiene distancia central (core distance) indefinida quiere decir que es un objeto central para la distancia generada dada y se procesan los vecinos del objeto y el objeto `Object` con el algoritmo 4 `OrderSeeds::update` [Ankerst et al., 1999].

El algoritmo 4 `OrderSeeds::update`, toma como entrada los vecinos y el objeto `CenterObject`, éste maneja la lista `OrderSeeds` y las distancias de alcanzabilidad. Para todos los objetos en el vecindario de `CenterObject` checa que no haya sido procesado, en caso de no haber sido procesado recopila de manera iterativa los objetos directamente alcanzables por densidad alcanzables y se anotan en la lista `OrderSeeds` y se van ordenando según su distancia de alcanzabilidad al objeto central más cercano. Para los objetos que aún no están en la lista `OrderSeeds` se escribe su distancia de alcanzabilidad directamente, para los objetos que ya pertenecían a la lista se mueven hacia la parte superior en caso de que su nueva distancia de alcanzabilidad sea menor a la anterior [Ankerst et al., 1999].

Algorithm 4 OrderSeeds::update [Ankerst et al., 1999]

Data: neighbors, CenterObject
c_dist := CenterObject.core_distance
forall Object FROM neighbors **do**
 if NOT Object.Processed **then**
 new_r_dist:=max(c_dist,CenterObject.dist(Object))
 if Object.reachability_distance=UNDEFINED **then**
 Object.reachability_distance := new_r_dist
 insert(Object, new_r_dist)
 else
 Object already in OrderSeeds
 if new_r_dist<Object.reachability_distance **then**
 Object.reachability_distance := new_r_dist
 decrease(Object, new_r_dist)
 else
 end
 end
 else
end

El algoritmo 5 ExtractDBSCAN-Clustering generó la lista ordenada de distancias de alcanzabilidad y distancia al centro de cada objeto de la base de datos, con esto podemos extraer el agrupamiento al que presuntamente pertenece cada objeto basándonos en densidad. Para todos los objetos en ClusterOrderedObjs checamos si la distancia de alcanzabilidad es mayor que la distancia del agrupamiento ϵ' , en este caso el objeto no es alcanzable por densidad desde ninguno de los objetos que están antes del objeto actual en el ordenamiento y procedemos a comprobar el valor de su distancia central en caso de ser menor o igual que ϵ' entonces se crea un nuevo agrupamiento de lo contrario se define al objeto como ruido. Ahora, si la distancia de alcanzabilidad del objeto es menor o igual que ϵ' , se asigna el objeto al agrupamiento actual [Ankerst et al., 1999].

Algorithm 5 ExtractDBSCAN-Clustering [Ankerst et al., 1999]

Data: ClusterOrderedObjs, ϵ' , MinPts
// Precondition: $\epsilon' \leq$ generating dist ϵ for ClusterOrderedObjs
ClusterId := NOISE
for i FROM 1 TO ClusterOrderedObjs.size **do**
 Object := ClusterOrderedObjs.get(i)
 if Object.reachability_distance $> \epsilon'$ **then**
 if Object.core_distance $\leq \epsilon'$ **then**
 ClusterId := nextId(ClusterId)
 Object.clusterId := ClusterId
 else
 Object.clusterId := NOISE
 end
 else
 Object.clusterId := ClusterId;
 end
end

Una forma de analizar los resultados de OPTICS es con la gráfica de alcanzabilidad (reachability plot). Donde en el eje x se muestran los objetos ordenados (como los ordenó OPTICS) y en el eje y se presenta su alcanzabilidad, en este tipo de gráficas los agrupamientos serían las depresiones en la gráfica. Por ejemplo, en la figura (4.10) el agrupamiento abarcaría del objeto 3 al 16, cabe mencionar que el objeto 3 se considera parte del agrupamiento a pesar de presentar una alta alcanzabilidad debido a que, como OPTICS ordena los objetos, ésta es la alcanzabilidad del objeto 2 al 3, entonces podemos ver que la alcanzabilidad del objeto 4 al 3 es baja. Es importante notar que en la figura (4.10) se aprecia perfectamente el inicio y fin del agrupamiento, sin embargo en datos tomados del mundo real no siempre es tan marcado el inicio y fin de un agrupamiento [Ankerst et al., 1999].

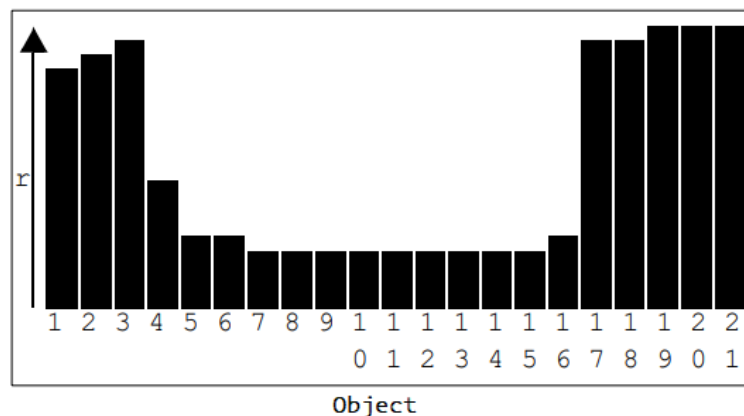


Figura 4.10: Gráfica de alcance (reachability plot). Fuente: [Ankerst et al., 1999].

4.1.3. HDBSCAN

En el artículo [Campello, 2013], publicado en 2013, proponen un método de agrupamiento jerárquico basado en densidad llamado HDBSCAN (por su acrónimo en inglés: Hierarchical Density-based spatial clustering of applications with noise) el cual genera una jerarquía de agrupamientos basada en densidad a partir de la que se puede extraer una jerarquía simplificada compuesta por los agrupamientos más significativos. Métodos de agrupamiento como DBSCAN o DENCLUE sólo proporcionan una etiquetación de datos “plana” (no jerárquica) basada en un parámetro de densidad global, sin embargo como vimos en la sección del algoritmo OPTICS, no es ideal considerar un parámetro de densidad global. Para esto, en HDBSCAN proponen una medida de estabilidad del agrupamiento para extraer los más significativos de diferentes posibles niveles jerárquicos. HDBSCAN se puede ver como una mejora del algoritmo OPTICS, éste pide sólo un valor de entrada, el cual es el mínimo de puntos $minPts$ para considerar un agrupamiento ya que diferentes niveles de densidad en la jerarquía resultante corresponderán a diferentes valores del radio ϵ [Campello, 2013].

Para explicar el funcionamiento de HDBSCAN, en [Campello, 2013] nos dan las siguientes definiciones, las cuales recomiendo ampliamente que se lean con detenimiento para una fácil comprensión del algoritmo.

Def 4.1.7: Distancia central (Core distance)

Dado un conjunto de datos X , ϵ y $MinPts$. Definimos la distancia central $d_{core}(x_p)$ como la distancia radial mínima con centro en x_p hasta donde en su vecindario tenga contenido $MinPts$ objetos.

Def 4.1.8: Distancia de alcanzabilidad mutua (Mutual Reachability Distance)

Dado un conjunto de datos X , los objetos $p, q \in X$ y los parámetros ϵ y $MinPts$. La distancia de alcanzabilidad mutua de p con respecto a q se define como:

$$d_{mreach}(p, q) = \max\{d_{core}(p), d_{core}(q), d(p, q)^a\}$$

^aDonde $d(p, q)$ denota la distancia “normal” bajo la métrica elegida, por ejemplo: Euclídeana.

De la definición 4.1.8 podemos notar que separan los puntos dispersos de los demás al menos por su distancia central, haciendo que sea más robusto frente al ruido. También definen la gráfica de alcanzabilidad mutua (Mutual Reachability Graph, MRG), en la cual cada objeto de la base de datos X son vértices conectados por aristas cuyo peso es la distancia de alcanzabilidad mutua entre el respectivo par de objetos. Con esta gráfica podemos construir un árbol de expansión mínimo (minimum spanning tree, MST³) como se muestra en la figura (4.11) y al ordenar sus aristas por la distancia de

³Es un subconjunto de un gráfico, que tiene todos los vértices cubiertos con el mínimo número posible de aristas con el menor peso posible.

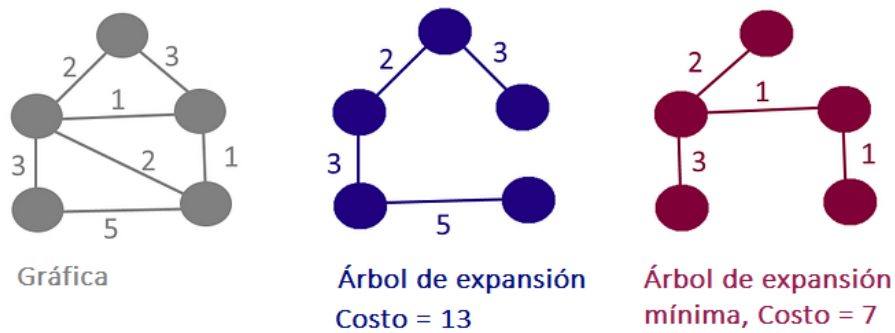


Figura 4.11: Ejemplo donde la primer figura muestra una gráfica de alcanzabilidad mutua con los objetos como vértices y se muestra el peso de cada arista. La segunda figura muestra un árbol de expansión (spanning tree) de la primer figura. Finalmente, la última figura muestra el árbol de expansión mínima (minimum spanning tree) de la primer figura. Fuente [Jain, 2022].

alcanzabilidad mutua, se obtiene una estructura jerárquica en forma de árbol, llamada dendrograma⁴. En la figura (4.13) se muestra un ejemplo esquemático de como se construye un dendrograma. HDBSCAN considera una división de agrupamiento como verdadera sólo si ambos subagrupamientos contienen al menos minPts objetos; si ninguno de los dos subagrupamientos cumple con tener por lo menos minPts objetos, entonces se considera que en ese nivel de densidad desaparece el agrupamiento. Por otro lado, si sólo uno de los dos subagrupamiento cumple con tener al menos minPts objetos, entonces se considera que es el mismo agrupamiento anterior sólo que se perdieron objetos de este en forma de ruido [Malzer and Baum, 2020].

A continuación analizaremos el algoritmo (6) el cual es el pseudocódigo de HDBSCAN proporcionado por [Campello, 2013]. Comienza calculando la distancia central para todos los objetos de la base de datos proporcionada. Luego calcula la gráfica de alcanzabilidad mutua (MRG) de la cual obtiene el árbol de expansión mínima (MST). A partir de esto crea una extensión de un árbol de expansión mínima (MST_{ext}) donde añaden autoaristas que conectan cada vértice consigo mismo y a estas autoaristas se les asigna la distancia central del vértice como peso. Después se extrae la jerarquía como un dendrograma a partir de MST_{ext} . Comenzamos etiquetando todos los objetos como pertenecientes al mismo agrupamiento. Posteriormente iteramos para eliminar aristas en orden decreciente y para esto se establece el valor del nivel jerárquico del dendrograma actual como el peso de las aristas que se eliminaran. Al terminar de eliminar las aristas se etiquetan los datos como el agrupamiento al que pertenecen, para esto se checa si el objeto aún tiene una arista, y en ese caso se etiqueta como parte de un agrupamiento; en caso de no tener arista que la conecte se considera ruido [Campello, 2013].

⁴Gráfico en forma de árbol invertido el cual muestra los datos agrupados jerárquicamente de forma divisiva.

Algorithm 6 HDBSCAN [Campello, 2013]

Data: X =conjunto de datos, MinPts

forall $\text{objects} \in X$ **do**

1. Calcular la distancia central d_{core}

2. Calcular el MST de la gráfica de alcanzabilidad mutua (MRG).

3. Se extiende el MST para obtener MST_{ext} , agregando para cada vértice una “autoarista” con la distancia central del objeto correspondiente como peso.

4. Se extrae la jerarquía de HDBSCAN como un dendrograma a partir de MST_{ext} :

for $x \in X$ **do**

4.1. Se le asigna a todos los objetos la misma etiqueta (teniendo un sólo agrupamiento en la raíz del árbol).

4.2 Iterativamente, eliminar todas las aristas de MST_{ext} en orden decreciente de pesos (en caso de empates, las aristas deben eliminarse simultáneamente):

4.2.1 Antes de cada eliminación, establecer el valor de escala del dendrograma del nivel jerárquico actual como el peso de la(s) arista(s) que se van a eliminar.

4.2.2 Después de cada eliminación, asignar etiquetas a la(s) componente(s) conectada(s) que contenga(n) el (los) vértice(s) final(es) de la(s) arista(s) eliminada(s), para obtener el siguiente nivel jerárquico:

if todavía tiene al menos una arista **then**

| asignar una nueva etiqueta de agrupamiento a una componente

else

| asignarle una etiqueta nula (ruido)

end

end

end

Result: Entrega un árbol de agrupamiento que contiene todas las particiones obtenidas con respecto a minPts de manera jerárquica y anidada y su solución plana

Es importante recalcar que extraer los agrupamientos significativos de un dendrograma no es tarea fácil, por ello propusieron una simplificación de la jerarquía del HDBSCAN en la que se toma en cuenta las tres posibilidades para la evolución de los objetos conectados al aumentar el nivel de densidad, es decir, disminuir ϵ : el agrupamiento se reduce perdiendo objetos en los bordes en forma de ruido, el agrupamiento se divide en subagrupamientos o el agrupamiento simplemente desaparece. Por lo que se modifica el paso 4.2.2 para obtener los siguientes niveles jerárquicos como se muestra en el algoritmo 7. Considerando estas evoluciones en HDBSCAN se seleccionan sólo los niveles jerárquicos en los que surgen nuevos agrupamientos o en los que desaparecen debido a que son los cambios más significativos, para los casos en donde el agrupamiento sólo pierde datos se seguirá considerando el mismo

Algorithm 7 HDBSCAN paso 4.2.2 con el parámetro opcional $m_{clSize} \geq 1$ [Campello, 2013]

4.2.2 Después de cada eliminación, procesar uno por uno cada agrupamiento que contenía la(s) arista(s) recién eliminada(s), mediante el etiquetado de sus subcomponentes conectados resultantes:

if Si hay subcomponentes que no se consideran significativas en el agrupamiento y se desconectan **then**

 Etiquetarlas como ruido

if Si todas las subcomponentes son marcadas como ruido **then**

 | El agrupamiento desaparece

else if Si hay una subcomponente que no es marcada como ruido **then**

 | Se conserva su etiqueta original (el agrupamiento disminuyó su tamaño)

else if Si hay dos o más subcomponentes que no fueron marcadas como ruido **then**

 | Se asignan nuevas etiquetas de agrupamiento a cada una (el agrupamiento se dividió)

else

end

agrupamiento y se agrega el parámetro opcional m_{clSize} que establece el tamaño mínimo de un agrupamiento para que se considere como válido con la finalidad de suavizar el árbol resultante [Campello, 2013].

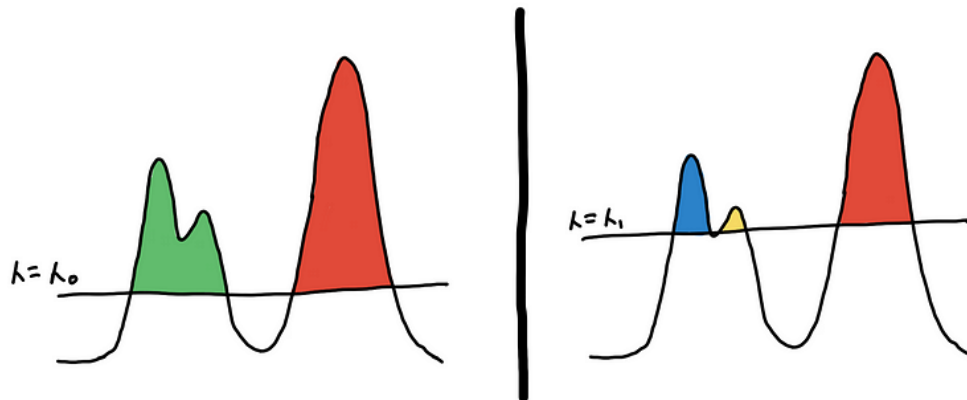


Figura 4.12: Ejemplo de como se puede determinar los agrupamientos utilizando la función de densidad de probabilidad dependiendo del parámetro global que se utilice. Fuente: [Berba, 2020].

Ahora, dado un árbol podemos simplemente tomar todos los nodos hoja como los agrupamientos finales, es decir, los ϵ con valores más bajos en la jerarquía que cumplan con $MinPts$, sin embargo HDBSCAN considera otra opción a la cual llama “eom” (por su acrónimo en inglés, Excess Of Mass), basándose en una observación de las funciones de densidad de probabilidad (pdf) de valores continuos. En la figura (4.12) nos muestran dos resultados de agrupamientos que dependen del umbral global que se elija (de eso depende si el algoritmo lo considera 2 agrupamientos o 3). Sin embargo, un ejemplo de lo que hace HDBSCAN es evitar usar un parámetro global,

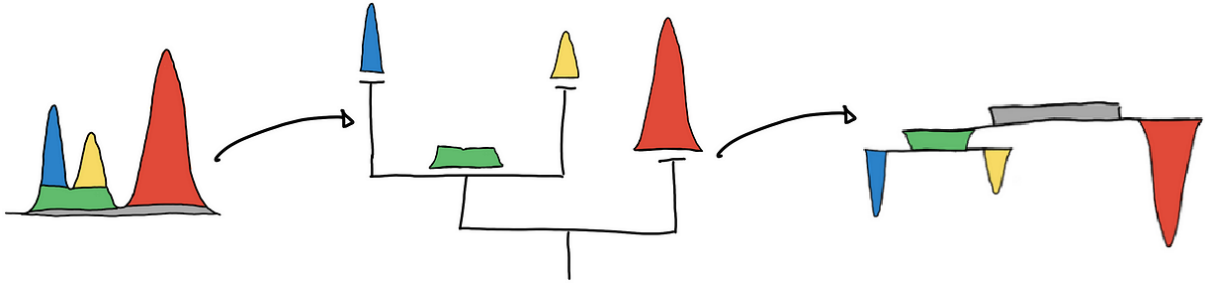


Figura 4.13: La primera imagen es un ejemplo de cómo se ve una función de densidad de probabilidad (pdf). La segunda imagen muestra cómo se dividen los datos en diferentes agrupamientos. La última imagen muestra cómo se vería el dendrograma. Fuente: [Berba, 2020].

así checa la estabilidad del agrupamiento, por ejemplo en ese caso el área de la densidad de probabilidad del verde (sin la parte del azul y amarillo) es mayor por lo que se consideraría ese agrupamiento como más estable y sería seleccionado. Así que la propuesta es encontrar los agrupamientos más estables, para esto definen agrupamientos estables como [Malzer and Baum, 2020],

$$S(\mathbf{C}_i) = \sum_{x_j \in \mathbf{C}_i} (\lambda_{\max}(x_j, \mathbf{C}_i) - \lambda_{\min}(\mathbf{C}_i)) = \sum_{x_j \in \mathbf{C}_i} \left(\frac{1}{\varepsilon_{\min}(x_j, \mathbf{C}_i)} - \frac{1}{\varepsilon_{\max}(\mathbf{C}_i)} \right). \quad (4.6)$$

Donde la densidad está dada por $\lambda = \frac{1}{\varepsilon}$, λ_{\min} es el nivel de densidad en que el agrupamiento aparece por primera vez y λ_{\max} es el último nivel en el que aparece. Esto nos proporciona información de la vida del agrupamiento y entre más vida tenga, mayor estabilidad tendrá. Luego, sean $\{\mathbf{C}_2, \dots, \mathbf{C}_k\}$ todos los agrupamientos en el árbol jerárquico simplificado (exceptuando \mathbf{C}_1 que es el tronco que contiene todos los objetos); para obtener un resultado plano de los agrupamientos más significativos y el ruido buscan maximizar las estabilidades de los agrupamientos de la siguiente forma [Campello, 2013]:

$$\max_{\delta_2, \dots, \delta_k} J = \sum_{i=2}^k \delta_i S(\mathbf{C}_i) \quad \text{con} \quad \begin{cases} \delta_i \in \{0, 1\}, & i = 2, \dots, k, \\ \sum_{j \in \mathbf{I}_h} \delta_j, & \forall h \in \mathbf{L}, \end{cases}$$

donde δ_i es booleano e indica si el i -ésimo agrupamiento fue seleccionado en el resultado plano ($\delta_i = 1$) o no ($\delta_i = 0$), \mathbf{L} es el índice de los agrupamientos más finos (hojas) y \mathbf{I}_h son los índices desde el agrupamiento hojas hasta el último antes de llegar al tronco. A continuación, el algoritmo de selección de HDBSCAN, como se muestra en el algoritmo 8, comienza estableciendo todos los nodos como seleccionados; luego recorre el árbol de abajo hacia arriba y compara la estabilidad de cada nodo con la suma de estabilidad de sus subagrupamientos, de tal forma que las estabilidades se propagan y se actualizan al subir por el árbol hasta encontrar y seleccionar el agrupamiento con la mayor estabilidad de cada rama. Para lograr esto localmente

Algorithm 8 HDBSCAN solución plana [Campello, 2013]

Establece $\delta_2 = \dots = \delta_k = 1$

forall Agrupamientos más finos (hoja) **do**

 | $\hat{S}(\mathbf{C}_h) = S(\mathbf{C}_h)$

end

Iniciando desde los niveles más bajos hacia el tronco:

if $S(\mathbf{C}_i) < \hat{S}(\mathbf{C}_{il}) + \hat{S}(\mathbf{C}_{ir})$ **then**

 | Asigna $\hat{S}(\mathbf{C}_i) = \hat{S}(\mathbf{C}_{il}) + \hat{S}(\mathbf{C}_{ir})$ y asigna $\delta_i = 0$

else

 | Asigna $\hat{S}(\mathbf{C}_i) = S(\mathbf{C}_i)$ y para todos los subagrupamientos ramificados abajo de este
 | asigna $\delta = 0$

end

en \mathbf{C}_i , actualizamos la estabilidad total $\hat{\mathbf{C}}_i$ de los agrupamientos seleccionados en el subárbol con raíz en \mathbf{C}_i de la siguiente forma [Campello, 2013]:

$$\hat{S}(\mathbf{C}_i) = \begin{cases} \hat{S}(\mathbf{C}_i), & \text{si } \mathbf{C}_i \text{ es un nodo hoja (último),} \\ \max\{S(\mathbf{C}_i), \hat{S}(\mathbf{C}_{il}) + \hat{S}(\mathbf{C}_{ir})\}, & \text{si } \mathbf{C}_i \text{ es un nodo interno.} \end{cases} \quad (4.7)$$

En este ejemplo escogen un árbol binario en el que cada los agrupamientos se dividen sólo en 2 subagrupamientos a los que llaman izquierdo \mathbf{C}_{il} y derecho \mathbf{C}_{ir} sin embargo el algoritmo funciona para N subagrupamientos.

4.1.4. DRL-DBSCAN

DRL-DBSCAN (por su acrónimo en inglés: "Deep Reinforcement Learning - Density-Based Spatial Clustering of Applications with Noise") fue propuesto en 2022 en el artículo [Zhang et al., 2022b], y busca parámetros óptimos para DBSCAN mediante Aprendizaje por Refuerzo Profundo Recursivo.

Recordemos que las ventajas de DBSCAN son que puede determinar la cantidad de agrupamientos (de diferentes formas) para datos con distintos tipos de distribuciones y es poco sensible al ruido, por lo que considero el algoritmo ideal para nuestro problema, sin embargo es bastante sensible a sus parámetros de entrada (eps y MinPts) es por eso que lo óptimo es buscar mejoras de éste como OPTICS o HDBSCAN, sin embargo ambos siguen requiriendo por lo menos un parámetro de entrada, por lo que con DRL-DBSCAN proponen encontrar estos parámetros con aprendizaje profundo por reforzamiento, con el objetivo de no requerir asistencia manual para ello, no requerir datos etiquetados y encontrar los parámetros óptimos, probando la menor cantidad de parámetros para evitar un alto costo computacional. A grandes rasgos, para obtener los parámetros óptimos utilizan un proceso de decisión de Markov (MDP), donde se considera a las modificaciones de los agrupamientos en cada paso como

el estado, luego utiliza la dirección de ajustes como la acción y los agentes de DRL perciben autónomamente el entorno para tomar decisiones. Posteriormente mediante supervisión débil, se construye una recompensa basada en un pequeño número de índices de agrupamiento externos. Para mejorar la eficiencia de aprendizaje y encontrar agrupamientos más estables, utilizan un mecanismo recursivo basado en agentes con diferentes precisiones de búsqueda. En este artículo trabajan con datos Online y Offline que pueden ser con o sin etiquetar con datos iniciales o incrementales en el caso Online, por lo que diseñaron cuatro modos de trabajo los cuales son: modo de entrenamiento, modo de entrenamiento continuo, modo de prueba de pre-entrenamiento y modo de prueba de mantenimiento [Zhang et al., 2022b, He, 2021b]. En éste trabajo no nos adentraremos en el funcionamiento del algoritmo DRL-DBSCAN sin embargo nos pareció interesante mencionarlo ya que en el futuro se espera trabajar con éste.

Capítulo 5

Cálculo de funciones de correlación de dos puntos

En este trabajo “simulamos” las posiciones de galaxias que, como se explicó en el capítulo 1, en promedio se agrupan alrededor de esferas debido a las Oscilaciones Acústicas de Bariones, también conocido como BAO. Nuestra simulación será en 2 dimensiones representando un corte del Universo, suponiendo que tiene curvatura cero. Sabemos que el radio de la esfera fue cambiando con el tiempo hasta el día de hoy, actualmente mide 150Mpc, también sabemos que diferentes corrimientos al rojo representan diferentes tiempos por lo que el radio dependerá del corrimiento al rojo. En nuestra simulación se podrá proporcionar el tamaño del radio, es decir se presentará la simulación de galaxias para un sólo corrimiento al rojo.

A continuación, con el fin de obtener información de las distribuciones de galaxias, se analizarán diferentes estimadores para encontrar la función de correlación de dos puntos de la simulación, después se utilizará un sólo estimador con diferentes configuraciones de las galaxias. Posteriormente, se explorará la posibilidad de que dichos patrones puedan ser identificados con técnicas de machine learning, en particular con algoritmos de agrupamiento. Se utilizó el siguiente código ¹, el cual es libre y se puede usar fácilmente.

5.1. DISTRIBUCIÓN DE GALAXIAS

A continuación programamos una función a la cual nombramos `Puntos_circulos` le entregamos los datos que se muestran en el siguiente renglón de código:

```
1 def Puntos_circulos(ncentros, ncircle_points, radio, tcaja, pert, pcenter=0.1, disp_cent=3.0):
2     points_center = int(ncircle_points*pcenter)
3     ran_centros = np.random.rand(ncentros, 2)*tcaja
4     x_1, y_1 = ran_centros.T[0], ran_centros.T[1]
5     pts = 0
6     circulo = []
```

¹https://github.com/SamanthaRizo/Tesis/blob/main/2pcf_clustering.ipynb

```

7     label = []
8     for i in range(ncentros):
9         cx = []
10        cy = []
11        ran_pcircles = np.random.uniform(0, 2*np.pi, ncircle_points)
12        for j in ran_pcircles:
13            r = radio + random.uniform(-radio*pert, radio*pert) #Pertubaci n
14            c_1 = x_1[i] + r*np.cos(j)
15            c_2 = y_1[i] + r*np.sin(j)
16            #Como los centros de los circulos pueden estar en la orilla, muchos
17            #puntos del ciculo deseado se pueden salir de area a estudiar deseada,
18            #por lo que solo guardamos los que quedan dentro
19            if (0 < c_1 < tcaja) & (0 < c_2 < tcaja):
20                cx.append(c_1)
21                cy.append(c_2)
22                label.append(i)
23                pts+= 1
24            #Agregamos puntos en el centro de los circulos con distribucion gaussiana
25            x = []
26            y = []
27            x = np.random.normal(x_1[i], disp_cent, points_center)
28            y = np.random.normal(y_1[i], disp_cent, points_center)
29            for k in range(points_center):
30                if (0 < x[k] < tcaja) & (0 < y[k] < tcaja):
31                    cx.append(x[k])
32                    cy.append(y[k])
33                    label.append(i)
34                    pts+= 1
35            # en caso de querer graficar un circulo en particular
36            circulo.append(list(zip(cx, cy)))
37        print('total puntos', pts)
38    return circulo, label

```

Donde **ncentros** son la cantidad de circunferencias, **ncircle_points** son la cantidad de galaxias que tendrá cada circunferencia, **radio** es el radio de las circunferencias, **tcaja** es el tamaño de la caja/Universo en la que se encuentran (elegimos cajas cuadradas entre $[0, R]$) y **pert** es el tamaño de la perturbación. Con perturbación nos referiremos al porcentaje del radio que una galaxia puede estar alejado de la circunferencia. Aunado a esto, hay dos variables que son opcionales: **pcenter** que es el porcentaje del número de galaxias en el centro respecto a las de la circunferencia (cuyo valor predeterminado es 0.01 que simboliza el 1%) y **disp_cent** es la dispersión en la distribución normal de puntos en el centro (cuyo valor predeterminado es 3 que corresponde a la desviación estándar). A partir de estos datos, siguiendo los pasos que se muestran en el diagrama de la figura (5.1), nos entrega una lista de las coordenadas de las galaxias de cada circunferencia y en los centros de cada circunferencia, así como la etiqueta de a que circunferencia pertenecen, de esta forma podemos graficar y analizar los datos de todas las galaxias o elegir ciertas circunferencias.

Debido a que la función **Puntos_circulos** entrega como primer salida el conjunto de coordenadas de puntos que pertenecen a cada circunferencia, programamos la siguiente función para obtener todas las coordenadas de puntos en una sola lista de dos columnas.

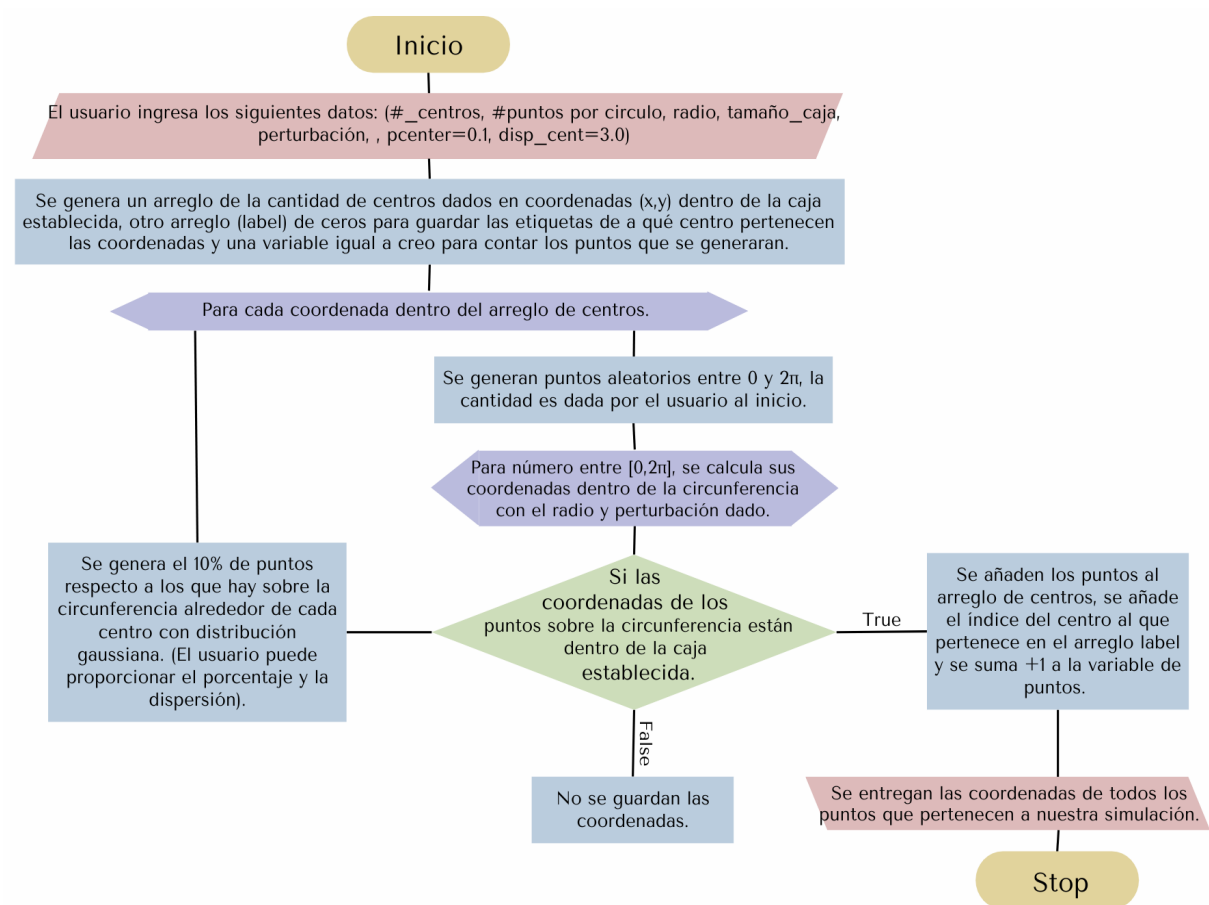


Figura 5.1: Diagrama de flujo del proceso para crear datos sintéticos de galaxias.

```

1 def DD(circulo):
2     # Save them all in one array
3     all_data = []
4     for i, c in enumerate(circulo):
5         for _, j in enumerate(c):
6             all_data.append(j)
7     return all_data
  
```

5.1.1. Datos aleatorios

El fin de este trabajo es obtener información de las distribuciones de galaxias a partir de los estimadores de la función de correlación de dos puntos, pero antes se analizarán dos distribuciones aleatorias para comprobar que su función de correlación efectivamente es cero cómo lo establece la teoría, es decir, se verá el caso en que las galaxias tuvieran una distribución aleatoria. Para este tipo de distribución utilizaremos la función:

```

1 rr = np.random.uniform(low=0.0, high=200.0, size=((1000,2)))
2 Nr = len(rr)
  
```

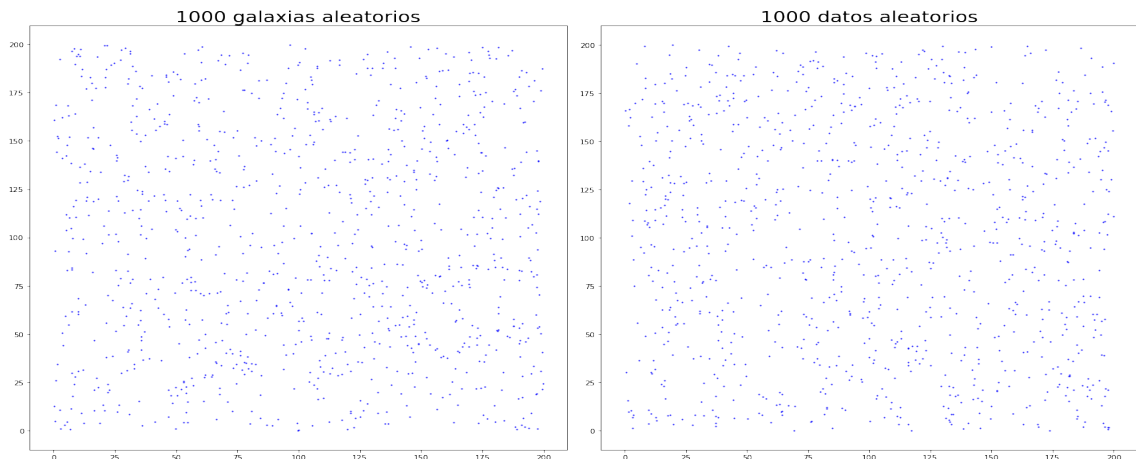


Figura 5.2: Dos distribuciones de 1000 datos aleatorios dentro de un Universo cuadrado de 200 unidades de longitud de lado.

Comenzaremos analizando las dos distribuciones aleatorias de la figura (5.2), y a continuación obtendremos las distancias Euclidianas entre todos los puntos de estas distribuciones y para esto programaremos la siguiente función:

$$\bar{r} = d(\bar{x}, \bar{y}) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (5.1)$$

```

1 def distancia(x1, x2):
2     diff_x = (x1[0] - x2[0])**2
3     diff_y = (x1[1] - x2[1])**2
4     return np.sqrt(diff_x+diff_y)

```

Posteriormente, con ayuda de la función **distancia** programamos otra para obtener la lista de distancias entre todos los puntos.

```

1 def comp_all_distances(sample):
2     dist_random = []
3     for i, _ in enumerate(sample):
4         g = partial(distancia, sample[i]) #partial() recibe una funcion A con
5             #sus respectivos argumentos y retorna una nueva funcion B que, al ser
6             #llamada, equivale a llamar a la funcion A con los argumentos provistos.
7         d = list(map(g, sample[i+1:]))
8         #La funcion map() toma una funcion y una lista y aplica esa funcion
9             #a cada elemento de esa lista, produciendo una nueva lista.
10        dist_random.extend(d)
11    return dist_random

```

Para calcular la cantidad total de distancias que se tendrán, utilizamos una combinatoria sin repetición, con esta se determina el número de subgrupos de x elementos que se pueden formar con los n elementos de una muestra, con la condición de que al menos un elemento sea diferente en cada subgrupo,

$$C_{n,x} = \binom{n}{x} = \frac{n!}{x!(n-x)!}. \quad (5.2)$$

Donde n es la cantidad de datos totales y x el número de elementos seleccionados. En

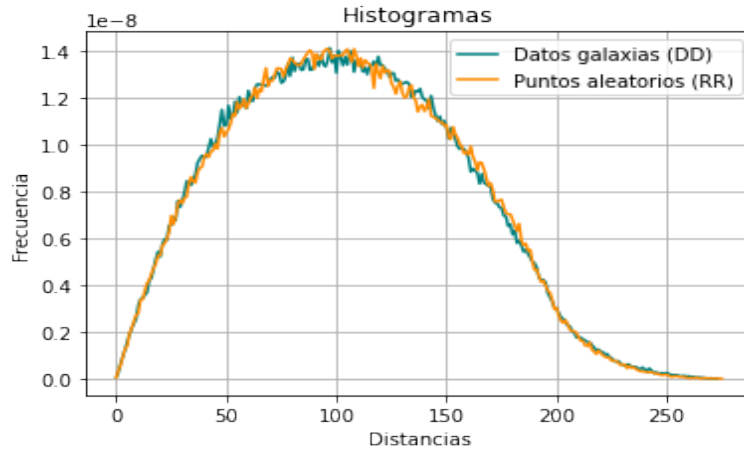


Figura 5.3: Histograma de 1000 galaxias con distribución aleatoria contra 1000 puntos con distribución aleatoria, mostrados en la figura (5.2).

nuestro caso tenemos 1,000 datos tanto en la distribución de galaxias aleatorias como en la distribución de datos aleatorios, por tanto $C_{1000,2} = 499,500$ y efectivamente se calcularon 499,500 distancias. Debido a que en este caso siempre tenemos $x = 2$ podemos simplificar la ecuación (5.2) de la siguiente forma,

$$C_{n,2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)(n-2)!}{2!(n-2)!} = \frac{n(n-1)}{2}. \quad (5.3)$$

A continuación se graficaron sus histogramas, para esto utilizamos la función **sorted()** para ordenar las distancias entre galaxias aleatorias y las distancias de los puntos aleatorios en orden específico ascendente, esto lo devuelve como una lista. Luego, con la función **round(r,0)** se redondearon las distancias a enteros. Después, se contó con la función **counter()** cuántas veces se repite cada distancia y se volvió a ordenar de manera ascendente. A continuación, en la figura (5.3) se muestran los histogramas de cada distribución, es decir, las gráficas de las distancias contra la frecuencia. En la misma figura se puede apreciar que a pesar de que las distribuciones de datos usaron la función **np.random.uniform** para sus coordenadas, cuya distribución es uniforme, las distribuciones de las distancias aleatorias se asemejan a una distribución de Poisson. Esto se puede deber a que es más probable que haya puntos con distancias de alrededor de la mitad del tamaño de la caja a diferencia de puntos que tengan una gran distancia entre ellos porque deberían estar ambos en las orillas, es decir se debe a sus fronteras. Para obtener una distribución uniforme se deberían establecer condiciones a la frontera tal que el espacio fuera un toro.

Para cada distancia se resta la frecuencia de las galaxias menos la frecuencia de los datos aleatorios y esta resta se divide entre la frecuencia de los datos aleatorios, lo cual es equivalente a calcularlo con el estimador **Peebles-Hauser** dado por la ecuación (3.11).

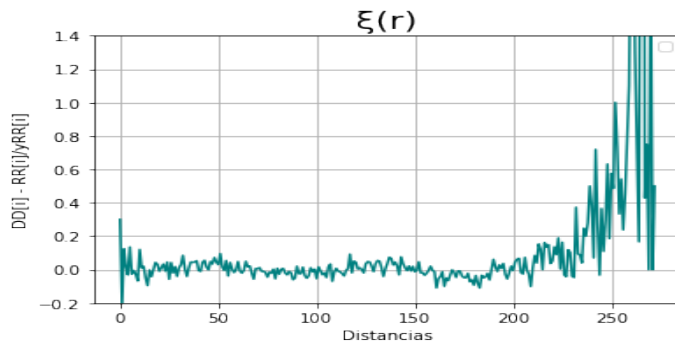


Figura 5.4: Función de correlación con estimador Peebles-Hauser de 1000 galaxias con distribución aleatorio contra 1000 puntos con distribución aleatoria, mostrados en la figura (5.2).

En la figura (5.4) se observa como la función oscila alrededor de cero, esto era esperado por la teoría. También podemos ver que en los bordes se crea ruido, debido a que a grandes distancias (superiores a las 200 unidades) hay poco conteo y por tanto más ruido. A continuación se graficarán las funciones de correlación obtenidas con diferentes estimadores presentados en el artículo [Vargas-Magaña et al., 2013].

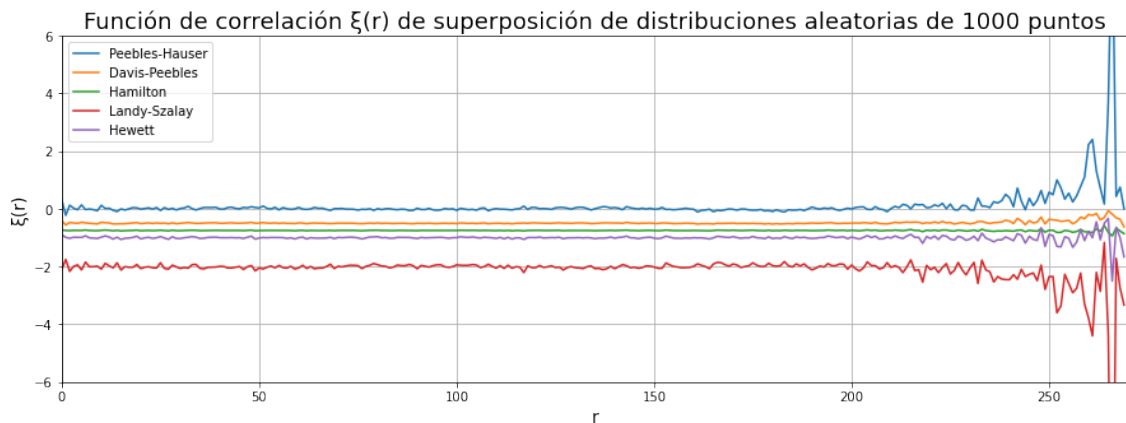


Figura 5.5: Funciones de correlación de una distribución aleatoria de 1000 galaxias y una distribución aleatoria de 1000 datos obtenida con 5 estimadores distintos.

Excluyendo las escalas mayores, ~ 200 unidades, en la figura (5.5) se puede apreciar que las funciones de correlación obtenidas oscilan alrededor de diferentes valores constantes, por lo que no hay ningún sobreagrupamiento de datos o una escala preferida. A continuación, para poder comparar los estimadores se normalizaran de tal manera que oscilen alrededor de 0. De la figura (5.6) podemos concluir que los estimadores **Peebles-Hauser** y **Landy-Szalay** son los que más ruido presentan a grandes distancias.

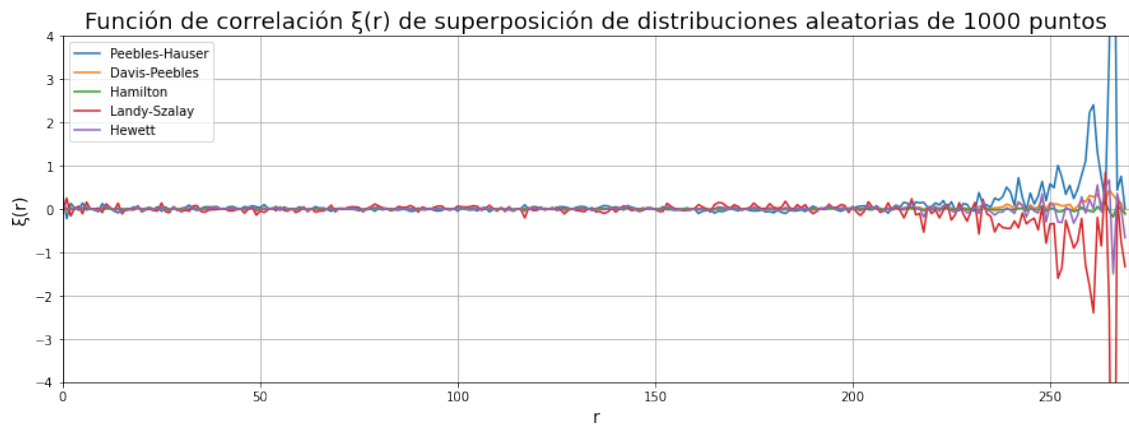
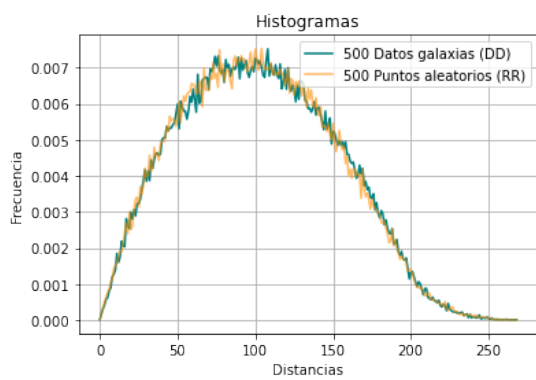
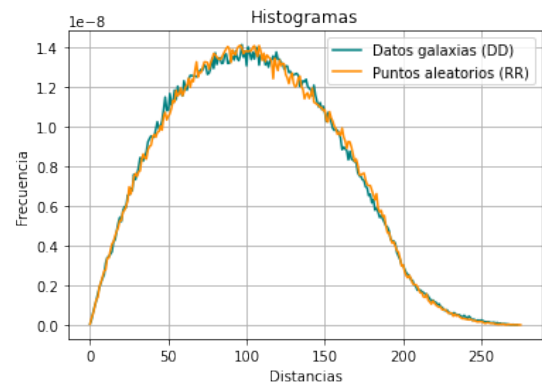


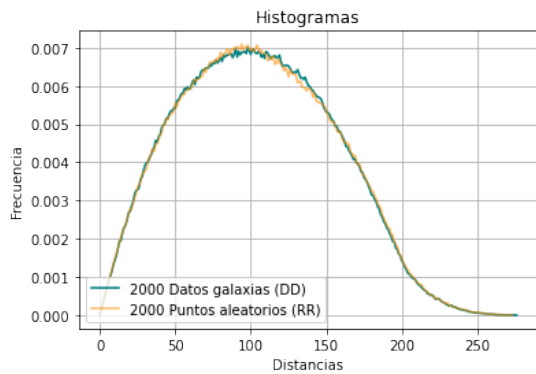
Figura 5.6: Funciones de correlación (trasladadas) de una distribución aleatoria de 1000 galaxias y una distribución aleatoria de 1000 datos obtenida con 5 estimadores distintos.



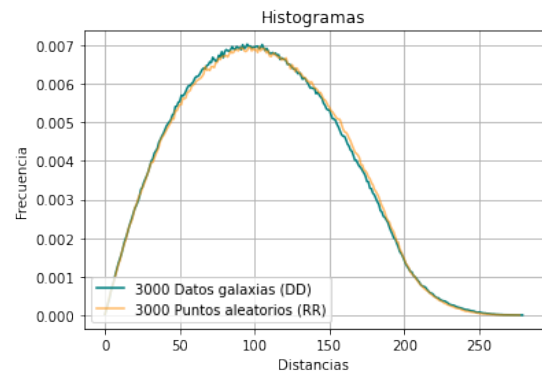
(a) 500 datos aleatorios y galaxias aleatorias.



(b) 1000 datos aleatorios y galaxias aleatorias.



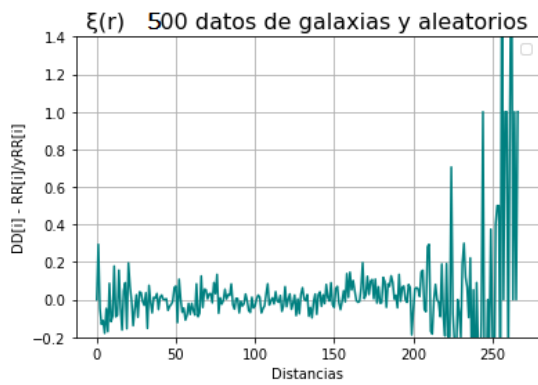
(c) 2000 datos aleatorios y galaxias aleatorias.



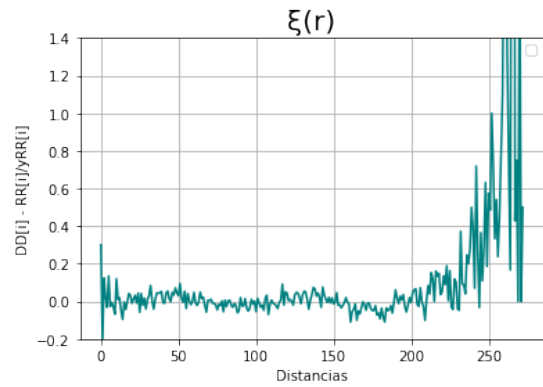
(d) 3000 datos aleatorios y galaxias aleatorias.

Figura 5.7: Histogramas de galaxias con distribución aleatoria contra puntos con distribución aleatoria.

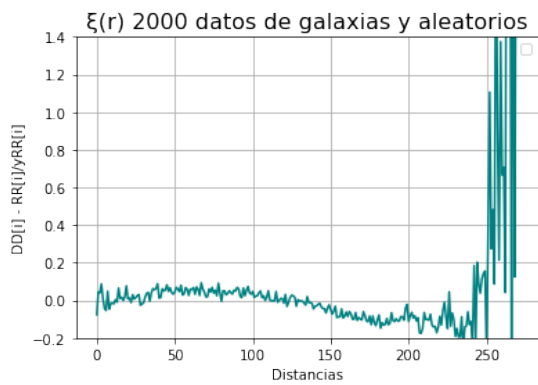
En la figura (5.7) podemos observar que los histogramas presentan menos ruido entre más datos se consideren (derecha-abajo).



(a) 500 datos aleatorios y galaxias aleatorias.



(b) 1000 datos aleatorios y galaxias aleatorias.



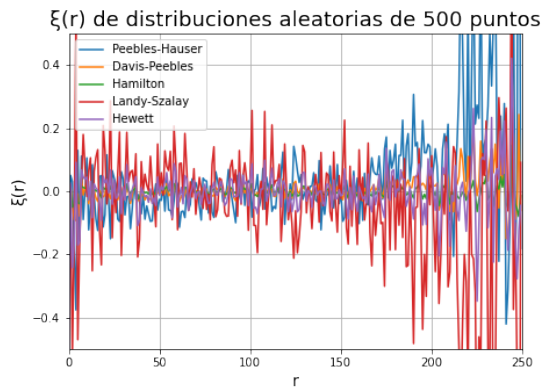
(c) 2000 datos aleatorios y galaxias aleatorias.



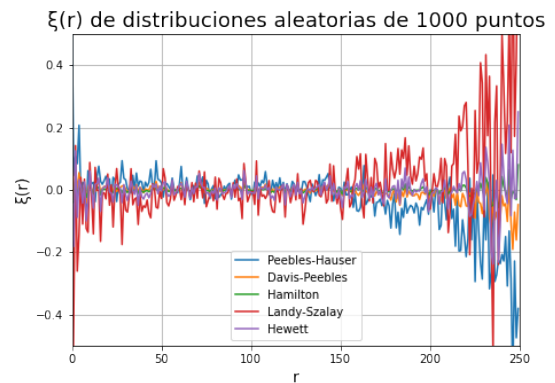
(d) 3000 datos aleatorios y galaxias aleatorias.

Figura 5.8: *Función de correlación con estimador Peebles-Hauser de distribuciones aleatoria contra puntos con distribución aleatoria.*

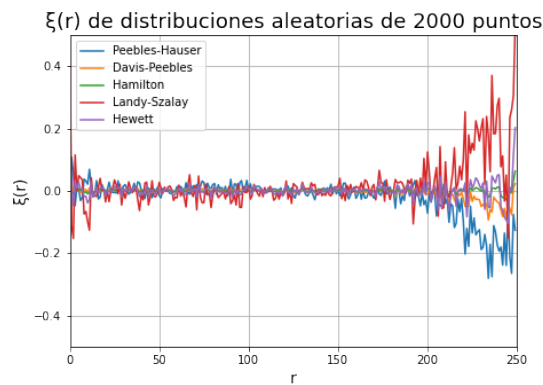
De la figura (5.8) se puede concluir que entre mayor cantidad de datos se consideren, menor será el ruido en la función de correlación, por lo tanto mayor será su precisión y resolución.



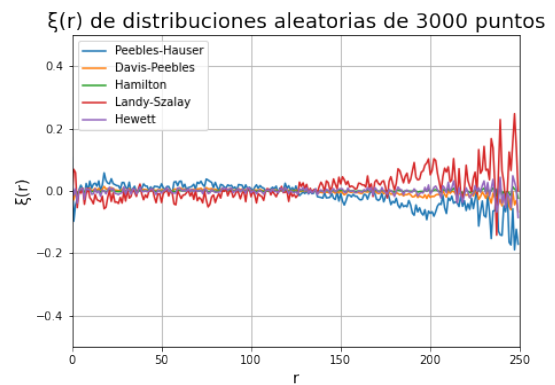
(a) 500 datos aleatorios y galaxias aleatorias.



(b) 1000 datos aleatorios y galaxias aleatorias.



(c) 2000 datos aleatorios y galaxias aleatorias.

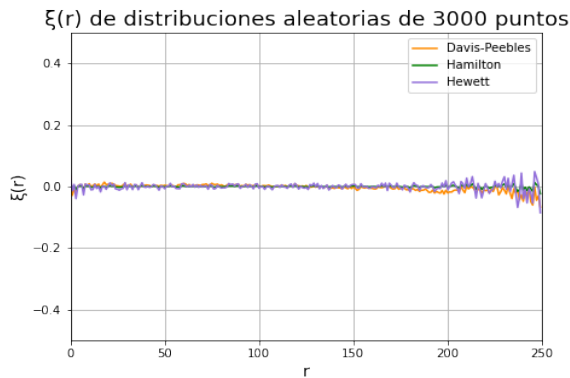


(d) 3000 datos aleatorios y galaxias aleatorias.

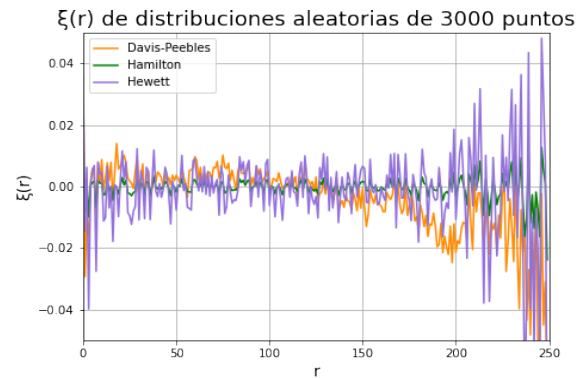
Figura 5.9: *Función de correlación con distintos estimadores de distribuciones aleatorias.*

En la figura (5.9) se compararon distintos estimadores y se puede apreciar que el ruido para pequeñas y grandes distancias disminuye conforme los datos aumentan, nuevamente se observa que entre mayor cantidad de datos se consideren, más preciso será el cálculo. El propósito de la comparación de diferentes estimadores es identificar aquel que proporciona la mayor cantidad de información, en otras palabras, determinar cuál es el más adecuado para el tipo de datos y el contexto en el que estamos trabajando. Sabemos que en el espacio que trabajamos, si comparamos dos distribuciones aleatorias, esperamos que la función de correlación de dos puntos con el estimador tienda a cero. En la figura (5.9(d)) se puede apreciar que los estimadores **Peebles-Hauser** y **Landy-Szalay** presentan bastante ruido para grandes distancias a comparación de los otros estimadores, es por eso que se repetirá la gráfica sin estos y se realizará un aumento en el eje y para determinar cual sería el mejor estimador es este caso.

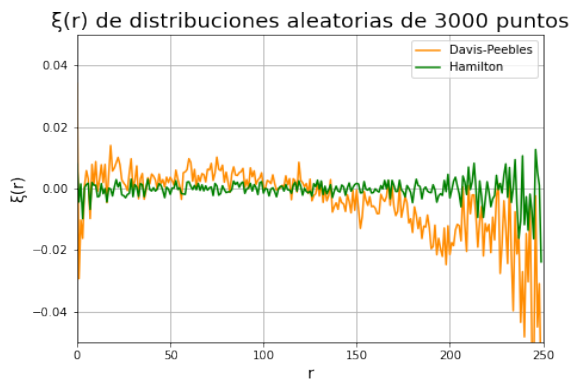
Si analizamos la figura (5.10(a)) podemos ver que es la figura (5.9(d)) pero en este caso sólo se analizan los estimadores **Davis-Peebles**, **Hamilton** y **Hewett** debido a que son los que menor ruido presentaron para grandes distancias. Luego, en la figura (5.10(b)) se realiza un aumento en el eje y para determinar el comportamiento de estos



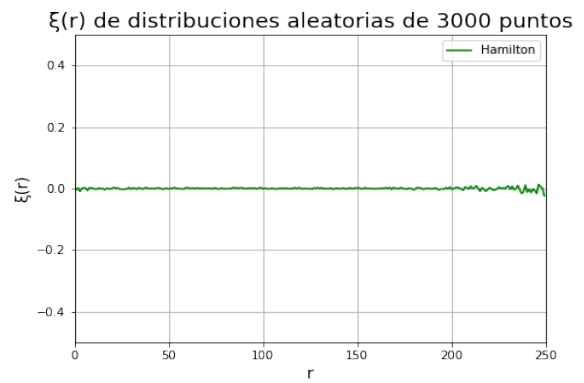
(a) 3000 datos aleatorios y galaxias aleatorias con un aumento en el eje y.



(b) 3000 datos aleatorios y galaxias aleatorias.



(c) 3000 datos aleatorios y galaxias aleatorias con aumento en eje y, analizando sólo los estimadores Davis-Peebles y Hamilton.



(d) Estimador Hamilton de la función de correlación de dos puntos para dos distribuciones aleatorias de 3000 datos cada una.

Figura 5.10: Función de correlación con distintos estimadores de dos distribuciones aleatorias con 3000 datos.

estimadores, ahora se aprecia que **Hewett** presenta mayor ruido para grandes distancias respecto a los otros dos estimadores. Se repite el análisis en (5.10(c)) y claramente se puede observar que el estimador **Davis-Peebles** presenta mayor ruido tanto para pequeñas como grandes distancias, incluso presenta mayor ruido en cualquier distancia en general. Finalmente, el estimador **Hamilton** de la función de correlación de dos distribuciones aleatorias es el que mejor funciona para nuestro espacio, como se ve en la figura (5.10(d)), debido a que es el que menor ruido presenta y oscila alrededor del cero.

5.1.2. Datos de galaxias

Anteriormente se realizó el análisis para una distribución de galaxias suponiendo que tienen una distribución uniforme, ahora se realizará el mismo análisis considerando que las galaxias se agrupan alrededor de circunferencias y de los centros de las circunferencias debido a los BAO. Iniciaremos comparando como se ve en la figura

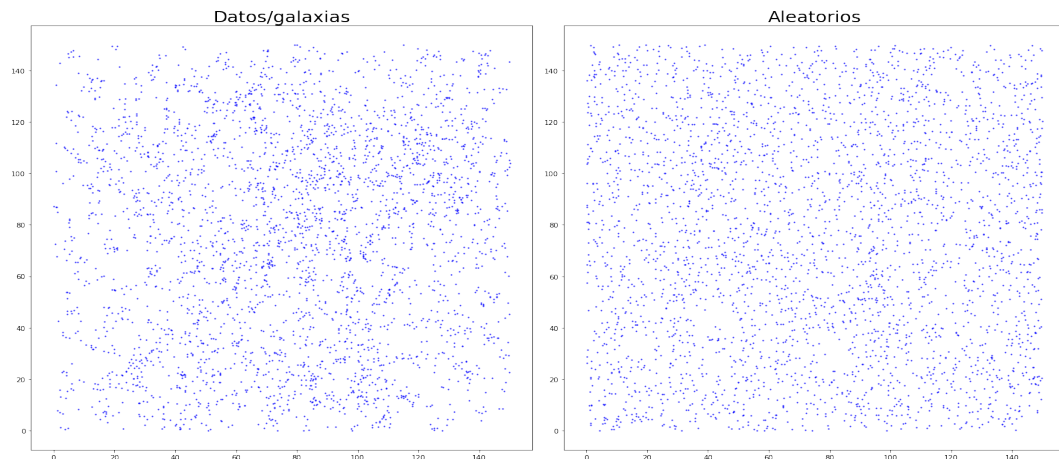


Figura 5.11: Lado izquierdo: distribución esquemática de 3290 galaxias distribuidas en 150 anillos de radio 15 con máximo 20 galaxias por anillo, con una perturbación del 10 % del radio y 4 galaxias en cada centro. Lado derecho: 3290 puntos con distribución uniforme. Ambos en un Universo cuadrado de 150 unidades de longitud de lado.

(5.11) la distribución esquemática de 3290 galaxias y 3290 puntos aleatorios. A simple vista no se observan los anillos de la distribución de galaxias incluso se podría pensar que ambas son distribuciones aleatorias por lo que es útil usar la función de correlación para detectar si hay alguna escala preferida y distinguir distribuciones de galaxias de distribuciones aleatorias.

El Universo con el que se está trabajando tiene condiciones de frontera no periódicas, en caso de ser periódicas sería equivalente a cubrir un espacio mayor. Se midió el espacio promedio disponible utilizando la integración de Monte Carlo, generando una distribución aleatoria de comparación sobre la misma área. En el artículo [Alonso, 2013] se menciona la importancia de generar catálogos aleatorios correctamente, es por eso que deben tener más partículas que los datos de galaxias para disminuir los errores de Poisson y se deben considerar todos los efectos observacionales que afecten en la distribución de galaxias como el corrimiento al rojo y los ángulos. En el artículo [He, 2021a] se menciona que para el estudio de distribuciones de galaxias con condiciones de límites no periódicos, se suele medir los promedios del Universo de estudio utilizando integración de Monte Carlo, donde se genera una distribución aleatoria de comparación que contiene un gran número de puntos sobre la misma área de estudio y que con el fin de reducir las fluctuaciones estadísticas, es estándar usar campos aleatorios densamente poblados, generalmente 50 veces más densos que los datos, sin embargo se recomienda utilizar la mayor cantidad de datos aleatorios posible para reducir el ruido del estimador, el límite está en la capacidad CPU que tenemos. En nuestro caso se trabaja con el 150 % de puntos respecto a los datos de galaxias pero con una distribución aleatoria.

Primero, se analizará el tamaño ideal de la porción del Universo/caja respecto al radio de los anillos para detectar la escala preferida. Por este motivo se compararan

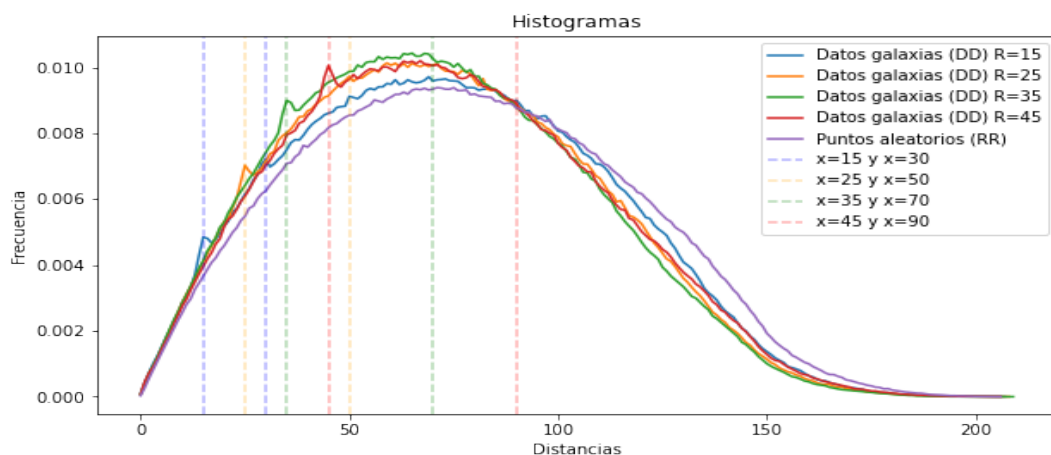


Figura 5.12: Histogramas de distintas distribuciones de galaxias, todas son superposiciones de 150 anillos, en un Universo cuadrado y plano de 150 de lado, con perturbación del 10% respecto al radio, con hasta entre 20 y 25 galaxias por anillo y distintos radios ($R = 15, 25, 35, 45$).

los histogramas y se aplicará el estimador Peebles-Hauser a distintas distribuciones de galaxias con cantidades similares de galaxias pero distintos radios de anillos. En la figura (5.12) se presentan los histogramas de estas distribuciones y se aprecia que todas las distribuciones de galaxias están recargadas a las menores distancias a diferencia de la distribución aleatoria. Incluso se tiene la misma cantidad de galaxias que puntos aleatorios y a diferencia de la figura (5.7) donde ambas distribuciones aleatorias se “traslapan”, en estos casos sus picos son distintos y en general los histogramas son distintos. En la distribución de galaxias se aprecian diferentes picos, en particular dos, el radio y el diámetro. El pico que mejor se ve es el de los datos de galaxias con radio BAO de 15 y efectivamente su pico está en 15 unidades de longitud, es decir, que si te encuentras en una galaxia es más probable encontrar otra a 15 unidades de longitud. Basándonos exclusivamente en cómo se ven los picos de los histogramas, debido a que los últimos dos no se detectan tan bien, para los siguientes análisis se usará un radio menor de 25 para cajas de 150 unidades por lado.

Luego en la figura (5.13) comparamos como cambia el pico de BAO en la función de correlación de dos puntos en función del radio y la cantidad de galaxias sobre un anillo de BAO. En el primer cuadrante se ven claramente los dos picos de BAO, mientras que para el segundo, tercero y cuarto cuadrante sólo se ve claro el primer pico BAO. Debido a esto, por fines prácticos, para los siguientes análisis se fijará un radio de 20 unidades y un Universo de 150 unidades por lado.

5.1.3. 2pcf en función de la cantidad de galaxias y anillos

En la figura (5.14), con el fin de analizar como afecta al análisis tener pocas y muchas galaxias en pocos y muchos anillos, se muestran cuatro distribuciones esquemáticas formadas al colocar las galaxias en anillos del mismo radio característico 20 con perturbación del 1% del radio. En el cuadrante superior izquierdo, vemos

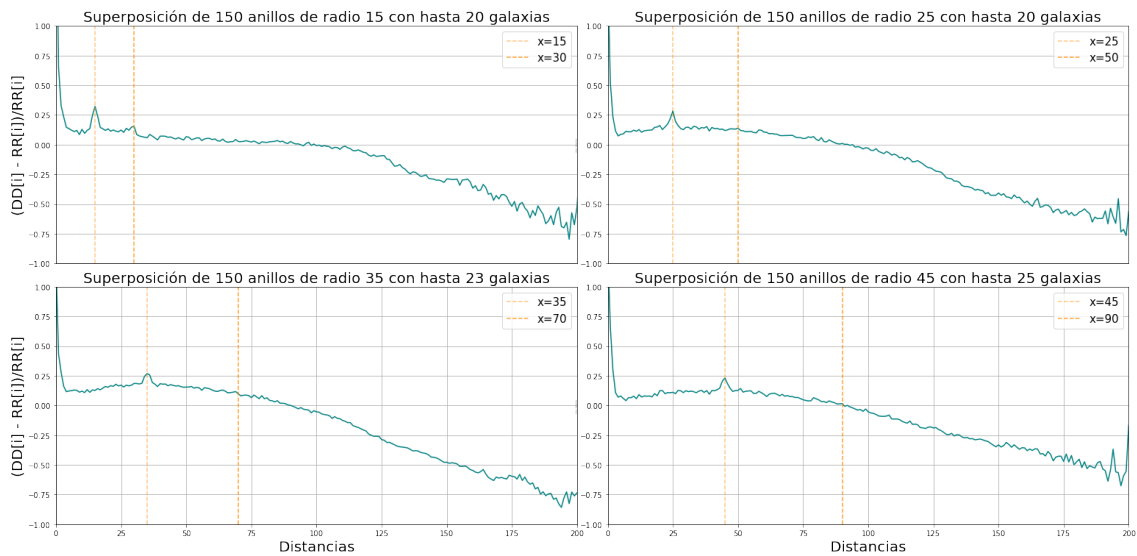


Figura 5.13: Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para distintas distribuciones de galaxias, todas son superposiciones de 150 anillos, en un Universo cuadrado y plano de 150 de lado, con perturbación del 10 % respecto al radio, con hasta entre 20 y 25 galaxias por anillo y distintos radios ($R = 15, 25, 35, 45$).

975 galaxias, distribuidas en 50 anillos, con máximo 25 galaxias en cada uno. En el cuadrante superior derecho, vemos 1982 galaxias, distribuidas en 50 anillos, con máximo 50 galaxias en cada uno. En el cuadrante inferior izquierdo, vemos 3133 galaxias, distribuidas en 150 anillos, con máximo 25 galaxias en cada uno. Por último, en el cuadrante inferior derecho vemos 6422 galaxias, distribuidas en 150 anillos, con máximo 50 galaxias en cada uno. Se puede apreciar que la escala radial característica es claramente visible en el cuadrante superior derecho. Mientras que en el primer cuadrante superior izquierdo a pesar de contener la misma cantidad de anillos en los que se distribuyen los datos, al disminuir la cantidad de datos por anillos no se logra apreciar la escala radial característica. Por otra parte, en el cuadrante inferior izquierdo se muestra un escenario más realista, con muchos anillos y relativamente pocas galaxias por anillo. Por lo tanto, vemos que conforme aumenta el número de anillos y/o se reduce la cantidad de galaxias sobre cada anillo, se “oculta” visualmente la escala preferida, lo que implica que solo puede recuperarse estadísticamente [Bassett and Hlozek, 2009]. A continuación, procedemos a graficar los histogramas de estas 4 distintas distribuciones de galaxias como se aprecia en la figura (5.15). El pico de la escala característica es muy notorio para pocos anillos con muchas y pocas galaxias, sin embargo al aumentar la cantidad de anillos vemos que el pico ya no es tan pronunciado.

Luego en la figura (5.16) podemos observar las funciones de correlación de dos puntos respectivas a cada cuadrante de la figura (5.14). En ellas podemos ver las regiones de sobre-densidad y notamos que el pico de BAO es más notorio entre mayor cantidad de galaxias por anillo BAO se consideren, aunado a que entre mayor cantidad de anillos BAO se consideren, menos notorio será el pico BAO.

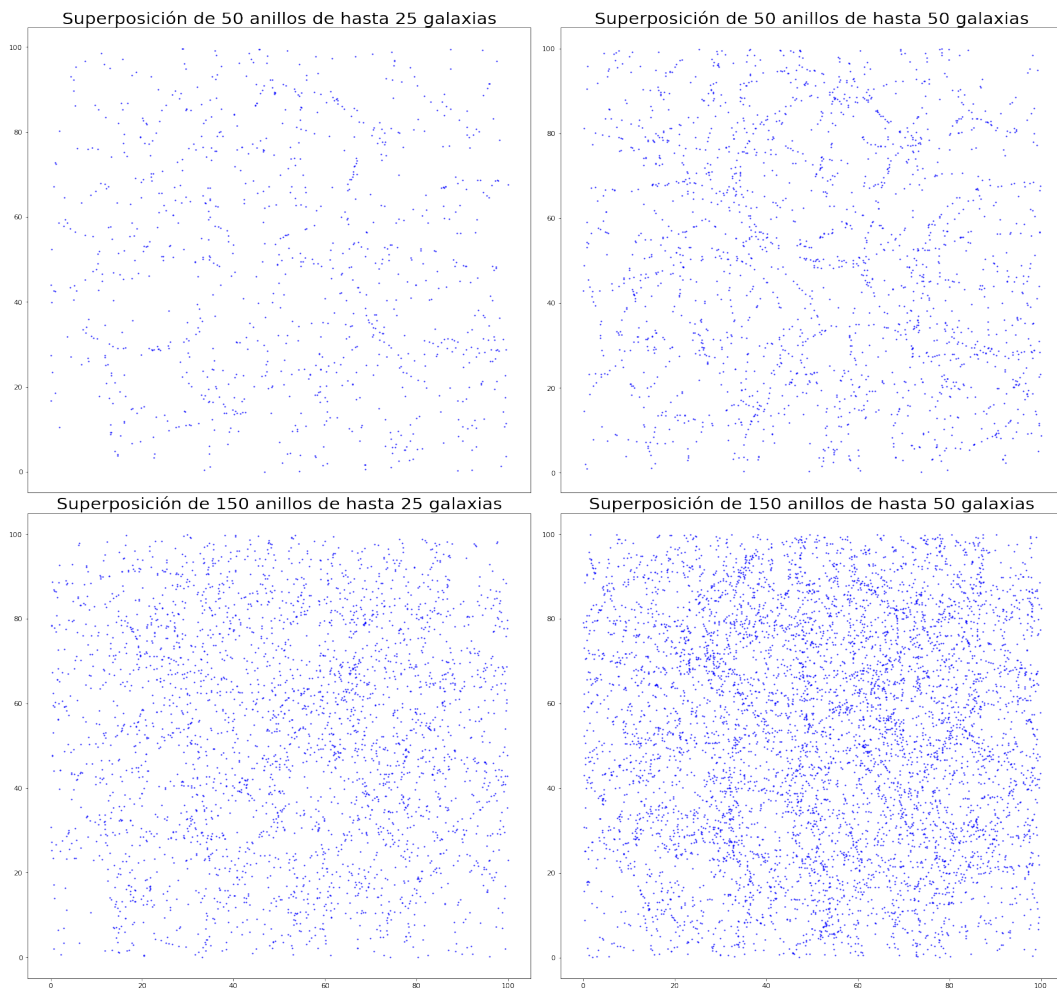


Figura 5.14: El primer cuadrante tiene 975 galaxias, el segundo 1982, el tercero 3133 galaxias y el cuarto cuadrante 6422. Esta distribución esquemática de galaxias muestra que entre más galaxias haya en menos anillos la escala radial característica se apreciara mejor. Entre menos galaxias haya por anillo o más anillos haya será más complicado detectarlo visualmente.

5.1.4. 2pcf en función de la perturbación

Recordemos que las galaxias se agrupan en promedio alrededor de circunferencias debido a los BAO y es en promedio debido a que presentan perturbaciones por velocidades peculiares, atracciones gravitacionales, choque de galaxias, entre otros motivos. A continuación analizaremos como el porcentaje de perturbación modifica la estructura de la distribución de galaxias, así como al pico BAO en la función de correlación de dos puntos. La cantidad de perturbación que presenta una distribución de galaxias respecto al BAO original es muy relevante ya que, como se puede observar en la figura (5.17) ésta modifica que tan aleatoria o no parece ser una distribución. En el primer cuadrante con perturbación del 0% se logran apreciar las circunferencias, sin embargo en el último cuadrante con perturbación del 10% es mucho más complicado apreciarlas.

A continuación en la figura (5.18) podemos apreciar como se modifica el pico del

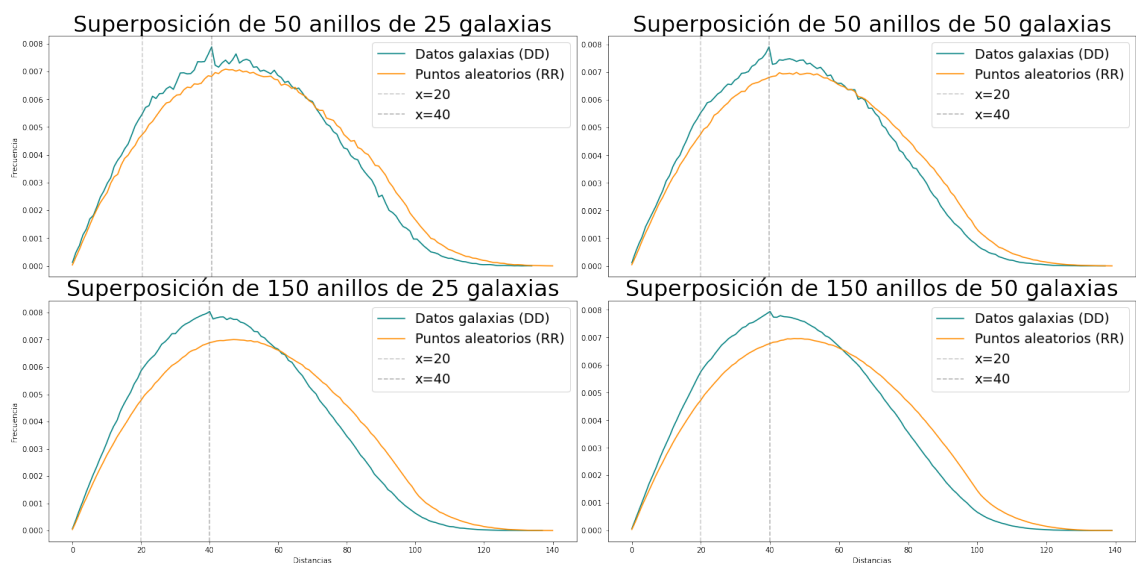


Figura 5.15: Cada histograma de esta figura corresponde al histograma de la distribución del respectivo cuadrante de la figura (5.14).

BAO en función de la perturbación que tienen las galaxias de las distribuciones de la figura (5.17). Podemos observar que entre mayor es la perturbación más disminuye su amplitud, es decir, menos predominante es el pico (*se achata*).

5.1.5. 2pcf con distinto porcentaje de galaxias en el centro respecto a las galaxias que hay sobre el anillo de BAO

En la presente sección analizaremos como se modifica el pico de la función de correlación de dos puntos en función del porcentaje de galaxias que hay en el centro del BAO respecto a las galaxias que se encuentran en el anillo. En la figura (5.19) no se observa diferencia relevante entre los histogramas de distribuciones de galaxias con diferentes porcentajes de galaxias en el centro respecto a las galaxias en los anillos.

Por otro lado, en la figura (5.20), en la gráfica izquierda podemos apreciar la función de correlación de dos puntos para diferentes porcentajes de galaxias en el centro, claramente se aprecian tres picos, es decir, tres zonas de sobredensidad. La primer zona es cercana al cero y corresponde a los puntos dentro del centro de los anillos (los cuales tienen poca distancia entre ellos). La segunda zona es alrededor de 20 y corresponde al radio de los BAO simulados. Por último, el tercer pico está alrededor de 40 y corresponde al diámetro de los BAO simulados. Luego, la gráfica de la derecha es un acercamiento en el eje y de la izquierda, lo cual es útil para apreciar como cambian los picos BAO dependiendo del porcentaje de galaxias en los centros respecto a los anillos. Podemos ver que entre menor es el porcentaje de galaxias en el centro, menos ancho y alto es el pico del radio mientras que menos ancho y más alto es el pico del diámetro. Esto se aprecia debido a que para el radio el pico más ancho y alto es para 30% de galaxias en el centro y el pico menos ancho y más bajo es para

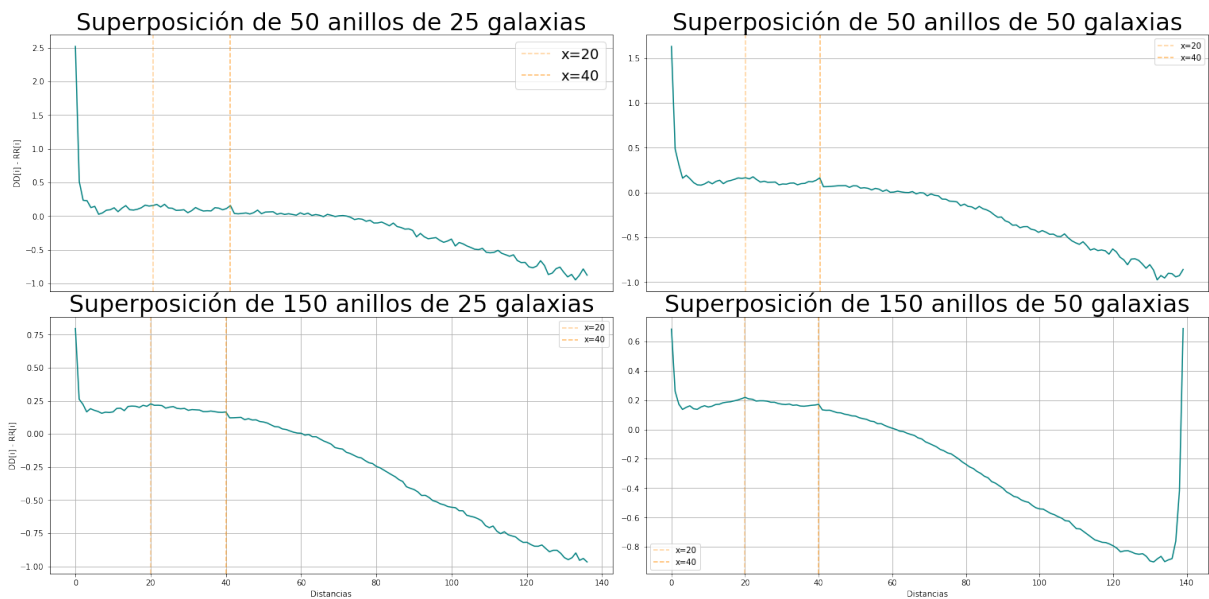


Figura 5.16: Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para distintas distribuciones de galaxias respectivas a cada cuadrante de la figura (5.14).

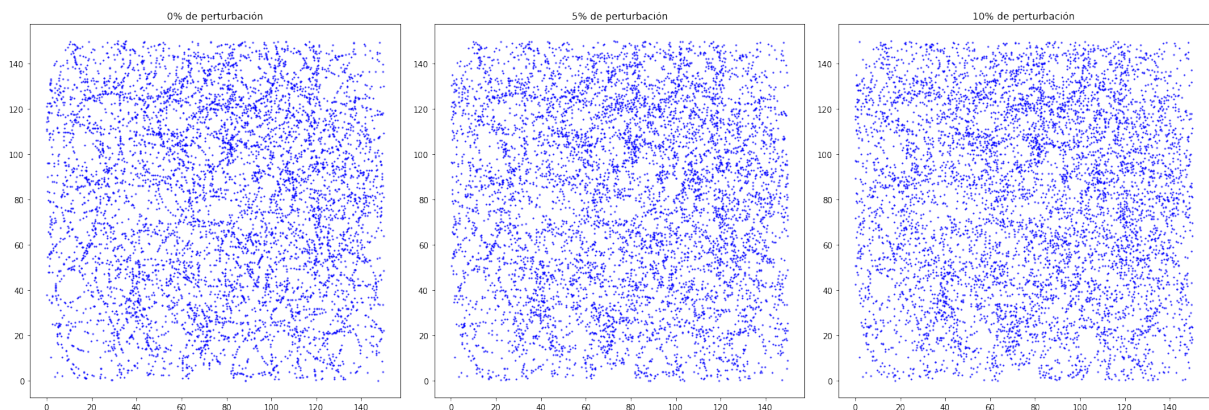


Figura 5.17: Distribuciones de galaxias con perturbaciones del 0 %, 5 % y 10 % respectivamente.

10 % de galaxias en el centro. Por otro lado, para el pico del diámetro el más alto es para 10 % de galaxias en el centro y el más bajo es para 30 % de galaxias en el centro.

5.1.6. 2pcf con distinta cantidad de galaxias sobre el anillo de BAO

A continuación analizaremos como la cantidad de galaxias por anillo modifica el pico BAO de la función de correlación de dos puntos. Para esto utilizamos distribuciones de galaxias con 25, 35, 45 y 55 galaxias por anillo BAO, como se muestra en la figura (5.21). En esta misma figura se muestra el histograma de diferentes distribuciones pseudoaleatorias uniformes con diferentes cantidades de datos (la menor con 5343 datos y la mayor con 11860 datos).

En la primer gráfica de la figura (5.22) se aprecian nuevamente los tres picos de

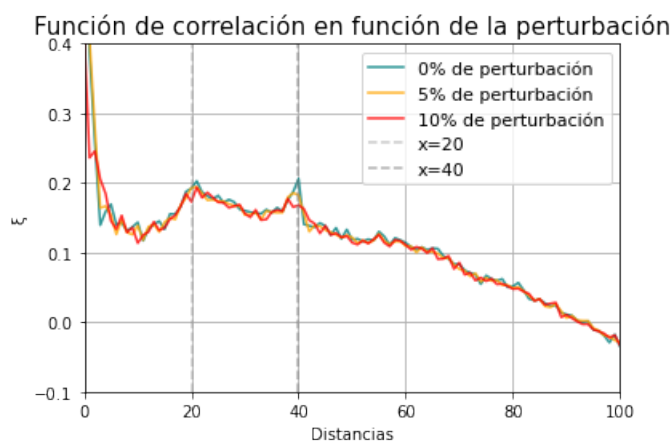


Figura 5.18: Función de correlación de dos puntos obtenidas con el estimador Peebles-Hauser para las distribuciones de la figura (5.17) que presentan distintas perturbaciones (0%, 5% y 10%).

sobre densidades de galaxias, los cuales corresponden a las galaxias del centro, el del radio de BAO y el del diámetro de BAO. Con ayuda del aumento de la gráfica (derecha), vemos que entre mayor cantidad de galaxias por anillo de BAO se consideren más aumentaran los picos del radio y diámetro y más definido será el pico del diámetro.

5.2. DISTINTAS FORMAS DE OBTENER LA DISTRIBUCIÓN ALEATORIA

Por otra parte en el artículo [He, 2021a] proponen que el conteo de pares aleatorio-aleatorio $RR(r)$ es una cantidad puramente geométrica, depende sólo de la geometría del catálogo de galaxias y la función de selección radial en el caso del estudio cosmológico, de manera que para un espacio de datos cuadrado de lado a , se propone la siguiente función:

$$RR_{\text{cuadrado}}(r) = \frac{2\pi}{a^2}r - \frac{8}{a^3}r^2 + \frac{2}{a^4}r^3. \quad (5.4)$$

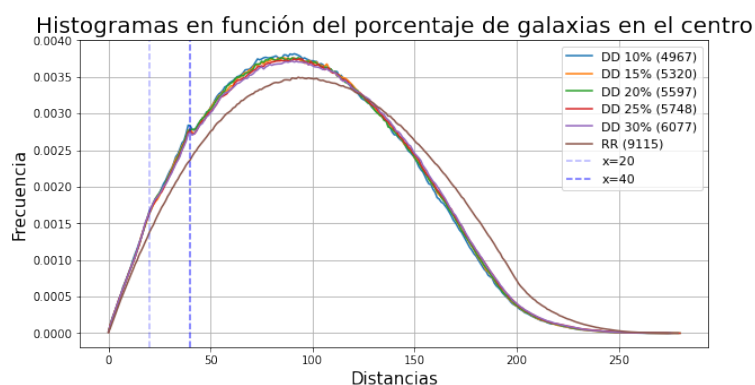


Figura 5.19: Histograma de distribuciones de galaxias con distintos porcentajes (10%, 15%, 20%, 25% y 30%) de galaxias en el centro respecto a los anillos. Entre paréntesis se muestra la cantidad total de galaxias con la que se trabajó cada distribución.

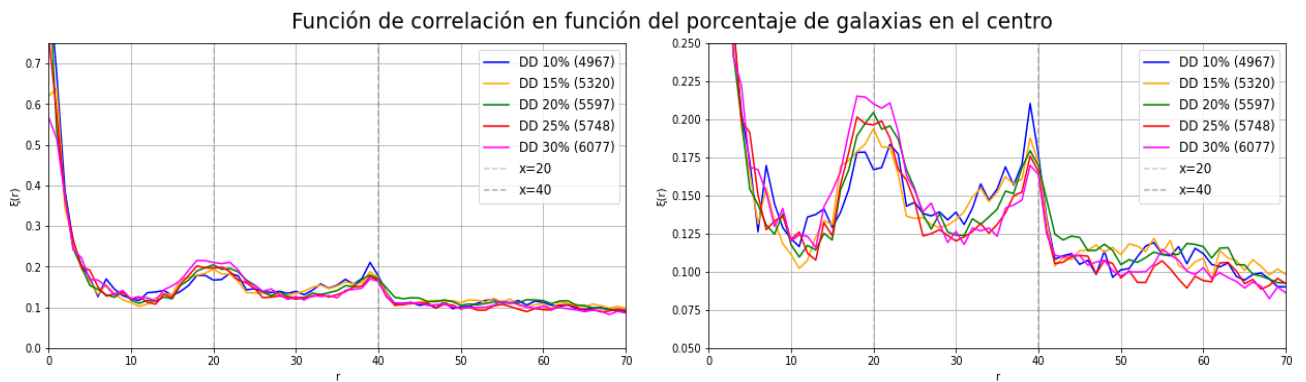


Figura 5.20: (Izquierda) Función de correlación de dos puntos obtenida con el estimador Peebles-Hauser para distribuciones de galaxias con distintos porcentajes (10 %, 15 %, 20 %, 25 % y 30 %) de galaxias en el centro respecto a los anillos. (Derecha) Aumento en el eje y.

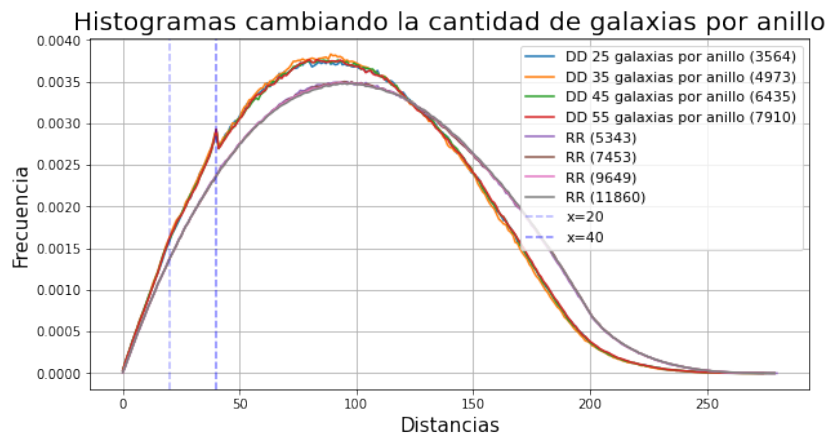


Figura 5.21: Histograma de distribuciones de galaxias con 5 % de perturbación y distinta cantidad de galaxias por anillo (25, 35, 45, 55). Entre paréntesis se muestra la cantidad total de galaxias con la que se trabaja cada distribución.

En éste mismo artículo, observaron que al analizar los datos con esta función concordaban muy bien con los resultados que con las estimaciones de integración de Monte Carlo, es equivalente a utilizar una infinidad de datos aleatorios, así que da una estimación perfecta. A continuación analizamos nuevamente la función de correlación de dos puntos con el estimador Peebles-Hauser utilizando esta nueva $RR(r)$ y se comparará el tiempo que toma contra el que usa un código Monte Carlo de fuerza bruta.

De la figura (5.23) podemos notar que ambas funciones tienen comportamientos distintos para diferentes escalas. La función geométrica presenta una infradensidad de galaxias para pequeñas distancias a diferencia de con el método Monte Carlo que indica una sobredensidad (la cual habíamos considerado que era debido a las galaxias en los centros). Con ambos métodos podemos observar los picos debido al radio y diámetro del BAO. Lo interesante es que con el método utilizado durante este trabajo

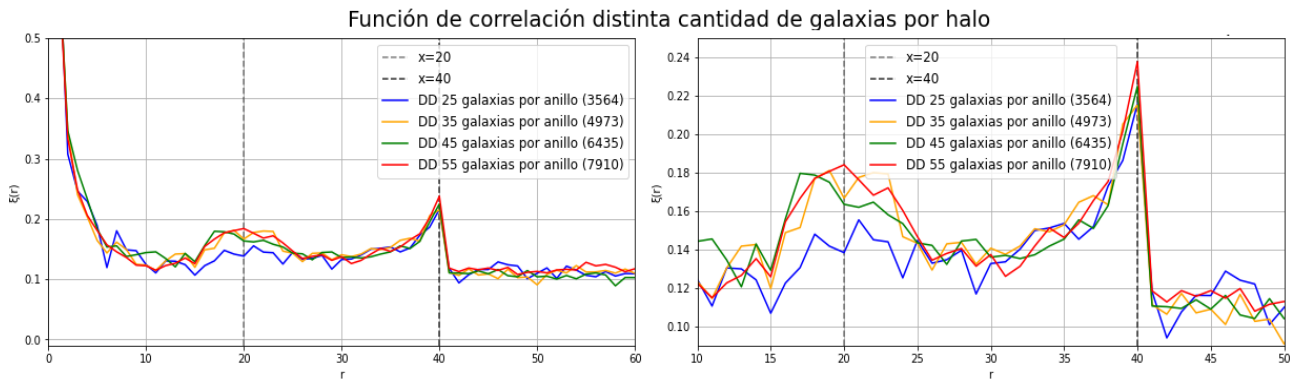


Figura 5.22: (Izquierda) Función de correlación de dos puntos obtenida con el estimador Peebles-Hauser para distribuciones de galaxias con 1% de perturbación, 10% de galaxias en el centro respecto a las galaxias sobre el anillo y distinta cantidad de galaxias por anillo (25, 35, 45, 55). (Derecha) Aumento tanto en eje x como eje y .

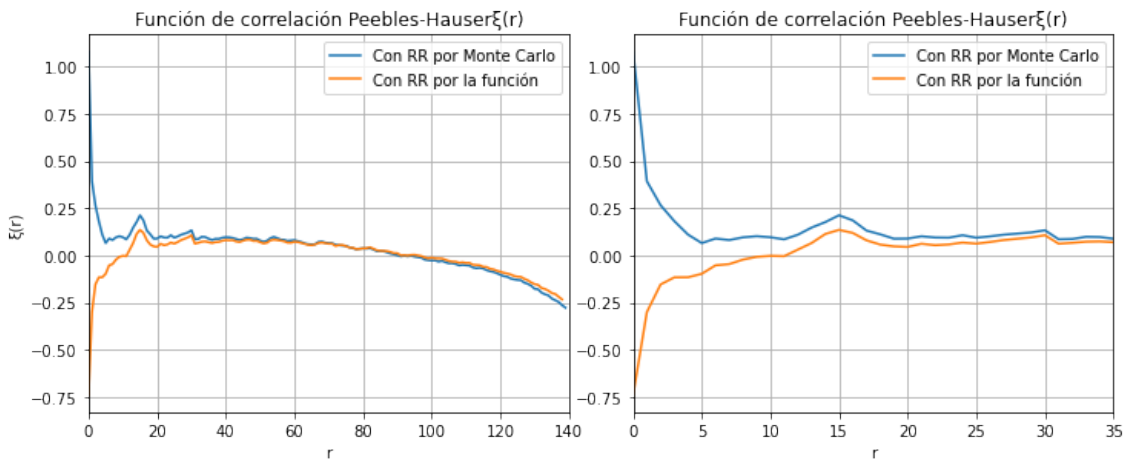


Figura 5.23: Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica. En este caso la distribución de galaxias consistió de 200 circunferencias de radio 15 con máximo 30 galaxias cada una, 30% de galaxias en los centros respecto a las circunferencias, en un Universo cuadrado de 150 unidades de longitud de lado.

de generar datos aleatorios en el espacio en el cual se encuentra la distribución de galaxias y posteriormente medir la distancia entre todos los datos aleatorios (en este caso para 7013 datos aleatorios) a la computadora Lenovo S145 con intel core i7 le tomó 160.6171s. Por otro lado analizarlo con la función geométrica le tomó 868 μ s, es decir, la función geométrica es mucho más rápida, es aproximadamente del orden de 10^5 veces más eficiente que el primer método en términos de tiempo de ejecución. A continuación lo repetiremos para distintos radios para analizar como cambian las funciones de correlación en función del radio con los dos métodos.

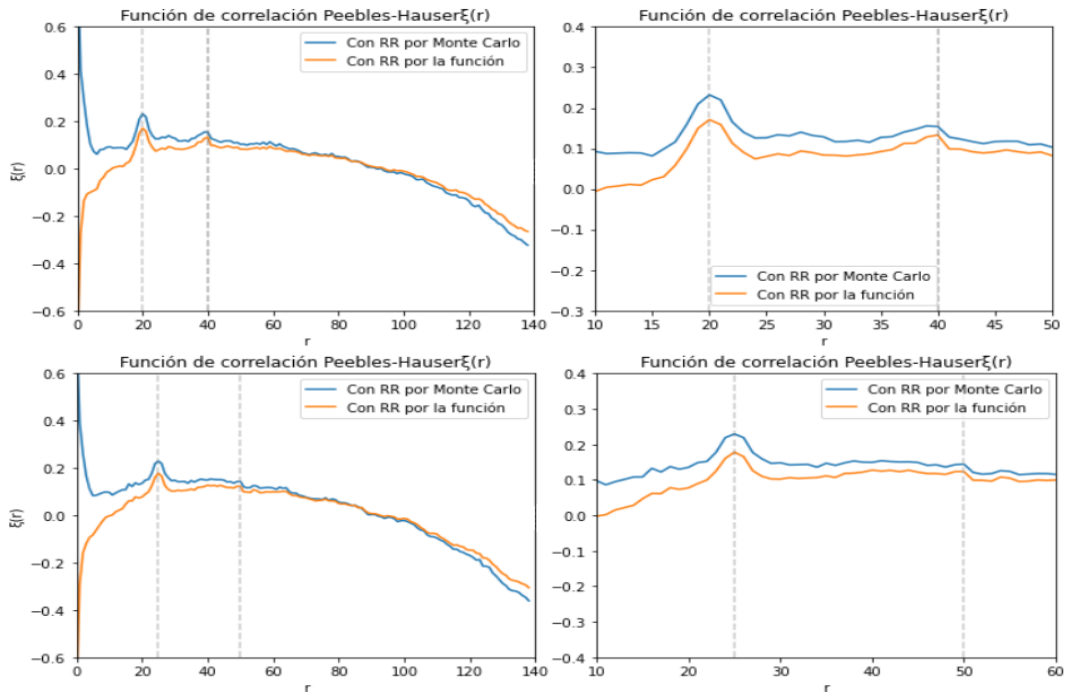


Figura 5.24: Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica. En este caso la distribución de galaxias consistió de 200 circunferencias de radio (Arriba) 20, (Abajo) 25 con máximo 30 galaxias cada una, 30 % de galaxias en los centros respecto a las circunferencias, en un Universo cuadrado de 150 unidades de longitud de lado.

De las figuras (5.23) y (5.24) podemos observar como se modifican los picos para diferentes radios en los dos métodos, en este caso vemos que para el tamaño de nuestro Universo, se aprecian mejor los picos de diámetro y radio para circunferencias de galaxias de radio 20 unidades de longitud. Es por esto que seguiremos utilizando un radio de 20 unidades de longitud pero ahora modificaremos el porcentaje de galaxias en el centro.

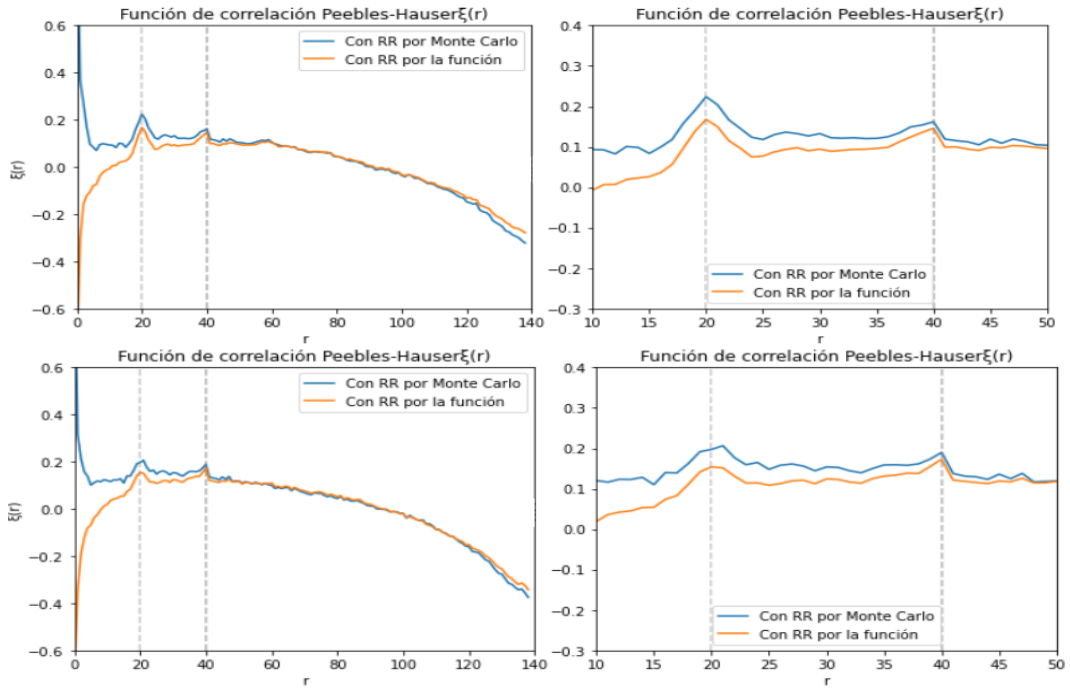


Figura 5.25: Comparación de la función de correlación de dos puntos obtenida con el estimador Peebles-Hauser utilizando Monte Carlo con una distribución aleatoria contra una función geométrica. En este caso la distribución de galaxias consistió de 200 circunferencias de radio 20 con máximo 30 galaxias cada una, en un Universo cuadrado de 150 unidades de distancia de lado, (Arriba) 20 % de galaxias en los centros respecto a las circunferencias, (Abajo) 10 % de galaxias en los centros respecto a las circunferencias.

De las gráficas de la figura (5.25) podemos apreciar que con ambos métodos al disminuir el porcentaje de galaxias en el centro, el pico del radio se achata (gráfica inferior derecha). A continuación, presentamos la tabla (5.1) en la que se compara el tiempo de ejecución para generar los datos aleatorios con ambos métodos para las diferentes figuras.

Figura	Monte Carlo	Función geométrica
5.23	$6.061710 \times 10^1 s$	$8.68 \times 10^{-4} s$
5.24 (Arriba)	$1.609125 \times 10^2 s$	$1.01 \times 10^{-3} s$
5.24 (Abajo)	$1.247000 \times 10^2 s$	$4.0 \times 10^{-3} s$
5.25 (Arriba)	$1.112000 \times 10^2 s$	$1.07 \times 10^{-3} s$
5.25 (Abajo)	$8.630692 \times 10^1 s$	$1.06 \times 10^{-3} s$

Tabla 5.1: Tiempos de ejecución.

Finalmente, vemos que al aumentar la cantidad de datos aleatorios efectivamente converge a a la función geométrica que es el caso ideal, por lo que si tenemos un espacio controlado y podemos utilizar alguna de las funciones geométricas presentadas en el artículo [He, 2021a] debemos utilizarlas ya que por definición reducen el trabajo

y disminuyen el ruido. Para distancias cercanas a cero vemos que hay un enorme error estadístico para ambos métodos de datos aleatorios.

5.3. ESTIMADORES DE LA FUNCIÓN DE CORRELACIÓN

A continuación analizaremos las diferencias de los estimadores **Peebles-Hauser**, **Davis-Peebles**, **Hamilton**, **Landy-Szalay** y **Hewett** para este tipo de espacio y distribución. Debemos considerar que para grandes distancias hay pocos conteos y por tanto se presentará mayor ruido. A continuación analizaremos como cambian los estimadores de la función de correlación de dos puntos en función de la cantidad de galaxias en anillos y centros, cantidad de anillos y porcentaje de la perturbación.

5.3.1. 2pcf en función de la cantidad de galaxias y anillos

Comencemos analizando los distintos estimadores de la función de correlación de dos puntos en función de la cantidad de galaxias y anillos. En las gráficas de la figura (5.26) se muestran distribuciones sintéticas de 50 anillos BAO con 20 galaxias cada uno, 50 anillos BAO con 40 galaxias cada uno y 150 anillos BAO con 40 galaxias cada uno respectivamente. Las tres figuras contienen el panorama de distancias completo y un zoom cerca de su pico BAO para un mejor análisis. Podemos observar que el ruido para grandes distancias disminuyó al aumentar los 50 anillos de BAO de 20 galaxias a 40 galaxias. Sin embargo el pico BAO de las gráficas de arriba y en medio, no cambió notoriamente, en ambas figuras los estimadores muestran el pico de BAO de manera mayor a menor de la siguiente forma: **Peebles-Hauser**, **Hewett**, **Landy-Szalay**, **Davis-Peebles** y finalmente el menor pico lo presentó el estimador **Hamilton**. Por otro lado, al aumentar la cantidad de anillos de BAO de 50 a 150 manteniendo 40 galaxias por anillo, en la gráfica de abajo el pico disminuyó. Por lo que podríamos deducir que entre mayor cantidad de anillos de BAO haya, más ténue será el pico de BAO.

5.3.2. 2pcf en función de la perturbación

Continuaremos el análisis con la figura (5.27) en la cual se muestran 150 anillos con máximo 40 galaxias por anillo pero ahora con 0 %, 5 % y 10 % de perturbación. En las gráficas del lado derecho, vemos que conforme la perturbación aumenta el pico de BAO se achata. Nuevamente, los estimadores muestran el pico de BAO de manera mayor a menor de la siguiente forma: **Peebles-Hauser**, **Hewett**, **Landy-Szalay**, **Davis-Peebles** y finalmente el menor pico lo presentó el estimador **Hamilton**. Finalmente, proponemos como mejor estimador a **Hewett** debido a que en diferentes circunstancias ha sido el segundo que más marcado muestra el pico BAO, aunado a que es junto con **Hamilton** de los que menos ruido presenta para grandes distancias.

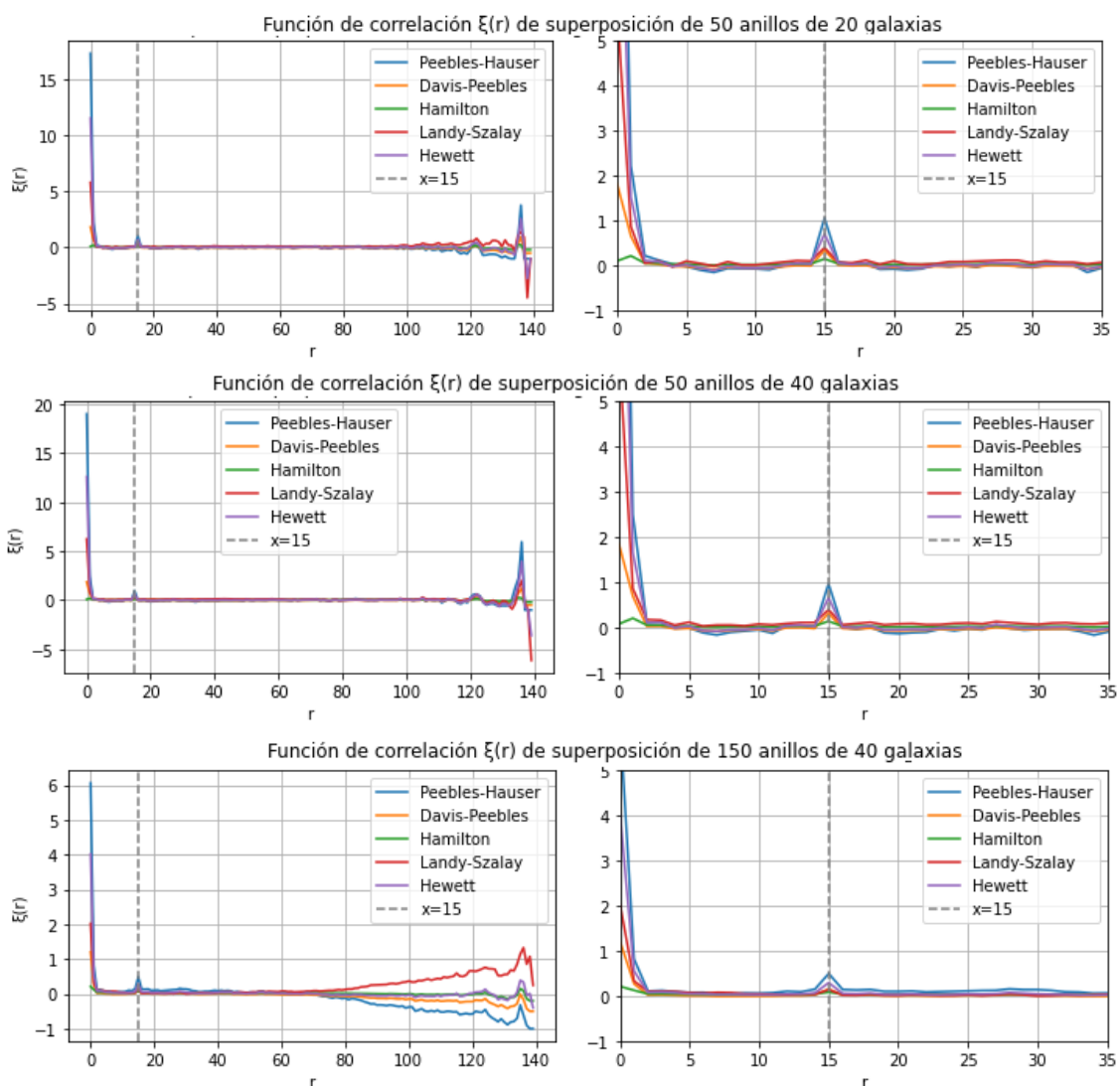


Figura 5.26: Función de correlación con distintos estimadores de una distribución de galaxias en un Universo cuadrado de 100 unidades de longitud de lado, con 10 % de galaxias en el centro respecto a las galaxias sobre el halo, perturbación del 1 % y formada (Arriba) 150 anillos con por lo menos 20 galaxias, (En medio) 50 anillos con 40 galaxias, (Abajo) 150 anillos con 40 galaxias cada uno.

5.4. RECONSTRUCCIÓN DEL PICO DE BAO ORIGINAL

En este capítulo se realizó un análisis de la función de correlación de dos puntos utilizando diferentes estimadores y diversas distribuciones de galaxias sintéticas. Como ejemplo final, se presenta la figura (5.28), que muestra cuatro cuadrantes con distribuciones de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado. Cada cuadrante contiene 150 anillos de BAO con un radio de 20 unidades de radio, y cada anillo tiene 40 galaxias, mas el 20 % de estas ubicadas en los centros de los anillos. Esta figura ilustra la reconstrucción del BAO cuando se calcula su función de correlación de dos puntos. Se realizó un acercamiento en la distribución y se colorearon las galaxias con diferentes colores según su nivel de perturbación

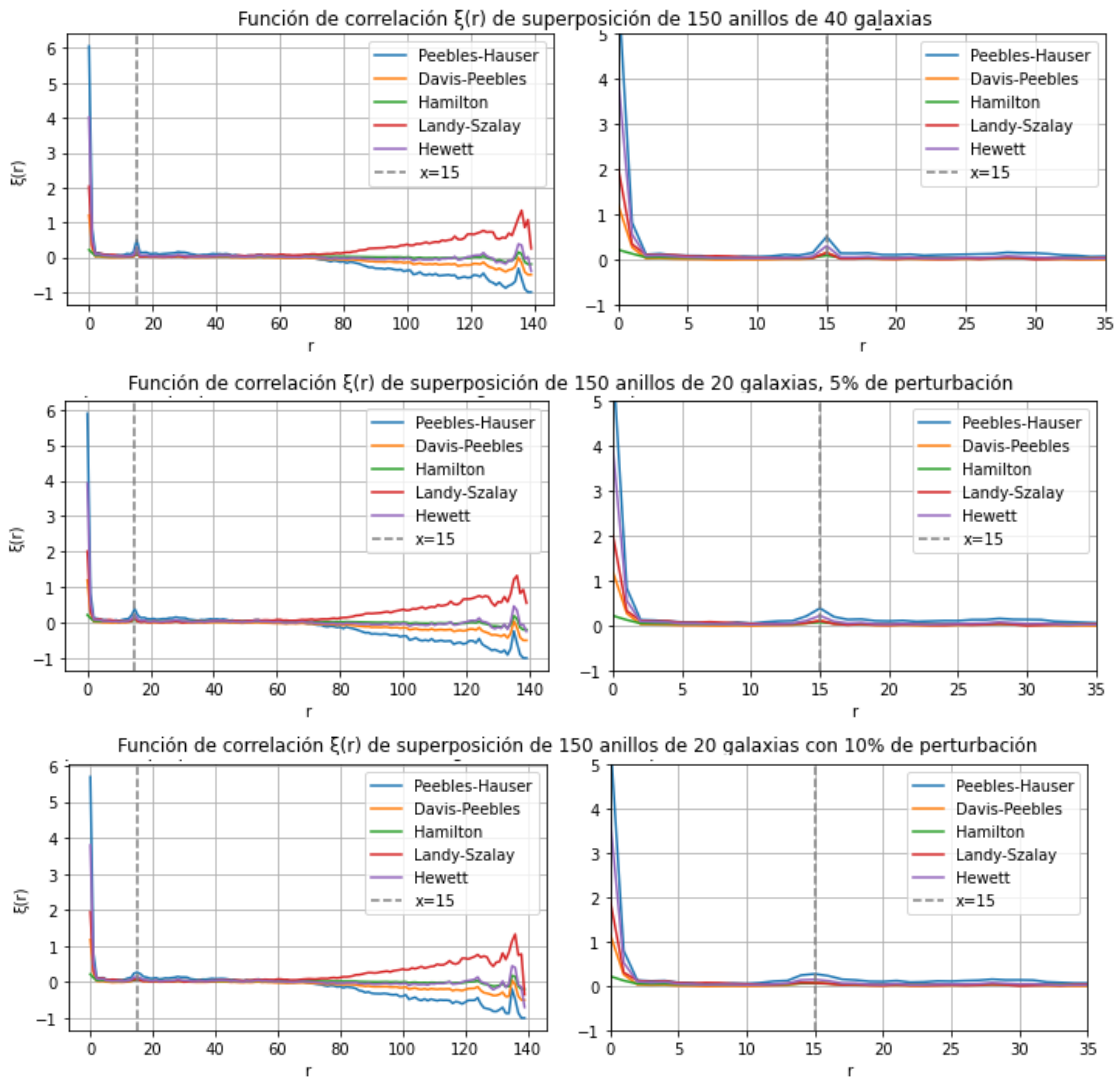
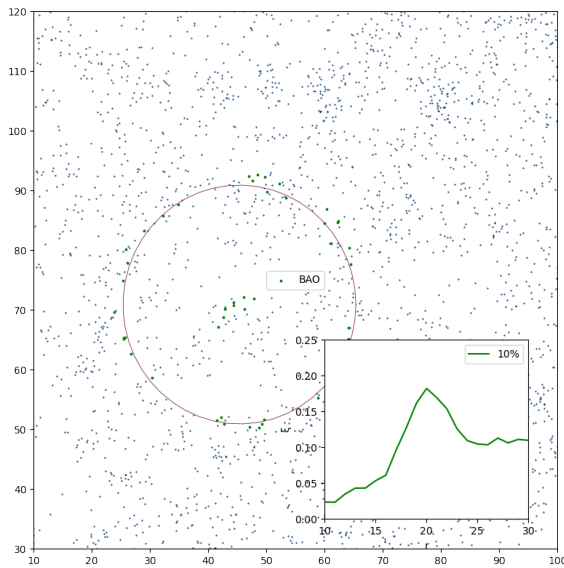


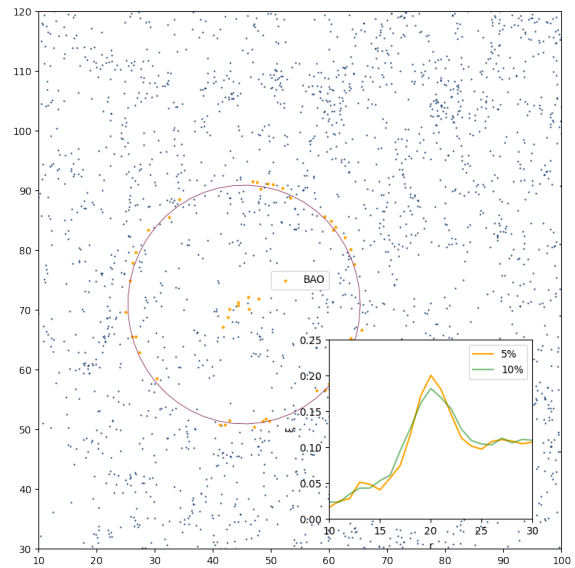
Figura 5.27: Función de correlación con distintos estimadores de una distribución de galaxias en un Universo cuadrado de 100 unidades de longitud de lado, con 10 % de galaxias en el centro respecto a las galaxias sobre el anillo, perturbación del 0 % y formada por (Arriba) 150 anillos con por lo menos 20 galaxias, (En medio) 50 anillos con 40 galaxias, (Abajo) 150 anillos con 40 galaxias cada uno.

(verde: 10 %, amarillo: 5 % y rojo: 0 %). El color morado representa el BAO original del cual las galaxias fueron desplazadas.

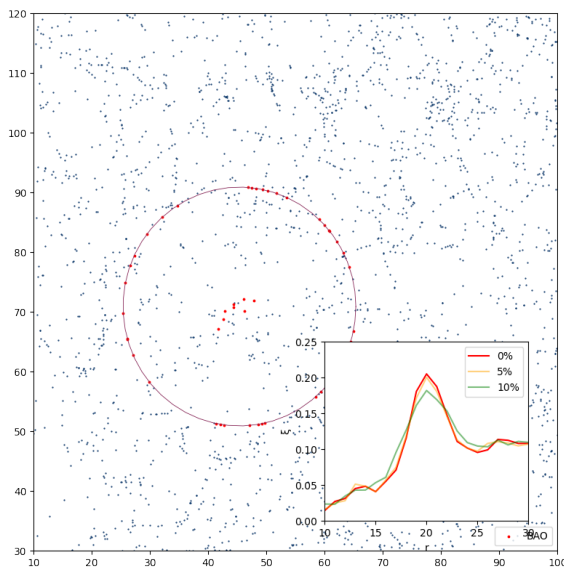
En el primer cuadrante, figura (5.28(a)), se observa una distribución bastante realista con una perturbación del 10 %. Se puede apreciar que las galaxias están significativamente desplazadas del BAO original. Será interesante analizar en futuros capítulos esta distribución de galaxias sintéticas con algoritmos de agrupamiento para detectar el centro. Sin embargo, este proceso puede resultar complicado debido a la presencia de numerosas galaxias cercanas al centro que no pertenecen a este BAO. Podemos apreciar que a pesar de la perturbación, la función de correlación es capaz de detectar muy bien la escala característica (radio) del BAO. En el segundo cuadrante, figura (5.28(b)),



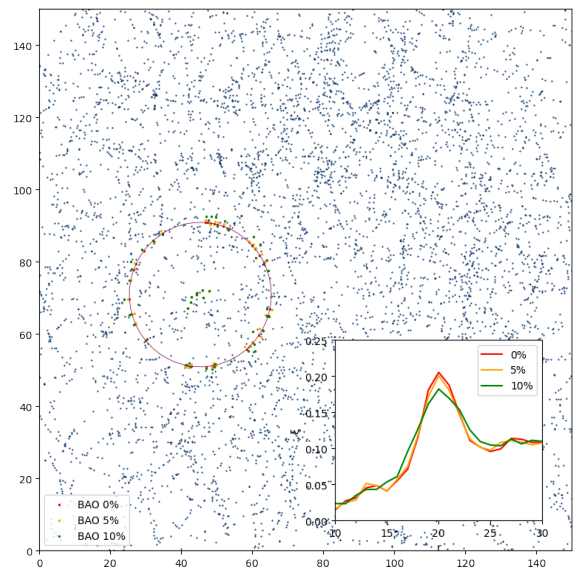
(a) 10% de perturbación.



(b) 5% de perturbación.



(c) 0% de perturbación.



(d) Distribución de la figura 5.28(a) con comparación de las funciones de correlación de dos puntos de todas las distribuciones de la figura 5.28.

Figura 5.28: Detección del radio del BAO con función de correlación de dos puntos. Todos los cuadrantes contienen distribuciones de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado, con 150 anillos de BAO de 20 unidades de radio, con 40 galaxias cada uno y 20% de galaxias en el centro respecto a las galaxias sobre el anillo.

se muestra una distribución similar pero con una perturbación menor (5%). A pesar que la perturbación disminuye, se observa que los centros permanecen inalterados, y en este caso, el pico del BAO aumenta su amplitud y se reduce su dispersión. En el tercer cuadrante, figura (5.28(c)), se presenta la reconstrucción del BAO original, donde las galaxias desplazadas se ajustan al BAO sin ninguna perturbación. Nuevamente, se

nota un aumento en la altura del pico del BAO y una disminución en su anchura, esto es, la distribución de galaxias se encuentra más concentrada alrededor del círculo.

Finalmente, en el último cuadrante, figura (5.28(d)), se muestra la distribución de la figura (5.28(a)) desde una perspectiva panorámica más amplia. Además, se incluyeron los datos de las galaxias que se encuentran en la misma región de sobredensidad resultantes de la oscilación acústica de bariones analizada para los casos de perturbación del 5% y 0%. Estas galaxias se representan en color rojo y amarillo, respectivamente, con el propósito de comparar sus funciones de correlación de dos puntos. A continuación, tanto en la distribución de las galaxias como en la función de correlación de dos puntos, se superponen los datos del BAO al que se hizo zoom con las distintas perturbaciones analizadas. Esto permite visualizar de manera esquemática la reconstrucción del BAO original, pasando de tener una perturbación del 10% (verde) a una perturbación del 5% (amarillo), y finalmente alcanzando una perturbación del 0% (rojo).

En conclusión, este capítulo demostró la importancia de analizar la función de correlación de dos puntos y utilizar distintos estimadores para estudiar las distribuciones de galaxias sintéticas. En la figura (5.28) se muestra cómo la perturbación, que afecta la posición de las galaxias, modifica la amplitud y la anchura del pico del BAO.

Capítulo 6

Algoritmos de agrupamiento con funciones de correlación

En este capítulo utilizaremos el algoritmo DBSCAN con el objetivo de encontrar los centros de los BAO y posteriormente reconstruir el pico de BAO con ayuda de las funciones de correlación de dos puntos previamente calculadas.

6.1. BÚSQUEDA DE CENTROS DE BAO CON DBSCAN

Es importante mencionar que difícilmente este algoritmo, o similares, detectaran todas las circunferencias sin embargo si se detecta una cantidad significativamente mayor de agrupamientos en la distribución de galaxias a comparación de los puntos aleatorios, sería un gran avance ya que demostraría que la distribución de galaxias no es aleatoria. A continuación se comparará el análisis con DBSCAN de una distribución de galaxias y una distribución aleatoria. Debido a lo complicado que sería detectar las circunferencias de los BAO, se buscará detectar sus centros.

En el panel izquierdo de la figura (6.1) podemos apreciar que de 200 posibles centros de BAO, el algoritmo DBSCAN con los parámetros (**eps=0.6, min_samples=4**) encontró 125 agrupamientos. Ahora analizaremos la distribución aleatoria con los mismos parámetros. Mientras que en la parte derecha de esta misma figura se puede observar que para los datos de las galaxias con un $eps = 0.6$ y un mínimo de 4 puntos por agrupamiento, DBSCAN detecta 125 clusters. Mientras que en la parte derecha de la figura, usando los puntos aleatorios con los mismos parámetros sólo detecta 20 agrupamientos. Esto nos indica que efectivamente nuestra distribución de “galaxias” no es aleatoria y que analizando mejor los parámetros y con ciertas modificaciones DBSCAN podría ser un buen algoritmo para esta tarea.

Para esto, se utilizó el siguiente comando:

```
1 from sklearn.cluster import DBSCAN
2 from sklearn import metrics
```

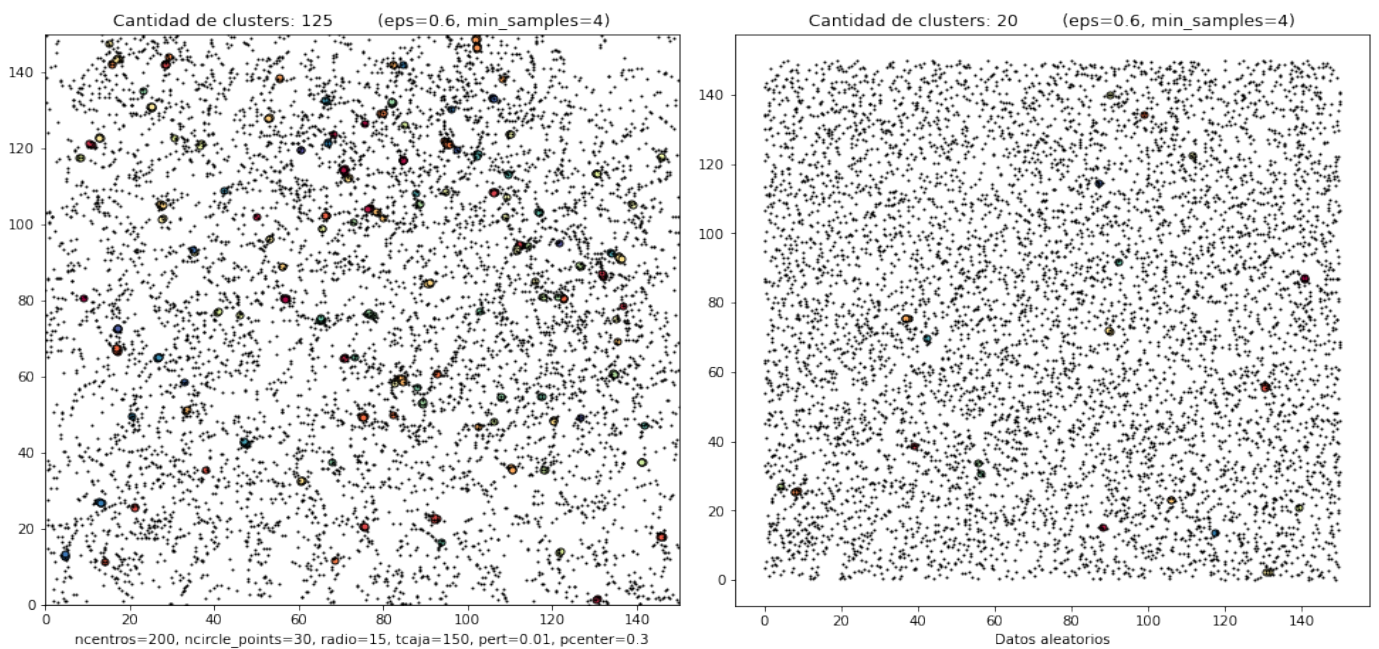


Figura 6.1: Izquierda: 7013 galaxias sintéticas distribuidas en 200 BAO de 15 unidades de longitud de radio con máximo 30 galaxias sintéticas cada una y 30% de galaxias sintéticas en el centro del BAO respecto a la circunferencia, en una caja cuadrada de 150 unidades de distancia de lado, analizados con DBSCAN, con los siguientes parámetros ($\text{eps}=0.6$, $\text{min_samples}=4$). Derecha: 7013 datos aleatorios, en una caja cuadrada de 150 unidades de longitud de lado, analizados con DBSCAN, con los siguientes parámetros ($\text{eps}=0.6$, $\text{min_samples}=4$).

```

3
4 # DBSCAN
5 db = DBSCAN(eps=0.6, min_samples=4).fit(data)
6 labels = db.labels_
7
8 # n mero de clusters, ignorando el ruido
9 n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
10 n_noise_ = list(labels).count(-1)

```

Con **labels** DBSCAN entrega una lista de la misma longitud que los datos donde cada término es la etiqueta numérica del agrupamiento al que pertenece el punto coordinado que pertenece a ese índice, **n_clusters_** nos entrega la cantidad de clústers que DBSCAN detectó sin considerar el ruido y **n_noise_** nos entrega la cantidad de datos que se consideran ruido.

6.2. RECONSTRUCCIÓN DE LA POSICIÓN ORIGINAL DE LAS GALAXIAS CON 2PCF Y DBSCAN

Con el fin de verificar si ubicamos el centroide de un BAO calcularemos el centro de masa de cada agrupamiento y posteriormente trazaremos circunferencias de radio 15 con el centro de masa como centro. Para esto se creó, **mydict**, un diccionario el cual nos otorga los índices correspondientes de los puntos de cada agrupamiento. Posteriormente, se creó un arreglo llamado **cluster_data** tal que para cada **cluster_data[i]**

entrega todos los puntos que pertenecen a ese agrupamiento.

```

1 # Para obtener los índices correspondientes de los puntos de cada cluster
2 mydict = {i: np.where(labels == i)[0] for i in range(n_clusters_)}
3
4 #Arreglo de datos que pertenecen al cluster
5 cluster_data = [] #Cada cluster_data[i] entrega todos los puntos que pertenecen a ese cluster
6 for j in mydict.values():
7     cluster_data.append(datos[j])

```

Nuestra distribución de galaxias conforman un sistema discreto, por lo que utilizamos el centro de masas como un punto geométrico que se mueve como una única partícula con la masa total del sistema. En el caso de un conjunto discreto de masas puntuales, el centro de masas es

$$\mathbf{r}_{\text{cm}} = \frac{\sum_i m_i \mathbf{r}_i}{\sum_i m_i} = \frac{1}{M} \sum_i m_i \mathbf{r}_i, \quad (6.1)$$

donde M es la masa total del sistema de partículas. m_i , masa de la partícula i -ésima. r_i , vector de posición de la masa i -ésima respecto al sistema de referencia supuesto. En este caso suponemos que $m_i = 1$ y procedemos a programar la función del centro de masa.

```

1 def centro_masa(n, r):
2     return sum(r)/n

```

Con ayuda de esta función creamos el arreglo **centros_clusters** en el que se guardarán las coordenadas del centro de cada agrupamiento.

```

1 centros_clusters = np.zeros((len(cluster_data),2))
2 for i in range(len(cluster_data)):
3     centros_clusters[i,0] = centro_masa(len(cluster_data[i][:,0]), cluster_data[i][:,0])
4     centros_clusters[i,1] = centro_masa(len(cluster_data[i][:,1]), cluster_data[i][:,1])

```

Ahora, si reconstruimos los BAO para los centros encontrados de la figura (6.1) obtenemos la figura (6.2), debido a la gran cantidad de datos que se están analizando no se puede apreciar que tantos centros fueron acertados y cuales no, sin embargo si se nota que muchos puntos están bastante distantes de cualquier BAO por lo que en la siguiente sección se analizarán distintas distribuciones para poder concluir si DBSCAN es una potencial herramienta útil.

6.3. ANÁLISIS DE DBSCAN CON DISTINTAS DISTRIBUCIONES DE GALAXIAS

En esta sección se llevarán a cabo pruebas utilizando el análisis de DBSCAN sobre diferentes distribuciones de galaxias sintéticas. El objetivo de estas pruebas es determinar la capacidad de DBSCAN para detectar los centros de los BAOs en función de la modificación de diferentes parámetros. Para esto, se modificarán parámetros como

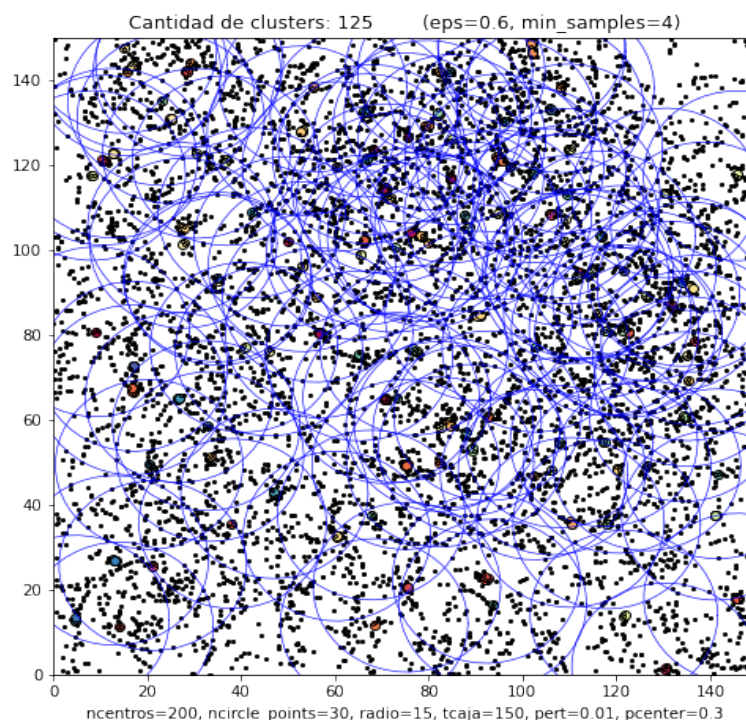


Figura 6.2: 7013 galaxias sintéticas distribuidas en 200 BAO de 15 unidades de longitud de radio con máximo 30 galaxias sintéticas cada una y 30% de galaxias sintéticas en el centro del BAO respecto a la circunferencia, en una caja cuadrada de 150 unidades de distancia de lado, analizados con DBSCAN, con los siguientes parámetros ($\text{eps}=0.6$, $\text{min_samples}=4$). Con reconstrucción de BAO.

la cantidad de BAOs, la perturbación o la cantidad de galaxias sobre la circunferencia del BAO, con el fin de analizar cómo afectan estos cambios en la detección de los centros por parte de DBSCAN. Con esto, se busca entender mejor la capacidad de DBSCAN para detectar estructuras cósmicas y su sensibilidad a los diferentes parámetros de las distribuciones de galaxias.

6.3.1. Modificando la cantidad de BAOs

Aquí analizaremos cómo el hecho de modificar la cantidad de BAOs afecta la detección de centroides por parte de DBSCAN. Comenzamos con 4 conjuntos de distribuciones de galaxias sintéticas que constan de diferentes cantidades de BAOs (30, 50, 75 y 100), con 50% de galaxias en el centro del BAO respecto a la circunferencia cuya dispersión del centro es de 1, y cuenta con 0% de perturbación. Esta distribución resulta poco realista debido a que, como se mencionó en el capítulo 1, como consecuencia de las interacciones gravitacionales, velocidades peculiares, choques entre otras cosas, las galaxias experimentan un grado de perturbación respecto al BAO original. Además, se está suponiendo que en el centro de los BAOs se acumula una gran cantidad de galaxias, lo cual no ha sido comprobado empíricamente y, por tanto, genera incertidumbre. En consecuencia, es importante considerar estas limitaciones al momento de interpretar los resultados obtenidos a partir de esta distribución de galaxias sintéticas.

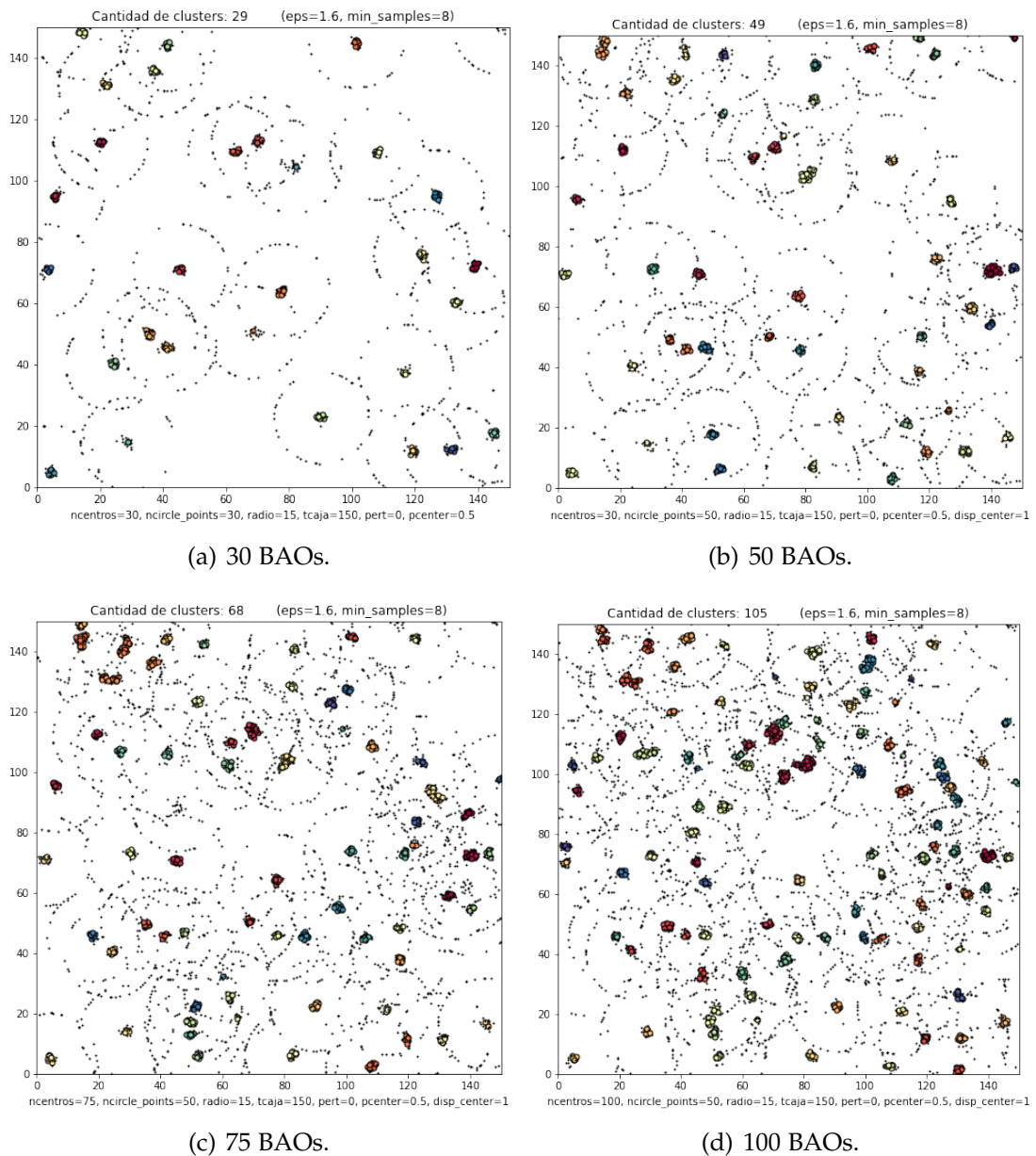
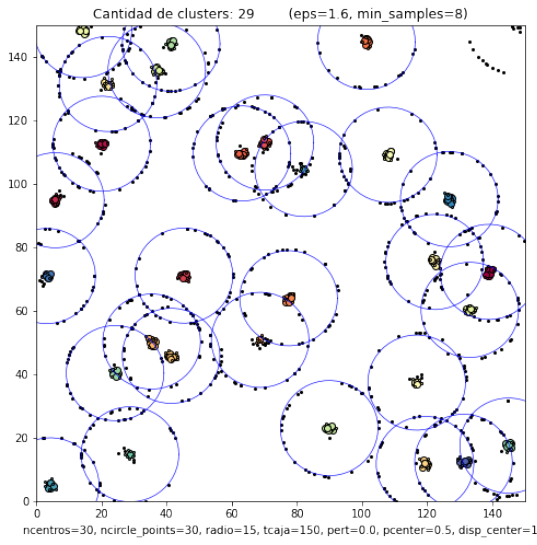
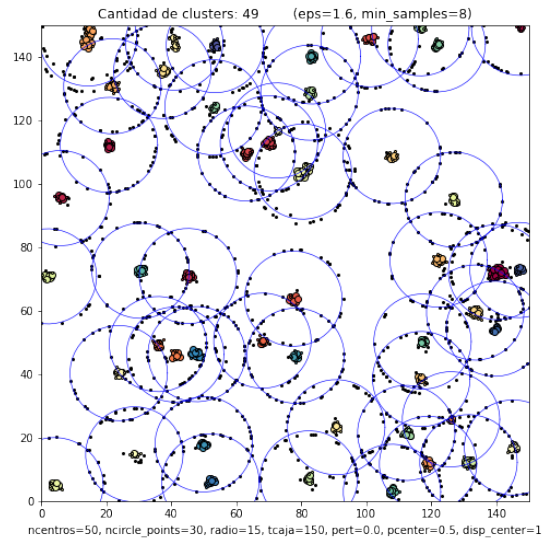


Figura 6.3: Datos de distintas cantidades de galaxias sintéticas con 50% de galaxias en el centro del BAO respecto a la circunferencia cuya dispersión es de 1, 0% de perturbación; analizados con DBSCAN, con los siguientes parámetros ($eps=1.6$, $min_samples=8$). Con agrupamientos detectados con DBSCAN.

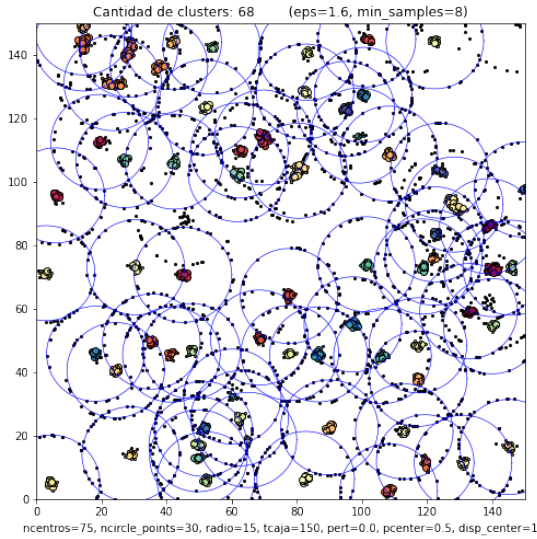
De la figura (6.3) podemos notar que las circunferencias de los BAO son apreciables a simple vista, sin embargo es importante comenzar con este tipo de distribuciones para ver que DBSCAN funcione para posteriormente pasar a distribuciones más realistas. Se observa que para 30 centros de BAO, DBSCAN detectó 29, debido a que un centro quedó en la esquina superior derecha lo cual nos muestra que para los datos en esquinas u orillas será complicado que DBSCAN los detecte. Para 50 centros de BAO, DBSCAN detectó 49 agrupamientos y como ahora la cantidad de galaxias sintéticas es mayor es difícil determinar a simple vista cual fue el BAO al que no



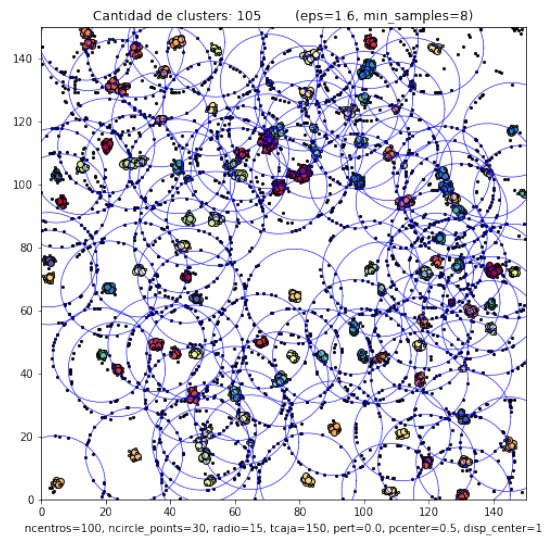
(a) 30 BAOs.



(b) 50 BAOs.



(c) 75 BAOs.



(d) 100 BAOs.

Figura 6.4: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.3). A partir de dicho análisis se procedió a trazar la reconstrucción del patrón de circunferencias de oscilaciones acústicas de bariones (BAO).

se le detectó su centro, por lo que nos servirá a continuación la reconstrucción de BAOs. Para 75 centros de BAO, DBSCAN detectó 68 agrupamientos y por último para 100 centros de BAO, DBSCAN detectó 105 agrupamientos lo cual no muestra que entre mayor cantidad de galaxias sintéticas se consideren, aumentará la posibilidad de que el algoritmo detecte como agrupamientos al traslape de galaxias en distintas circunferencias de BAO.

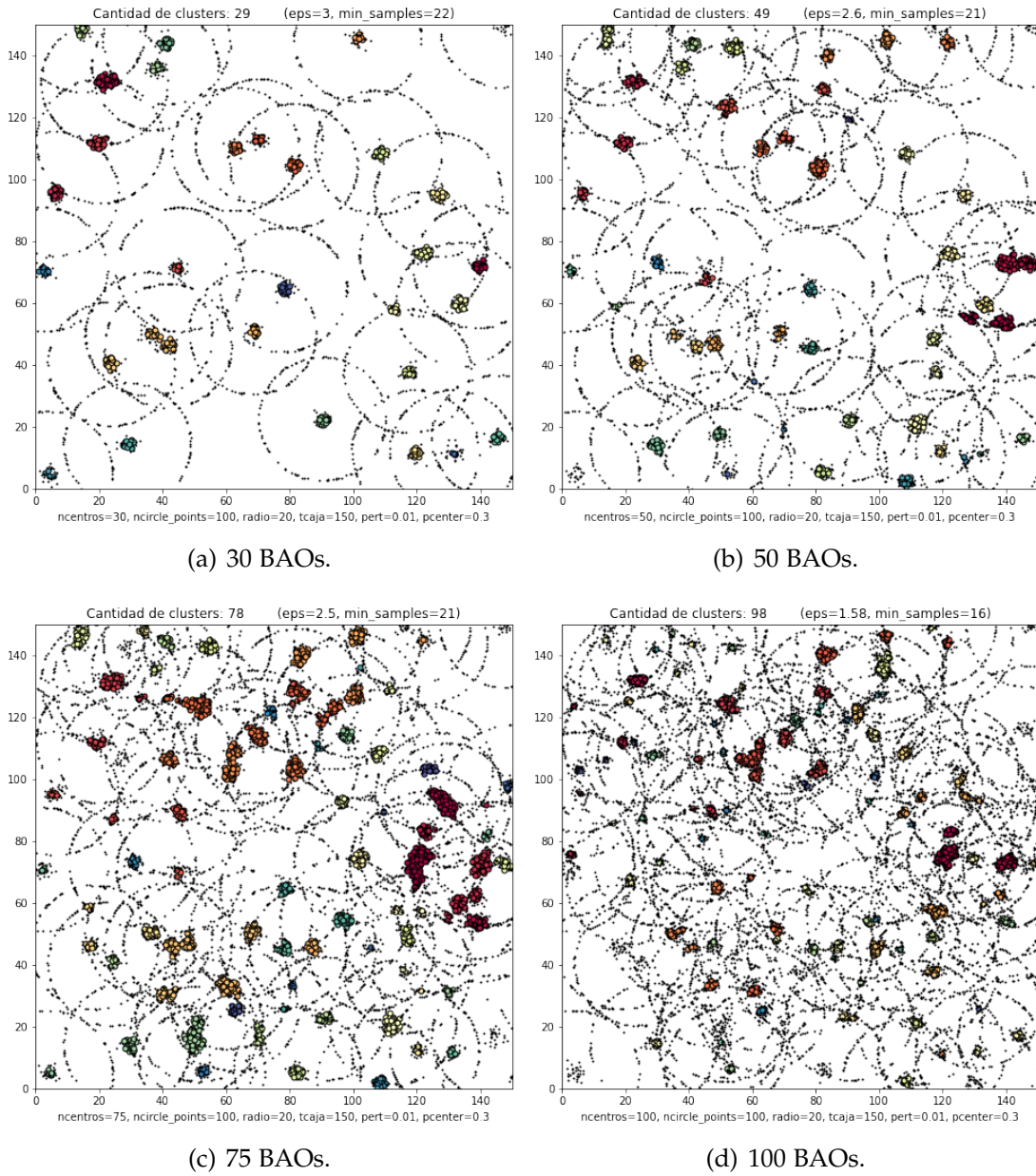


Figura 6.5: Distribuciones de distintas cantidades de galaxias sintéticas con 30% de galaxias en el centro del BAO respecto a la circunferencia de radio 20 unidades de longitud, perturbación de 1%, analizados con DBSCAN, con los siguientes parámetros ($eps=1$, $min_samples=8$).

En la figura (6.4) se muestra la reconstrucción del BAO original y podemos apreciar que DBSCAN detecta con bastante precisión los centros de los BAO. También podemos observar que para 50 centros de BAO, DBSCAN detectó como un mismo agrupamiento a dos centros distintos debido a que estaban bastante cerca (esquina superior izquierda).

A continuación trabajaremos con distribuciones más realistas, aumentaremos el tamaño del radio a 20 unidades de distancia, aumentaremos la cantidad de galaxias en las circunferencias de los BAOs, disminuirémos el porcentaje de puntos en el centro

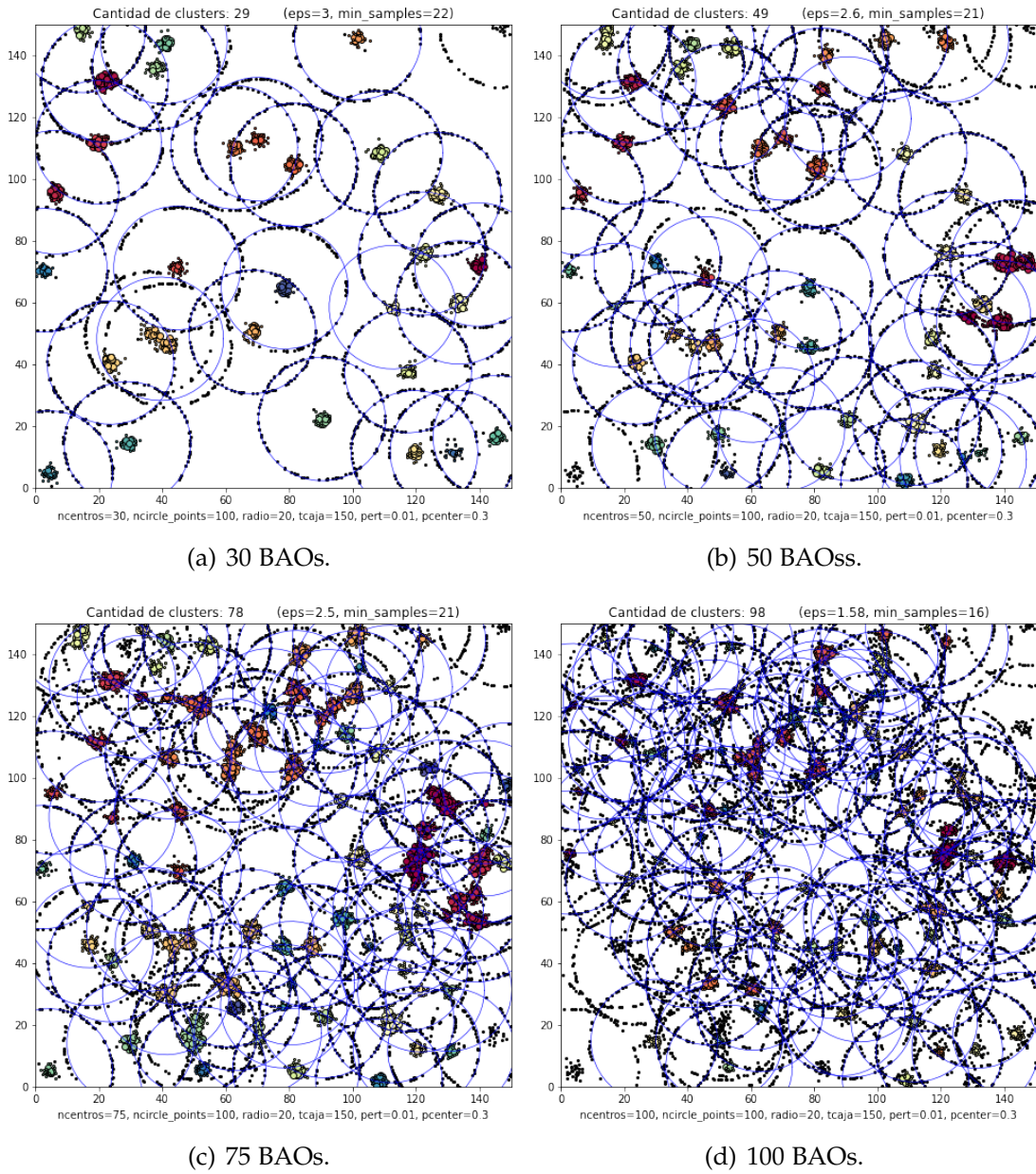


Figura 6.6: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.5). A partir de dicho análisis se procedió a trazar la reconstrucción del patrón de circunferencias de oscilaciones acústicas de bariones (BAO).

respecto a la circunferencia (ahora será de 30%), tendrán una perturbación de 1% y las galaxias de los centros tendrán dispersión de 1.5.

En la figura (6.5) aún se logran apreciar las circunferencias de los BAOs sin embargo ya para 100 centros de BAO vemos un panorama más realista.

En la figura (6.6) vemos que para 30 y 50 centros de BAO, detectó 29 y 49 grupa-

mientos respectivamente debido a que en ambas distribuciones un centro quedó muy cerca de la esquina. Aunque ya comienzan a existir confusiones, por ejemplo, en para 30 BAOs podemos notar que por las coordenadas (40,50) consideró a dos centros como uno sólo; luego en las coordenadas (115,60) consideró el traslape de dos círculos de BAO como un centro y finalmente, no consideró al centro ubicado en las coordenadas (150,150). A continuación, para 75 centros de BAOs, detectó 78 agrupamientos reafirmando que entre mayor cantidad de galaxias se consideren, DBSCAN comenzará a confundir en mayor medida traslape de circunferencias de BAOs con centros. Finalmente para 100 centros de BAO, detectó 98 agrupamientos de manera que aunque podría parecer contradictorio a lo que concluimos para 75 centros, vemos que llega un punto en que hay tantas galaxias que los argumentos comienzan a considerar varios centros en un mismo agrupamiento. Es importante mencionar que los parámetros tomados son manualmente por lo que es posible que no sean los óptimos.

6.3.2. Modificando la cantidad de galaxias sobre la circunferencia

En esta parte, procederemos a modificar la cantidad de galaxias que hay sobre las circunferencias de BAO, analizaremos distribuciones de 100 circunferencias de BAO con 20, 40, 60 y 80 galaxias máximas sobre cada circunferencia de BAO y veremos como esto afecta a la detección de centros con DBSCAN, es importante recordar que entre más galaxias se consideren sobre la circunferencia más galaxias habrá alrededor de los centros por como definimos nuestro generador de galaxias.

En las figuras (6.7) y (6.8) podemos apreciar que entre mayor sea la cantidad de galaxias sobre el BAO más sencillo será para DBSCAN detectar los centroides correctamente a pesar de que la cantidad de galaxias aumente en el espacio debido a que aumenta la cantidad de galaxias en cada centroide también.

6.3.3. Modificando la cantidad de galaxias y BAOs

A continuación se analizará como el aumentar o disminuir la cantidad de BAOs afecta la detección de centros con DBSCAN, así mismo veremos el efecto de aumentar o disminuir la cantidad de galaxias sobre la circunferencia del BAO.

En la figura (6.9) se muestran cuatro cuadrantes con diferentes distribuciones de galaxias. Los dos cuadrantes superiores contienen 50 círculos de BAO cada uno, donde cada círculo de BAO tiene un máximo de 25 y 50 galaxias en su interior. Por otro lado, los dos cuadrantes inferiores contienen 150 círculos de BAO cada uno, donde cada círculo de BAO tiene un máximo de 25 y 50 galaxias en su interior. En los cuadrantes a, b y d se pueden apreciar a simple vista los círculos de BAO, sin embargo en el cuadrante c es más complicado detectarlos debido a que hay pocas galaxias por círculo de BAO y muchos círculos de BAO, lo cual es una distribución más realista. Comenzamos nuestro análisis notando que para los cuadrantes con 50 circunferencias de BAO, DBSCAN detectó alrededor del 150 % de agrupamientos esperados mientras que los

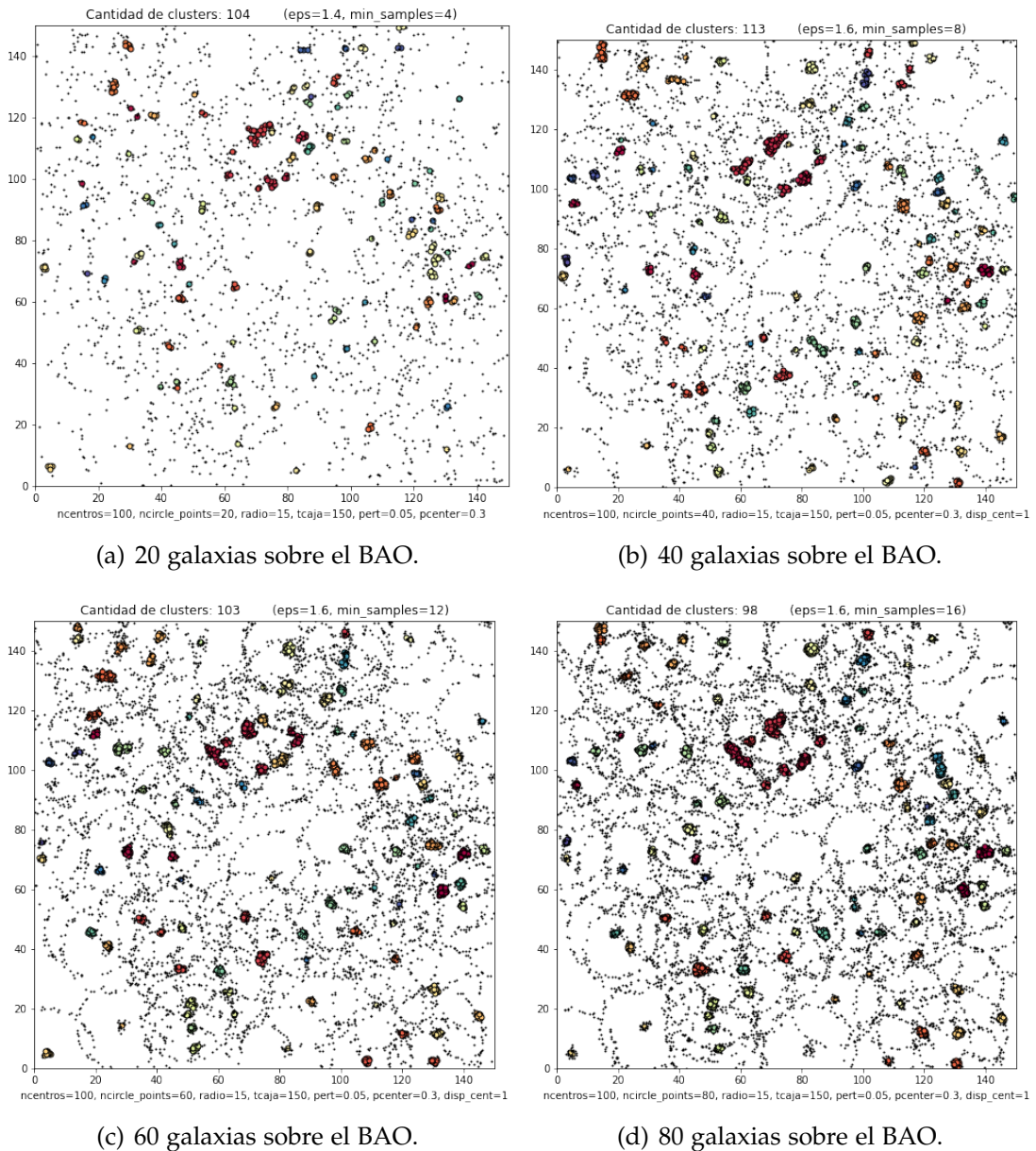


Figura 6.7: Distribuciones de distintas cantidades de galaxias sintéticas con 100 circunferencias BAO las cuales tienen 30 % de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud y perturbación de 5 %, analizados con DBSCAN.

cuadrantes con 150 circunferencias de BAO, detectó alrededor del 100 %; sin embargo es necesario ver la figura (6.10) donde se reconstruye el BAO para analizar que tan certero fue DBSCAN. En los cuadrantes a y b de la figura (6.10) se observa que DBSCAN encontró la mayoría de circunferencias BAO correctamente, pero es importante mencionar que hacia la parte superior y derecha debido a que las circunferencias de BAO están bastante cerca, éste detecta bastantes centros que no son correctos. También se puede apreciar que no detecta algunos centros que a simple vista se podría apreciar.

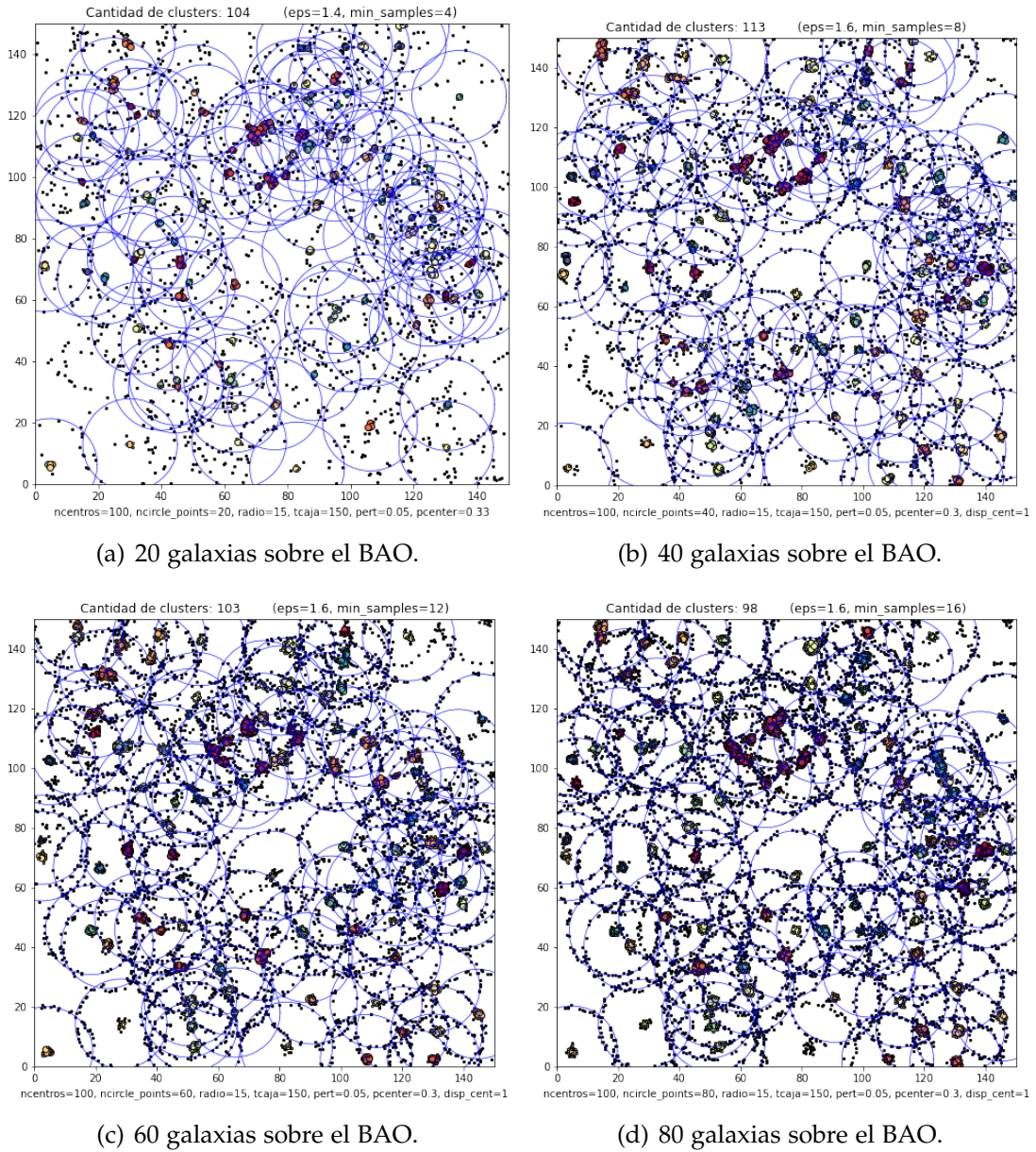


Figura 6.8: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.7). A partir de dicho análisis se procedió a trazar la reconstrucción del patrón de circunferencias de oscilaciones acústicas de bariones (BAO).

Por otro lado, en los cuadrantes c y d de la figura (6.10) a simple vista no es tan sencillo observar las circunferencias de BAO y al reconstruir los BAO originales vemos que el algoritmo detecta correctamente una gran parte.

6.3.4. Modificando la perturbación de galaxias sobre la circunferencia

Por último, analizaremos como modificar la perturbación afecta la detección de centroides. En la figura (6.11) vemos que para el primer cuadrante con 0% de pertur-

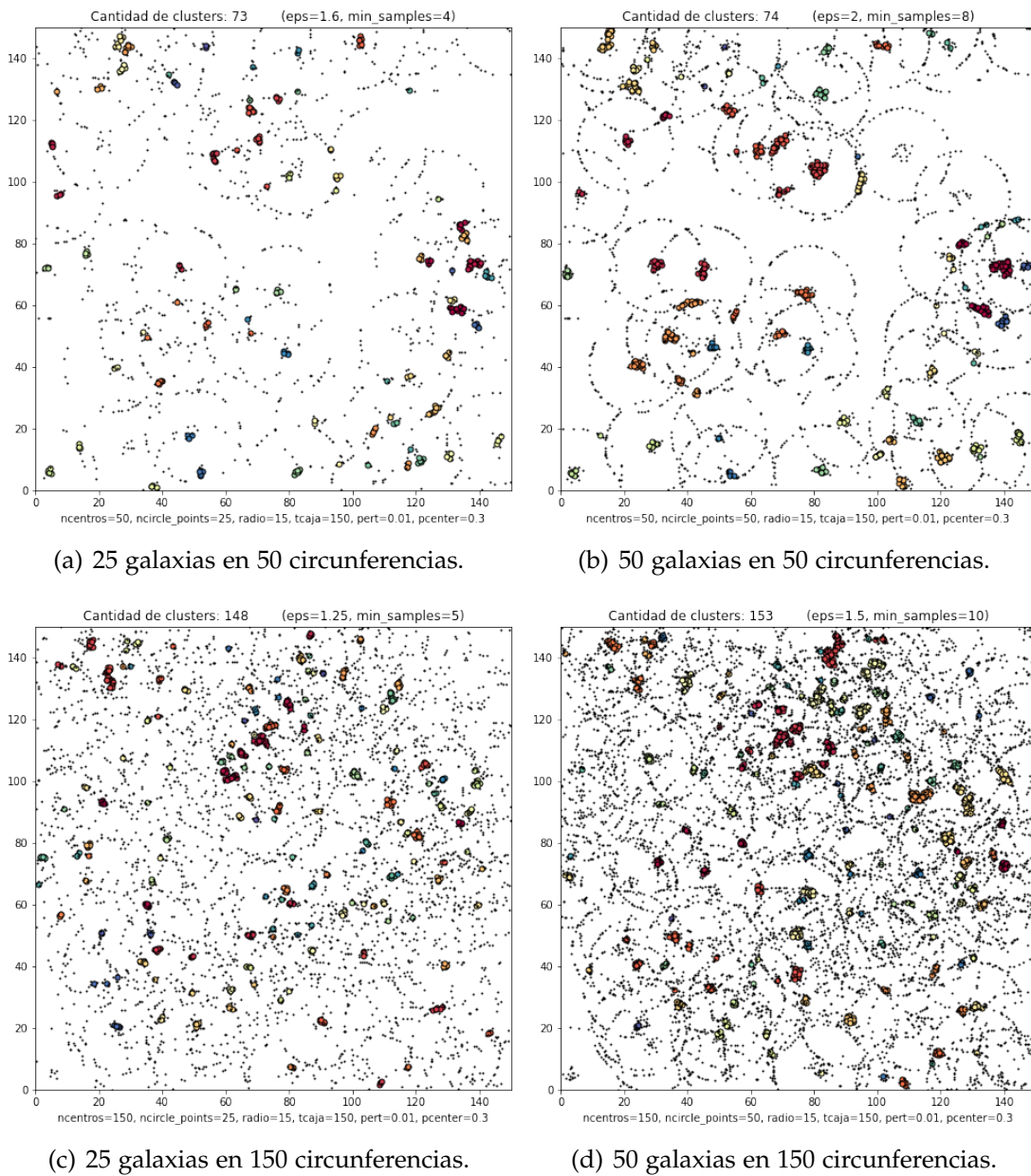


Figura 6.9: Distribuciones de distintas cantidades de galaxias sintéticas con 30% de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud, perturbación de 1%, analizados con DBSCAN.

bación es sencillo detectar las circunferencias de BAO, sin embargo conforme aumenta la perturbación se va complicando cada vez más esta detección para el ojo humano. Ahora en la figura (6.11) podemos ver que para todos los cuadrantes DBSCAN detectó la mayoría de las circunferencias de BAO correctamente.

Finalmente, podemos concluir que DBSCAN es un algoritmo bastante prometedor para la detección de centros de BAO, debido a que detecto gran parte de éstos. Es importante considerar que DBSCAN busca agrupamientos con densidades iguales por

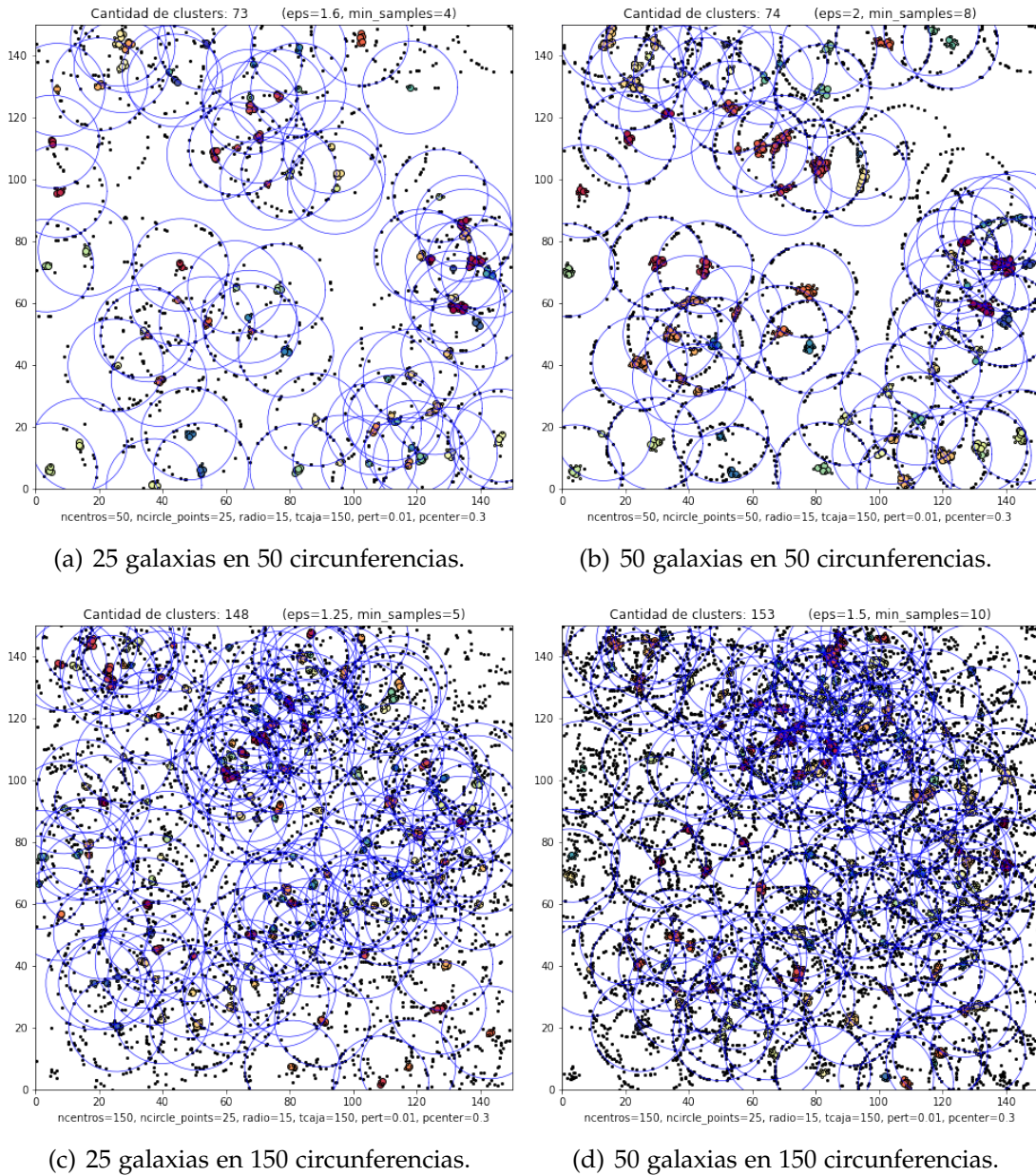


Figura 6.10: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.9). A partir de dicho análisis se procedió a trazar la reconstrucción del patrón de circunferencias de oscilaciones acústicas de bariones (BAO).

lo que incluso podría funcionar para clasificación utilizando diferentes parámetros de densidades. Así mismo, no es un algoritmo perfecto, es bastante sensible a los parámetros que le otorgamos manualmente, es posible que considere traslape de galaxias de diferentes circunferencias de BAO como centroides, o que considere a dos centroides cercanos como uno sólo.

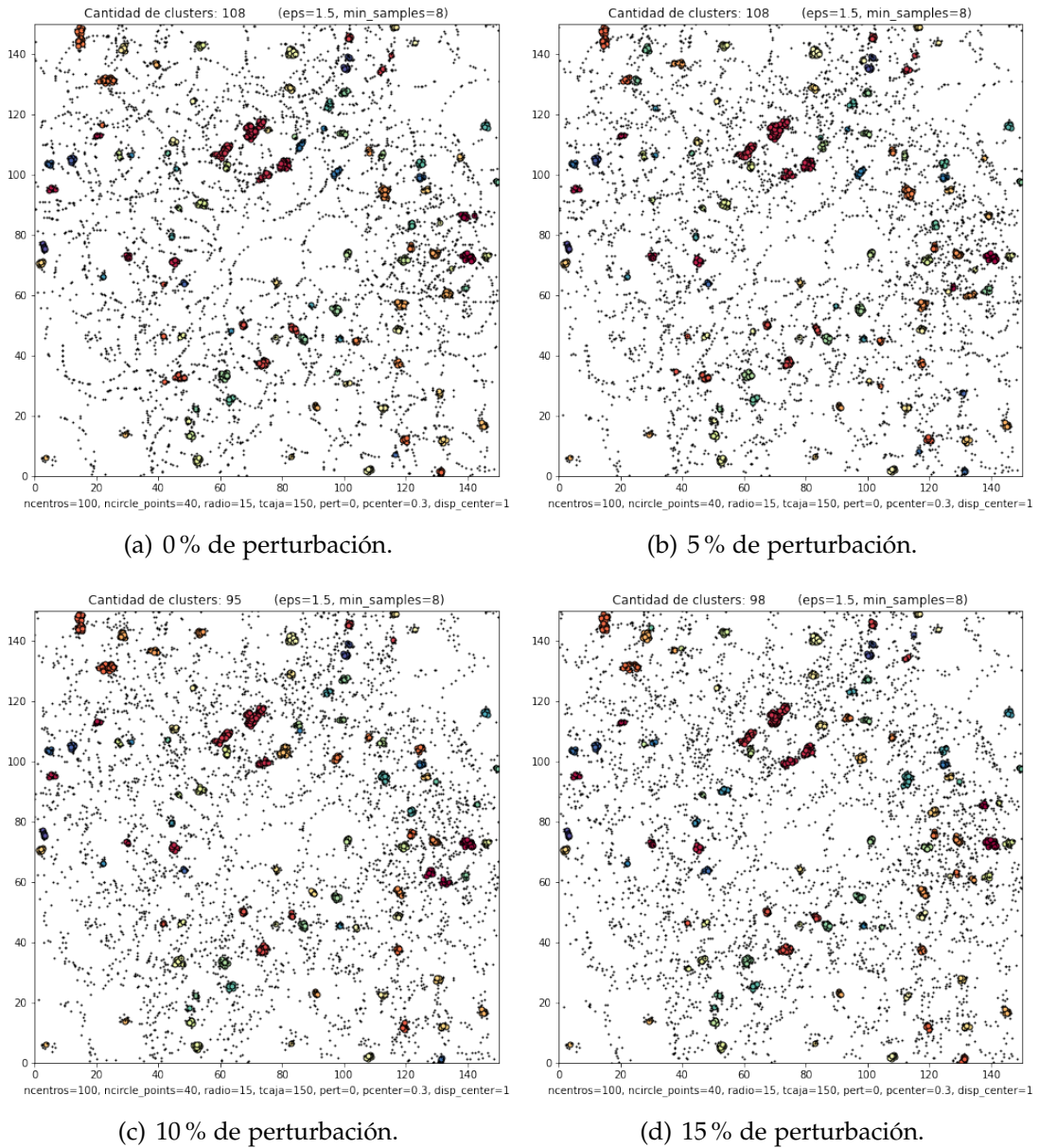


Figura 6.11: Distribuciones de distintas cantidades de galaxias sintéticas con máximo 40 galaxias sobre la circunferencia del BAO y 30 % de galaxias en el centro del BAO respecto a la circunferencia de radio 15 unidades de longitud, analizados con DBSCAN.

6.4. RECONSTRUCCIÓN DEL BAO ORIGINAL CON ALGORITMOS DE AGRUPAMIENTO

En el capítulo 5 de esta tesis, se obtuvo el pico de la función de correlación de dos puntos, a partir de la cual se obtuvo el radio característico de las circunferencias en las que en promedio se agrupan las galaxias debido a las oscilaciones acústicas de bariones. Durante este capítulo, se demostró que es posible detectar agrupamientos de galaxias (que pueden o no ser los centros de los BAO) de una distribución utilizando algoritmos

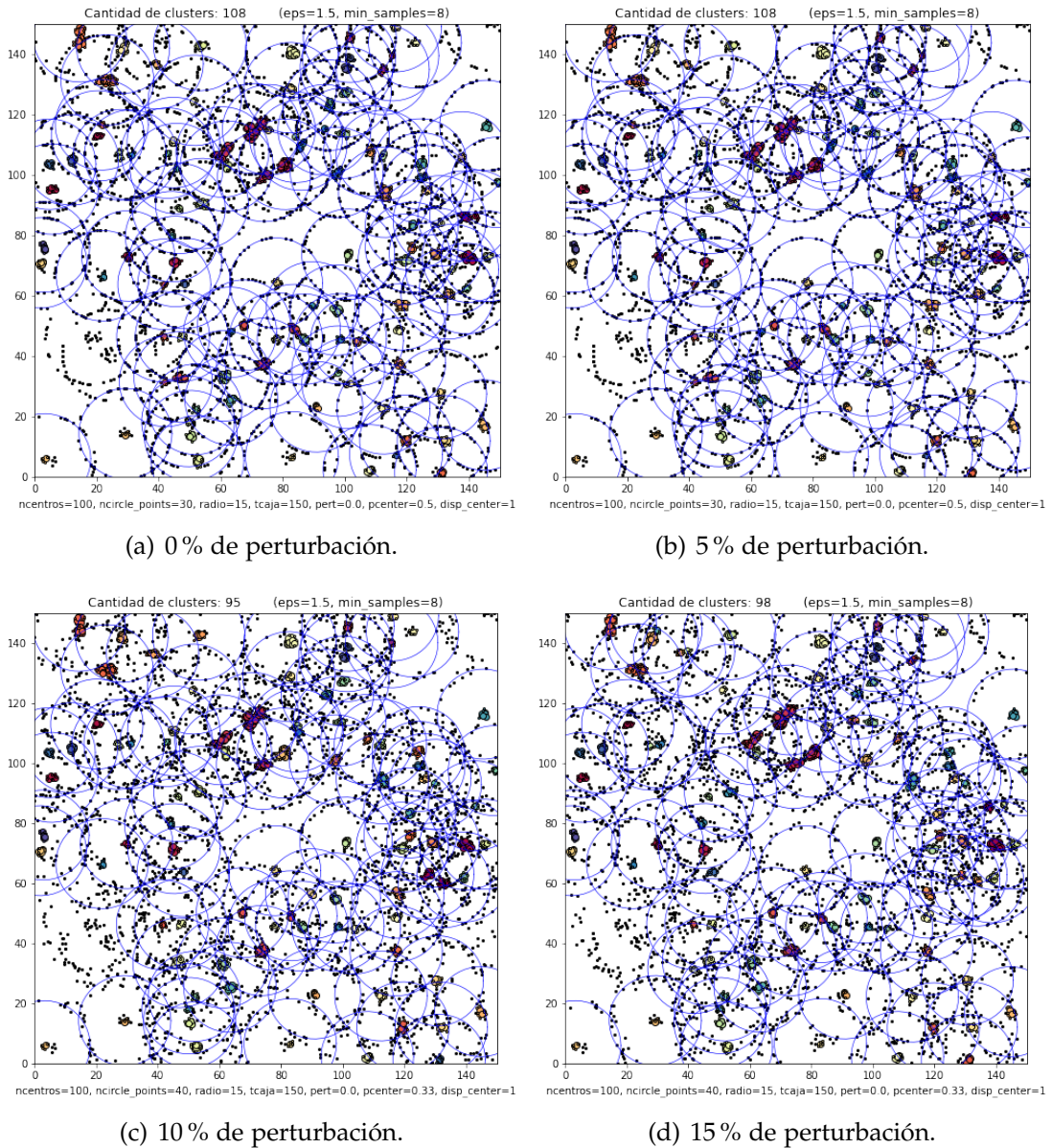


Figura 6.12: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.11). A partir de dicho análisis se procedió a trazar la reconstrucción del patrón de circunferencias de oscilaciones acústicas de bariones (BAO).

de agrupamiento basados en densidad, específicamente el algoritmo DBSCAN. En adelante, se explorará la posibilidad de reconstruir el BAO original de la distribución de galaxias sintéticas, determinando el radio característico utilizando algoritmos de agrupamiento. Para lograr este objetivo, supondremos que los agrupamientos detectados son centros de los BAO, utilizaremos tres anillos diferentes, conectados en sus extremos, de manera que todos los centros de masa de los agrupamientos serán considerados centros de los BAO. Denominaremos ring1 al anillo más pequeño con un radio inferior de $liminf$ y un radio superior de r . Luego, el anillo mediano

lo llamaremos *ring2*, y tendrá un radio inferior de *r* y un radio superior de *R*. Por último, el anillo más grande, *ring3*, tendrá un radio inferior de *R* y un radio superior de *limsup*. A continuación, recordemos que el área de un anillo con radio inferior *r* y radio superior *R* está dado por la siguiente ecuación,

$$Area_{anillo} = R^2\pi - r^2\pi = \pi(R^2 - r^2). \quad (6.2)$$

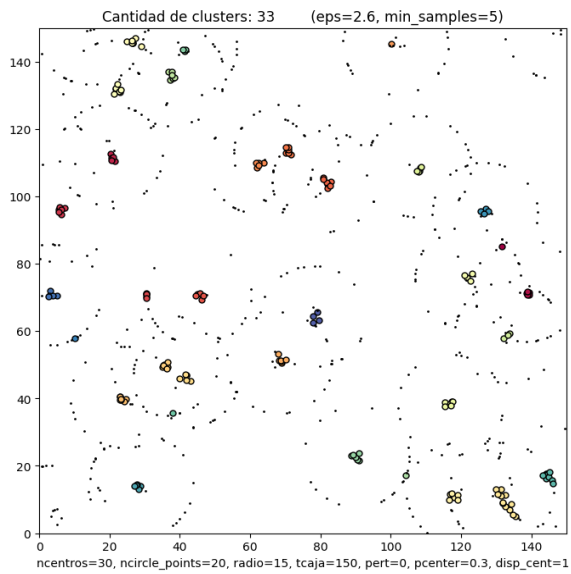
En nuestro caso, para un adecuado análisis requerimos que todos los anillos tengan la misma área por lo que se define *liminf* y *limsup* de la siguiente forma.

```
1 liminf = np.sqrt((r**2)-(ring_area(R,r)/np.pi))
2 limsup = np.sqrt((R**2)+(ring_area(R,r)/np.pi))
```

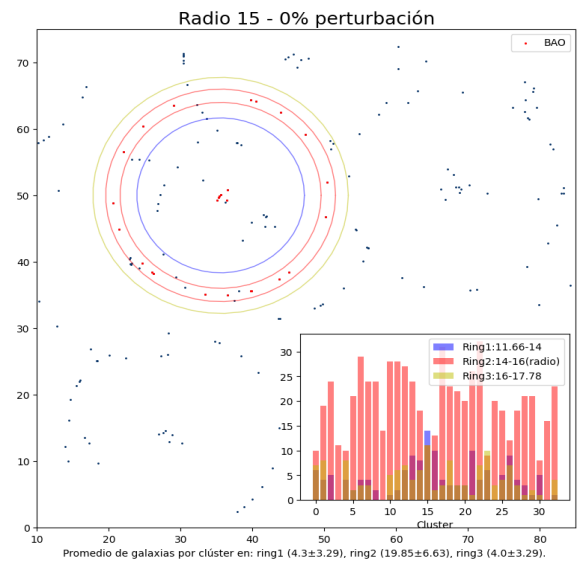
Ahora bien, nuestro interés se centra en determinar qué anillo contiene una densidad mayor número de galaxias sintéticas. Para lograrlo, utilizamos la distancia Euclideana (definida en el capítulo 5), la cual nos permitirá calcular qué galaxias cumplen con la condición de que su distancia desde el centro de masa de un agrupamiento de galaxias hasta la galaxia esté comprendida dentro de alguno de los anillos mencionados. Para cada agrupamiento detectado con DBSCAN se calculo cuantas galaxias hay dentro de cada anillo y se fueron guardando en los arreglos *ring1*, *ring2* y *ring3* respectivamente como se muestra a continuación.

```
1 ring1 = np.zeros(len(centros_clusters))
2 ring2 = np.zeros(len(centros_clusters))
3 ring3 = np.zeros(len(centros_clusters))
4 ring1_points = []
5 ring2_points = []
6 ring3_points = []
7 for i in range(len(centros_clusters)):
8     for j in range(len(datos)):
9         if liminf < distancia(centros_clusters[i], datos[j]) < r:
10             ring1[i] +=1
11             ring1_points.append(datos[j])
12         elif r <= distancia(centros_clusters[i], datos[j]) <= R:
13             ring2[i] +=1
14             ring2_points.append(datos[j])
15         elif R < distancia(centros_clusters[i], datos[j]) < limsup:
16             ring3[i] +=1
17             ring3_points.append(datos[j])
18
19 clusters = []
20 for i in range(len(centros_clusters)):
21     clusters.append(i)
```

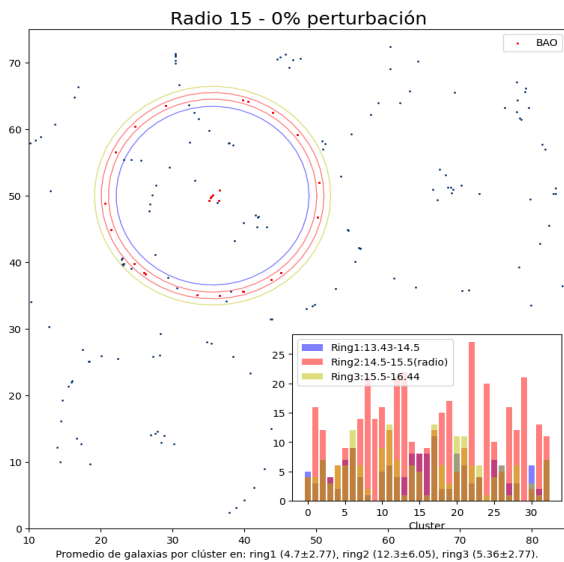
En la figura (6.13), se presenta una propuesta de detección del pico de la función de correlación de las BAO utilizando sólo algoritmos de agrupamiento. Comenzando con la figura 6.13(a), donde se muestra la detección de centros utilizando el algoritmo DBSCAN en una distribución de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado. En esta configuración, se tienen 30 anillos de BAO con un radio de 15 unidades de longitud, con 20 galaxias en cada anillo y una proporción del 30% de galaxias en el centro en comparación con las galaxias sobre el anillo. Además, no se aplicó ninguna perturbación en esta configuración. Observamos que se



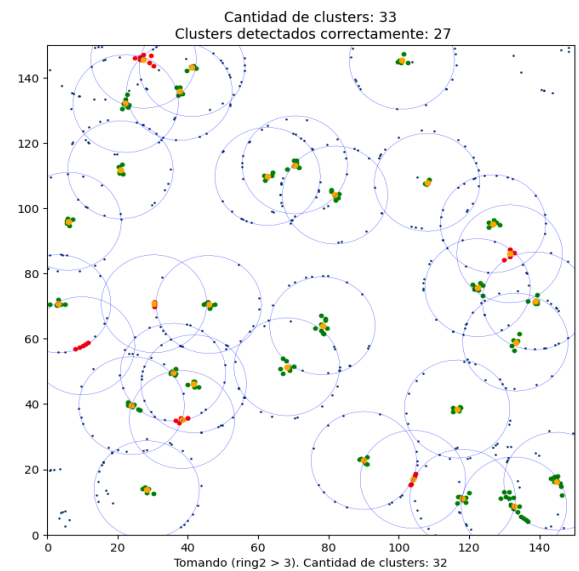
(a) Detección de centros con DBSCAN.



(b) Cálculo de cantidad de galaxias sintéticas en anillos de diferentes radios.



(c) Refinamiento de la figura 6.13(b).



(d) Trazo de BAO a partir de los centros.

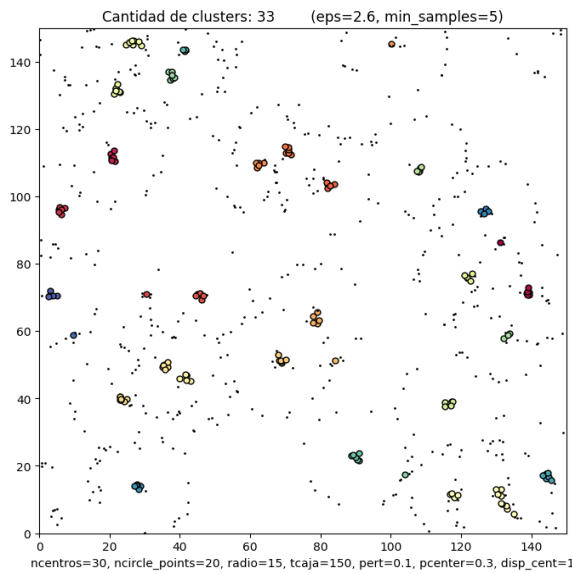
Figura 6.13: Reconstrucción del BAO original con algoritmos de agrupamiento, de una distribución de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado, con 30 anillos de BAO de 15 unidades de longitud de radio, con 20 galaxias cada uno y 30% de galaxias en el centro respecto a las galaxias sobre el anillo.

detectaron 33 agrupamientos, de los 30 centros de BAO esperados. Sin embargo, al inspeccionar visualmente la distribución, se observa que 3 centros cercanos a los bordes no fueron detectados (uno en la esquina inferior izquierda, otro en la esquina superior izquierda y un último en la esquina superior derecha). También se puede notar que en las coordenadas (105,15), (130, 85) y (40,35) se consideraron como agrupamien-

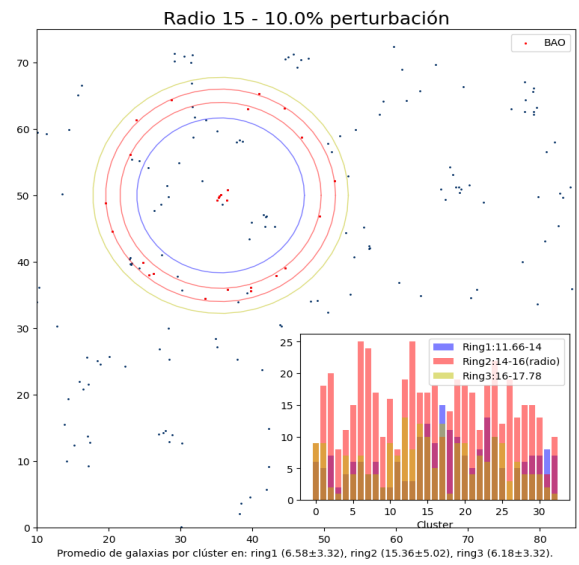
tos o centros debido a la superposición de galaxias sobre las circunferencias de los BAO.

Luego, en la figura 6.13(b), se presenta el análisis de los centros utilizando diferentes anillos con el fin de determinar el radio del BAO, que suponemos desconocido. Para ilustrar este proceso, se realizó un acercamiento a un BAO específico, donde las galaxias del BAO están representadas en color rojo. Utilizando el algoritmo DBSCAN, se detectó el centro, y se observa que diferentes galaxias caen sobre cada anillo, pero todas las galaxias pertenecientes a la circunferencia del BAO se encuentran dentro del rango de radio de 14 a 16 unidades de longitud. En la esquina inferior izquierda se muestra un histograma que representa el número de galaxias que caen dentro de cada anillo en función de los centros detectados por DBSCAN (en este caso, se detectaron 33 centros). Se puede apreciar que el anillo con un radio de 14 a 16 unidades de longitud contiene predominantemente más galaxias. Al calcular el promedio de la cantidad de galaxias por agrupamiento en toda la distribución, se obtuvo que para el Ring1 (azul) hubo en promedio (4.3 ± 3.29) galaxias por agrupamiento, para el Ring2 (rojo) hubo en promedio (19.85 ± 6.63) galaxias por agrupamiento, y para el Ring3 (amarillo) hubo en promedio (4.0 ± 3.29) galaxias por agrupamiento. De esto se concluye que el radio del BAO se encuentra efectivamente entre 14 y 16 unidades de longitud. Además, en la figura 6.13(c) se muestra un análisis similar al segundo cuadrante, pero con un refinado en el rango de radio, reduciendo el área de exploración. En este caso, se obtuvo que en promedio hubo (4.7 ± 2.77) galaxias por agrupamiento en el Ring1 (azul), (12.3 ± 6.05) galaxias por agrupamiento en el Ring2 (rojo) y (5.36 ± 2.77) galaxias por agrupamiento en el Ring3 (amarillo). Estos resultados nos permiten concluir que el radio del BAO se encuentra efectivamente entre 14.5 y 15.5 unidades de longitud. Finalmente, en la figura 6.13(d) se muestra la detección de los BAO trazada en azul, con un radio de 15 unidades de longitud. Esta representación visual nos permite apreciar la reconstrucción del BAO original. Además, se muestran los agrupamientos detectados en dos colores, rojo si no son un centro de BAO y verde si efectivamente pertenecen a un centro de BAO. Podemos ver que de los 33 agrupamientos detectados, 27 fueron centros de BAO. Se detectó un $\frac{27}{30} = 0.9 = 90\%$ de los esperados con una precisión de $\frac{27}{33} = 0.818 = 81.8\%$. Una posible forma de mejorar la precisión sería analizando el histograma presentado en la figura 6.13(c), por ejemplo en este caso se excluirán a los agrupamientos que en su anillo ring2 cuenten con menos de 3 galaxias. Así que se marcan con amarillo los agrupamientos que cumplen con este criterio, podemos observar que fueron 32 y el agrupamiento ubicado alrededor de (10,57), el cual era incorrecto, ya no se considera. De esta manera aumenta la precisión a $\frac{27}{32} = 0.8437 = 84.37\%$.

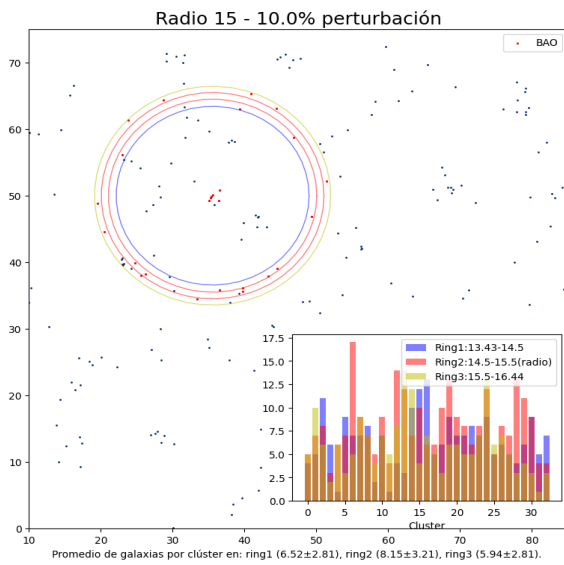
En la Figura 6.14, se realizó el mismo procedimiento que en la Figura 6.13, pero en este caso se agregó un 10% de perturbación a las galaxias sobre las circunferencias de BAO. En la parte superior izquierda de 6.14(a), podemos apreciar la importancia de la perturbación, ya que a pesar de la pequeña cantidad de BAOs, la distribución comienza a verse más realista. En este caso, se detectaron 33 agrupamientos de los



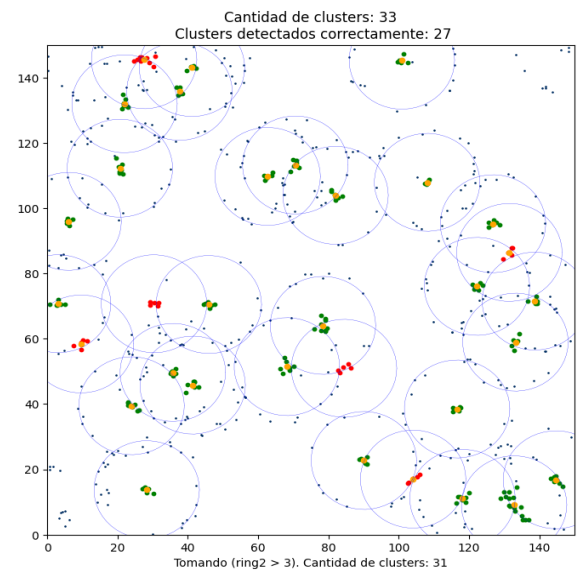
(a) Detección de centros con DBSCAN.



(b) Cálculo de cantidad de galaxias sintéticas en anillos de diferentes radios.



(c) Refinamiento de la figura 6.14(b).



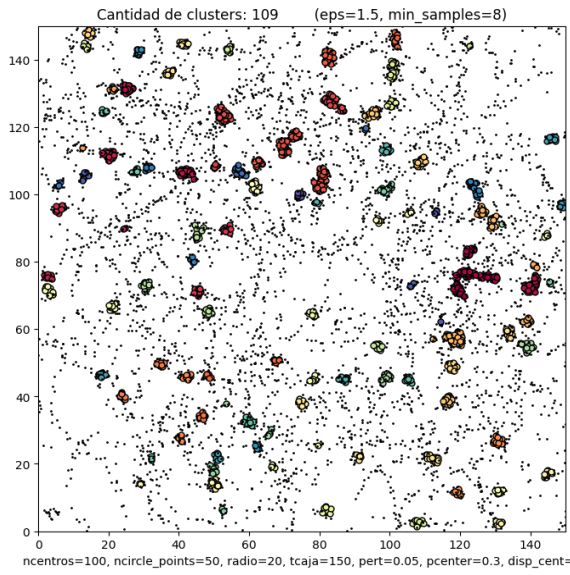
(d) Trazo de BAO a partir de los centros.

Figura 6.14: Reconstrucción del BAO original con algoritmos de agrupamiento. Todos los cuadrantes contienen distribuciones de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado, con 30 anillos de BAO de 15 unidades de longitud de radio y 10% de perturbación, con 20 galaxias cada uno y 30% de galaxias en el centro respecto a las galaxias sobre el anillo.

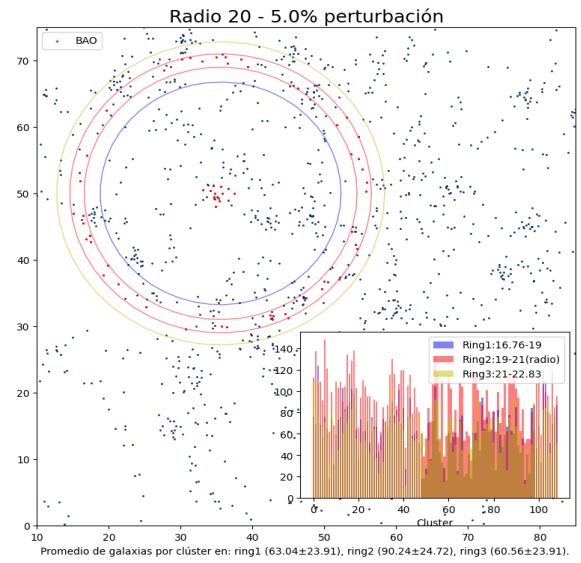
30 posibles centros de BAO. Es importante destacar que los centros detectados no coinciden con los de la Figura 6.13(a). En 6.14(b) se ilustra la dispersión de las galaxias pertenecientes al clúster (en rojo). Ya no se encuentran solo en el anillo del medio, sino que están más dispersas. Sin embargo, en el histograma podemos observar

que la cantidad de galaxias predominantes se encuentra en el ring2 (rojo). En este caso, en promedio hubo (6.58 ± 3.32) galaxias por agrupamiento en el Ring1 (azul), (15.36 ± 5.02) galaxias por agrupamiento en el Ring2 (rojo) y (6.18 ± 3.32) galaxias por agrupamiento en el Ring3 (amarillo). Estos resultados nos llevan a concluir que el radio del BAO se encuentra efectivamente entre 14 y 16 unidades de longitud. Luego, en la Figura 6.14(c), procedemos a refinar el área de los anillos para encontrar el radio característico y observamos en el histograma que hay agrupamientos donde hay más galaxias en el ring1 (azul), y el predominio del ring2 (rojo) es casi imperceptible, esto se puede deber a que las galaxias presentan gran perturbación o el refinamiento fue demasiado grande, sin embargo sigue sobresaliendo levemente el anillo ring2. En este caso, en promedio hubo (6.52 ± 2.81) galaxias por agrupamiento en el Ring1 (azul), (8.15 ± 3.21) galaxias por agrupamiento en el Ring2 (rojo) y (5.94 ± 2.81) galaxias por agrupamiento en el Ring3 (amarillo). Estos resultados nos permiten concluir que el radio del BAO se encuentra efectivamente entre 14.5 y 15.5 unidades de longitud. Posteriormente, en la figura 6.14(d) se muestra la reconstrucción de los BAOs con radio 15, y vemos que de los 33 agrupamientos detectados, 27 fueron centros de BAO. Se detectó un $\frac{27}{30} = 0.9 = 90\%$ de los esperados con una precisión de $\frac{27}{33} = 0.818 = 81.8\%$. Buscando mejorar la precisión se excluirán a los agrupamientos que en su anillo ring2 cuenten con menos de 4 galaxias. Así que se marcan con amarillo los agrupamientos que cumplen con este criterio, podemos observar que fueron 31 y los agrupamientos ubicados alrededor de (30,70) y (85,50), los cuales eran incorrectos, ya no se consideran. Así aumenta la precisión a $\frac{27}{31} = 0.8709 = 87.09\%$.

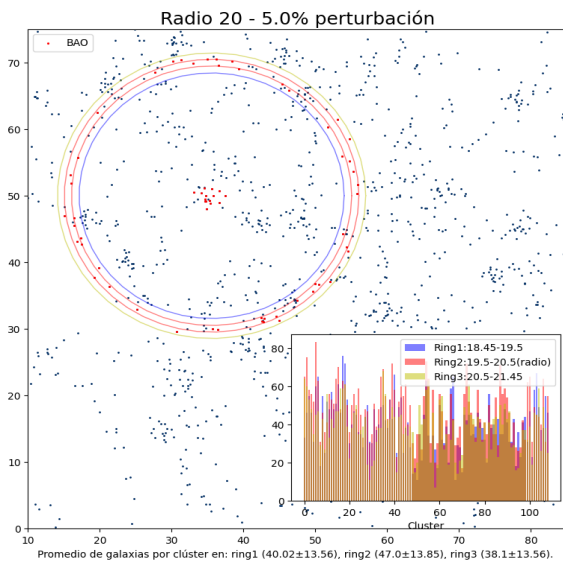
Por último, en la Figura 6.15, se realizó el mismo procedimiento que en la Figura 6.13, pero en este caso se aumentó la cantidad de galaxias y BAOs. En la parte superior izquierda de 6.15(a), podemos apreciar la importancia de la perturbación, ya que a pesar de la pequeña cantidad de BAOs, la distribución comienza a verse más realista. En este caso, se detectaron 109 agrupamientos de los 100 posibles centros de BAO. En 6.15(b) se ilustra la dispersión de las galaxias pertenecientes al clúster (en rojo). Ya no se encuentran solo en el anillo del medio, sino que están más dispersas. Sin embargo, en el histograma podemos observar que la cantidad de galaxias predominantes se encuentra en el ring2 (rojo). En este caso, en promedio hubo (63.04 ± 23.91) galaxias por agrupamiento en el Ring1 (azul), (90.24 ± 24.72) galaxias por agrupamiento en el Ring2 (rojo) y (50.56 ± 23.91) galaxias por agrupamiento en el Ring3 (amarillo). Estos resultados nos llevan a concluir que el radio del BAO se encuentra efectivamente entre 19 y 21 unidades de longitud. Luego, en la Figura 6.15(c), procedemos a refinar el área de los anillos para encontrar el radio característico y observamos en el histograma que hay agrupamientos donde hay más galaxias en el ring1 (azul), y el predominio del ring2 (rojo) es casi imperceptible, esto se puede deber a que las galaxias presentan gran perturbación o el refinamiento fue demasiado grande, sin embargo sigue predominando levemente el anillo ring2. En este caso, en promedio hubo (40.02 ± 13.56) galaxias por agrupamiento en el Ring1 (azul), (47.0 ± 13.85) galaxias por agrupamiento en el Ring2 (rojo) y (38.1 ± 13.56) galaxias por agrupamiento en el Ring3 (amarillo). Estos resultados nos permiten concluir que el radio del BAO se encuentra efectivamente



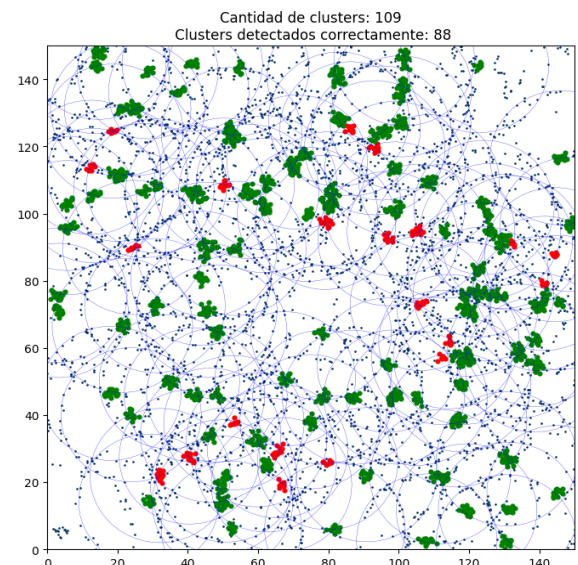
(a) Detección de centros con DBSCAN.



(b) Cálculo de cantidad de galaxias sintéticas en anillos de diferentes radios.



(c) Refinamiento de la figura 6.14(b).



(d) Trazo de BAO a partir de los centros.

Figura 6.15: Reconstrucción del BAO original con algoritmos de agrupamiento. Todos los cuadrantes contienen distribuciones de galaxias sintéticas en un Universo cuadrado de 150 unidades de longitud de lado, con 100 anillos de BAO de 20 unidades de longitud de radio y 10% de perturbación, con 50 galaxias cada uno y 30% de galaxias en el centro respecto a las galaxias sobre el anillo.

entre 19.5 y 20.5 unidades de longitud. Posteriormente, en la figura 6.15(d) se muestra la reconstrucción de los BAOs con radio 15, y vemos que de los 33 agrupamientos detectados, 27 fueron centros de BAO. Se detectó un $\frac{88}{100} = 0.88 = 88\%$ de las esperados con una precisión de $\frac{88}{109} = 0.8073 = 80.73\%$.

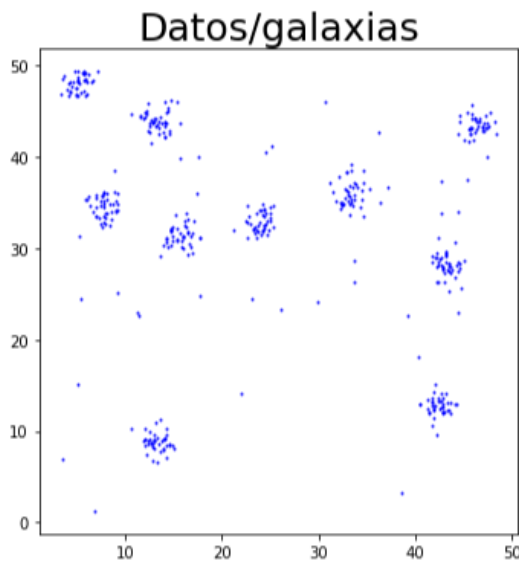


Figura 6.16: 428 galaxias sintéticas distribuidas en 10 circunferencias BAO de 10 unidades de distancia de radio con máximo 5 galaxias sintéticas cada una con perturbación del 5% y 800% de galaxias sintéticas en el centro del BAO respecto a la circunferencia, en una caja cuadrada de 50 unidades de longitud de lado.

En las figuras (6.13) y (6.14) se trabajó con distribuciones poco realistas pero bastante ilustrativas de cómo sería la reconstrucción del BAO original utilizando sólo algoritmos de agrupamiento, una forma alternativa a la función de correlación de dos puntos de calcular el radio, se continuará trabajando este método con mayor cantidad de galaxias sintéticas para determinar si es un método viable como la figura (6.15). Es importante destacar la necesidad de ser cuidadosos al seleccionar los parámetros adecuados y aplicar un riguroso proceso de refinamiento de anillos para detectar con precisión el radio característico del fenómeno BAO.

6.5. BÚSQUEDA DE CENTROS DE BAO COMPARANDO DBSCAN, OPTICS Y HDBSCAN

En esta sección del capítulo analizaremos una distribución de galaxias sintéticas como lo hicimos con DBSCAN pero ahora comparándolo con OPTICS y HDBSCAN debido a que estos algoritmos se consideran mejoras de DBSCAN. Antes de comenzar el análisis se presenta el ejemplo de una distribución de galaxias, mostrado en la figura (6.16) donde se pueden apreciar los agrupamientos a simple vista, es importante mencionar que esta distribución es poco realista ya que considera máximo 5 galaxias por circunferencia BAO y 40 galaxias en el centro, sin embargo se utilizará para un mejor entendimiento de cómo OPTICS y HDBSCAN trabajan, con el fin de posteriormente aplicarlo a distribuciones de galaxias más realistas.

6.5.1. OPTICS

Comenzamos analizando la distribución de la figura (6.16) con el algoritmo OPTICS para comprender su funcionamiento. En la figura (6.17) se muestran 4 subfiguras, en donde se compara la detección de agrupamientos con OPTICS y DBSCAN para esta distribución de galaxias. La gráfica superior es la gráfica de alcanzabilidad, la cual muestra como OPTICS ordeno y escogió los agrupamientos (para mayor detalle vuelva al capítulo 4), en ésta podemos apreciar los puntos que considera como ruido ya que están dispersos y son negros. En la gráfica inferior izquierda se muestran los datos analizados con OPTICS bajo el parámetro `min_samples = 18`, se encontraron los 10 agrupamientos esperados y están del mismo color que en la gráfica de alcanzabilidad, se sugiere ver la figura (6.18) para una mejor comprensión, ya que en esta se une cada agrupamiento con su región en la gráfica de alcanzabilidad. Recordemos que OPTICS no pide como parámetro un radio `eps` a diferencia de DBSCAN, por este motivo comparamos los resultados con DBSCAN con 2 diferentes `eps` (`eps= 1` y `eps= 1.4`). En la gráfica inferior media se muestran los datos analizados con DBSCAN bajo los parámetros `eps= 1` y `min_samples= 18` y podemos notar que de 10 agrupamientos sólo logró detectar 6. Por último, en la gráfica inferior derecha se muestran los datos analizados con DBSCAN bajo los parámetros `eps= 1.4` y `min_samples= 18` y podemos notar que efectivamente encontró los 10 agrupamientos, sin embargo se puede apreciar que no detecta todas las galaxias del agrupamiento. Finalmente, podemos también notar que con el análisis de OPTICS algunos datos ruido los considero dentro de agrupamientos, lo cual sesga los resultados.

Para evitar que OPTICS considere puntos ruido como parte de un agrupamiento, tiene un parámetro opcional llamado `max_eps` el cual a diferencia del parámetro `eps` de DBSCAN no limita al algoritmo a buscar agrupamientos con sólo una densidad (un mínimo de datos en cierto radio), éste considera a todos los agrupamientos con al menos `min_samples` en radios iguales o menores a `max_eps`, por lo que acepta agrupamientos con diferentes densidades. En la figura (6.19) se aprecia el mismo análisis que en la figura (6.17) pero, se establece el `max_eps= 1.4` para OPTICS, evitando así que considere puntos ruido como parte del agrupamiento.

6.5.2. HDBSCAN

A continuación analizamos la distribución de la figura (6.16) con el algoritmo HDBSCAN para comprender su funcionamiento. En la parte izquierda de la figura (6.20) se muestra el árbol de expansión mínima de la distribución de la figura (6.16) el cual se obtiene mediante el siguiente código.

```
1 hd.minimum_spanning_tree_.plot(edge_cmap='viridis', edge_alpha=0.6, node_size=20,
    edge_linewidth=1)
```

Por lo visto en el capítulo 4, sabemos que al ordenar las aristas del árbol de expansión mínima por distancia de alcanzabilidad mutua se obtiene su dendrograma el

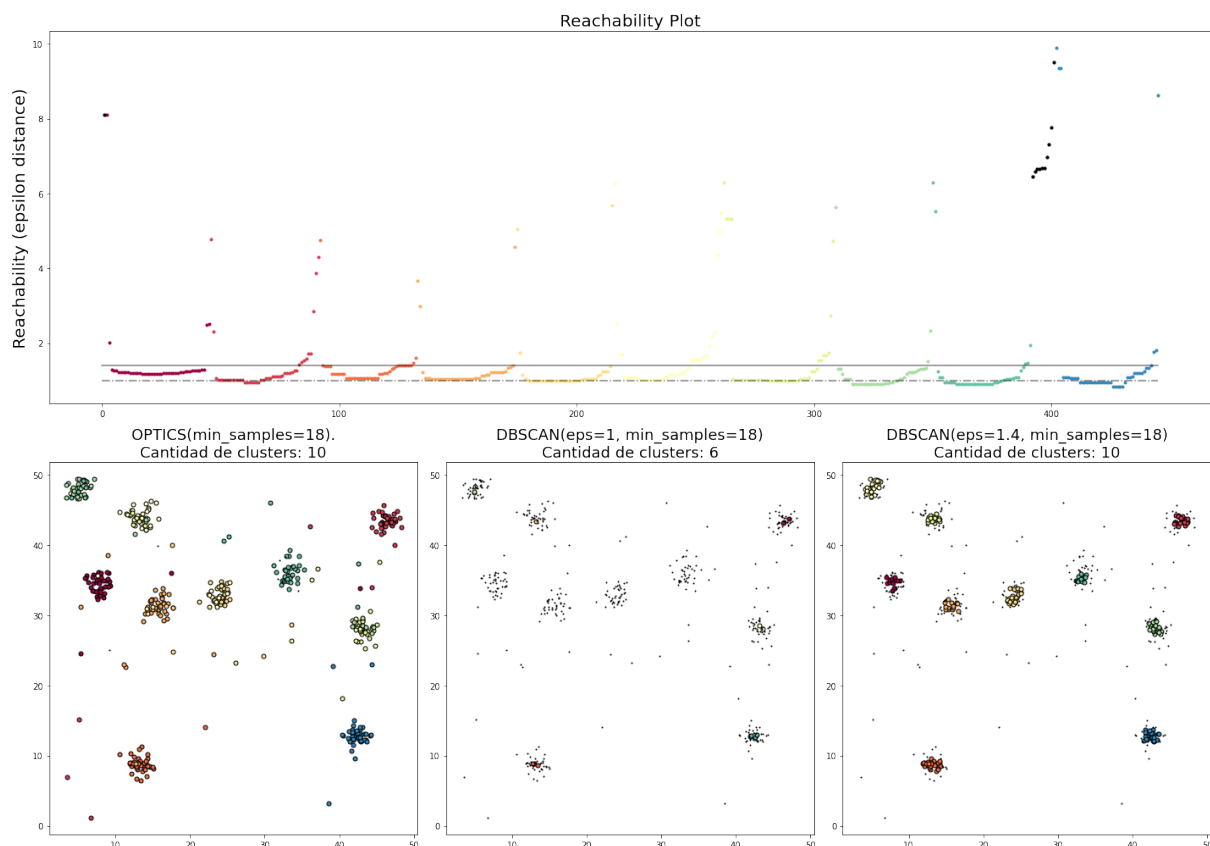


Figura 6.17: Gráfica de alcanzabilidad (*Reachability plot*) de la distribución de la figura (6.16), así como la detección de agrupamientos con los algoritmos OPTICS y DBSCAN con distintos radios ϵ s.

cual se muestra en la figura (6.21) (donde podemos ver los niveles de densidad elegidos debido a que los encierra con un ovalo) y se obtiene mediante el siguiente código.

```
1 hd.condensed_tree_.plot(select_clusters=True, selection_palette=sns.color_palette('deep',8))
```

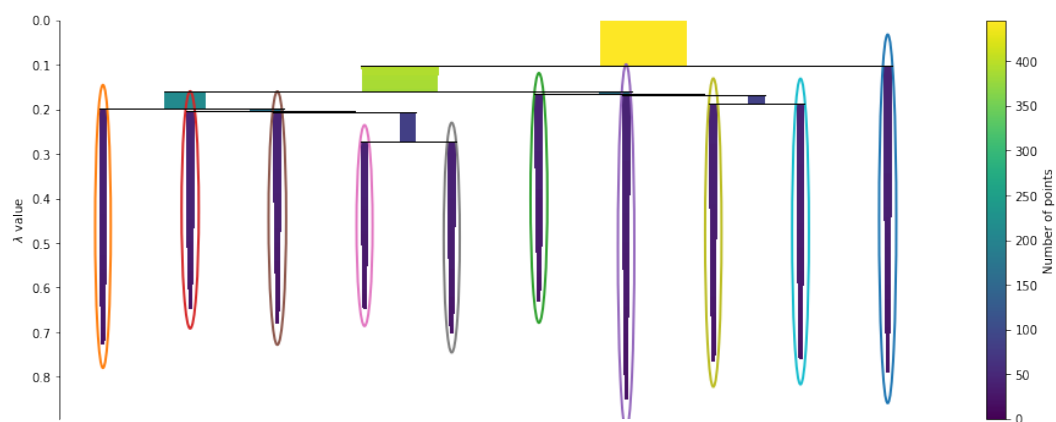


Figura 6.21: Dendrograma simplificado de la distribución de la figura (6.16).

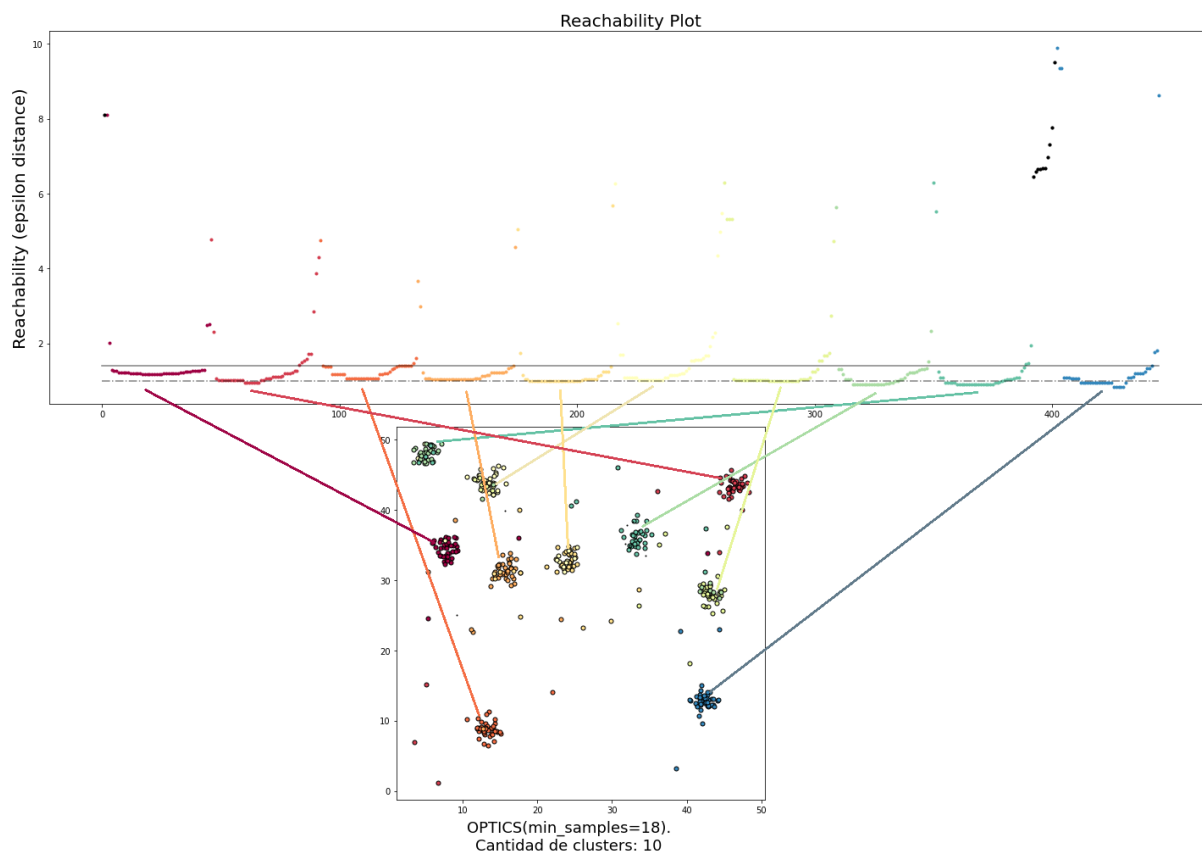


Figura 6.18: Análisis de la detección de agrupamientos con el algoritmo OPTICS y la gráfica de alcanzabilidad (Reachability plot) de la distribución de la figura (6.16).

El dendrograma de la figura (6.21) se ve bastante pequeño, esto se debe a que está simplificado, es decir solo se muestran las ramificaciones hasta las elegidas, las cuales se encierran dentro de óvalos, si contamos la cantidad de ramificaciones encerradas notaremos que son diez, es decir, detectó diez agrupamientos, que eran justo los esperados. En este problema trabajamos con pocos agrupamientos sin embargo si aumentamos la cantidad será complicado visualizar su dendrograma. En caso de querer visualizar todas las ramificaciones, como se muestra en la derecha de 6.20, se puede obtener mediante el siguiente código, sin embargo no es recomendable ya que conforme aumentemos la cantidad de agrupamientos por detectar, más complicado será su dendrograma.

```
1 hd.single_linkage_tree_.plot(cmap='viridis', colorbar= True)
```

Finalmente, en la figura (6.22) se muestran los agrupamientos detectados por HDBSCAN, podemos apreciar que considera prácticamente todas las galaxias de cada agrupamiento, no es afectado por el ruido y es un avance realmente grande debido a que sólo otorgamos el parámetro `min_cluster_size`.

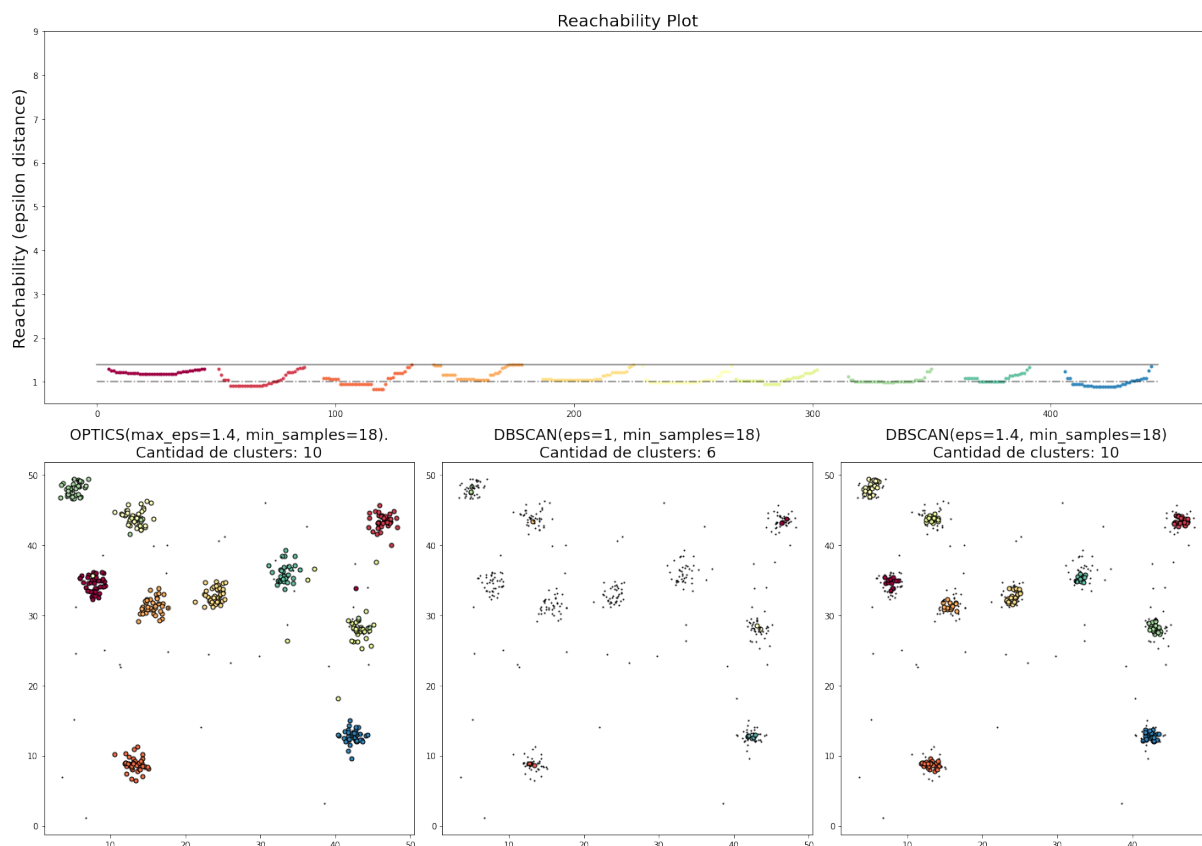


Figura 6.19: Gráfica de alcanzabilidad (Reachability plot) de la distribución de la figura (6.16), así como la detección de agrupamientos con los algoritmos OPTICS (agregando el parámetro max_eps) y DBSCAN con distintos radios eps .

6.5.3. Análisis con distintas distribuciones de galaxias

Ya que comprendemos el funcionamiento de tanto de OPTICS como HDBSCAN, procederemos a analizar una distribución de galaxias sintéticas.

En la figura (6.23) se muestra una comparación del análisis tanto con DBSCAN como OPTICS y HDBSCAN de la distribución de 30 BAOs de la figura (6.5). Ésta es una distribución muy poco realista sin embargo nos puede dar un panorama del funcionamiento de estos. Como ya lo habíamos analizado en la sección de análisis con DBSCAN (gráfica inferior izquierda) de este capítulo, vemos que con DBSCAN se detectan 29 agrupamientos de 30, y detecta exclusivamente los centroides como podemos corroborarlo al observar la reconstrucción de BAOs. Luego para OPTICS (gráfica inferior media) vemos que detecta 31 agrupamientos de 30, esta imagen es bastante interesante pues podemos notar que a pesar de que algunos centroides en el centro no son detectados, sí logra detectar agrupamientos completos, es decir, centroide y circunferencia incluso a pesar de que algunos están bastante juntos o se traslapan, lo cual es bastante prometedor. Aunado a esto la gráfica de alcanzabilidad (gráfica superior) nos puede dar una buena idea de cómo detectar estos agrupamientos de una

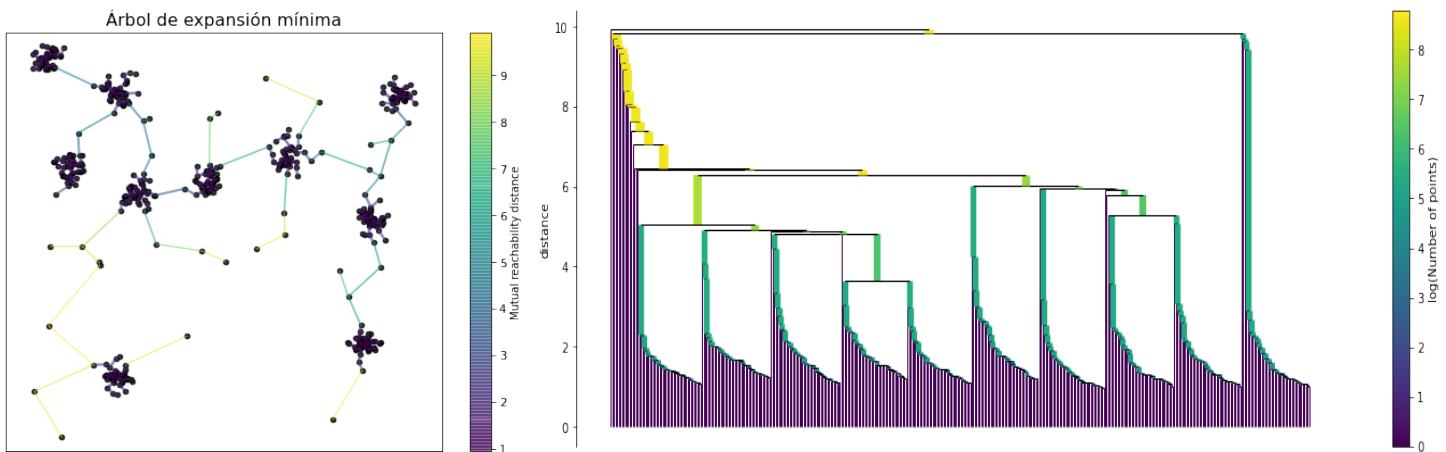


Figura 6.20: Izquierda: *Árbol de expansión mínima de la distribución* y Derecha: *Dendrograma de la distribución*. Ambos correspondientes a la figura (6.16).

manera mas eficiente, es importante mencionar que en este caso se decidió no dar un parámetro de `max_eps`. Por último, el análisis con HDBSCAN (gráfica inferior derecha), detecta 28 agrupamientos de 30 esperados, vemos en la reconstrucción de BAOs que aunque no es muy exacto, si detecta correctamente gran parte de los BAO, algunos un poco trasladados, esto debido a que también está considerando algunas galaxias en las circunferencias como parte del agrupamiento, aquí me gustaría recalcar que es muy interesante las detecciones que realizó con el único parámetro de `min_cluster_size` así como las detecciones de OPTICS con sólo el parámetros de `min_samples`. A pesar de que el análisis aquí realizado no es el más exhaustivo para OPTICS y HDBSCAN podemos determinar que son algoritmos bastante prometedores para el fin de detectar agrupamientos de galaxias, realmente piden muy pocos parámetros y en futuros trabajos se seguirá trabajando con éstos.

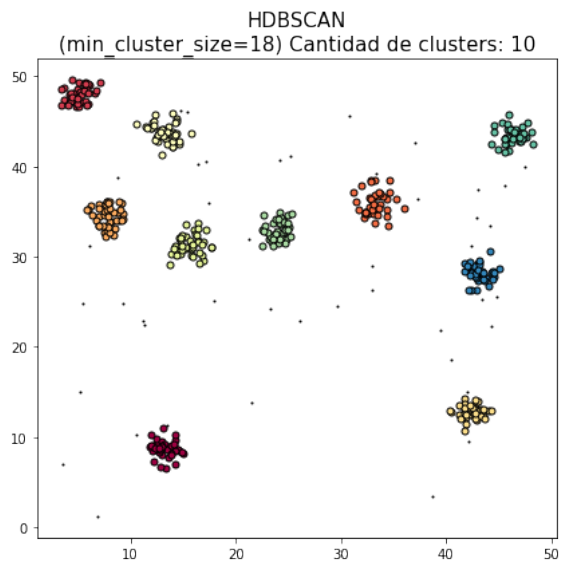


Figura 6.22: Análisis de los centroides detectados mediante el algoritmo de agrupamiento espacial basado en densidad (DBSCAN) de las distribuciones de galaxias sintéticas representadas en la figura (6.9).

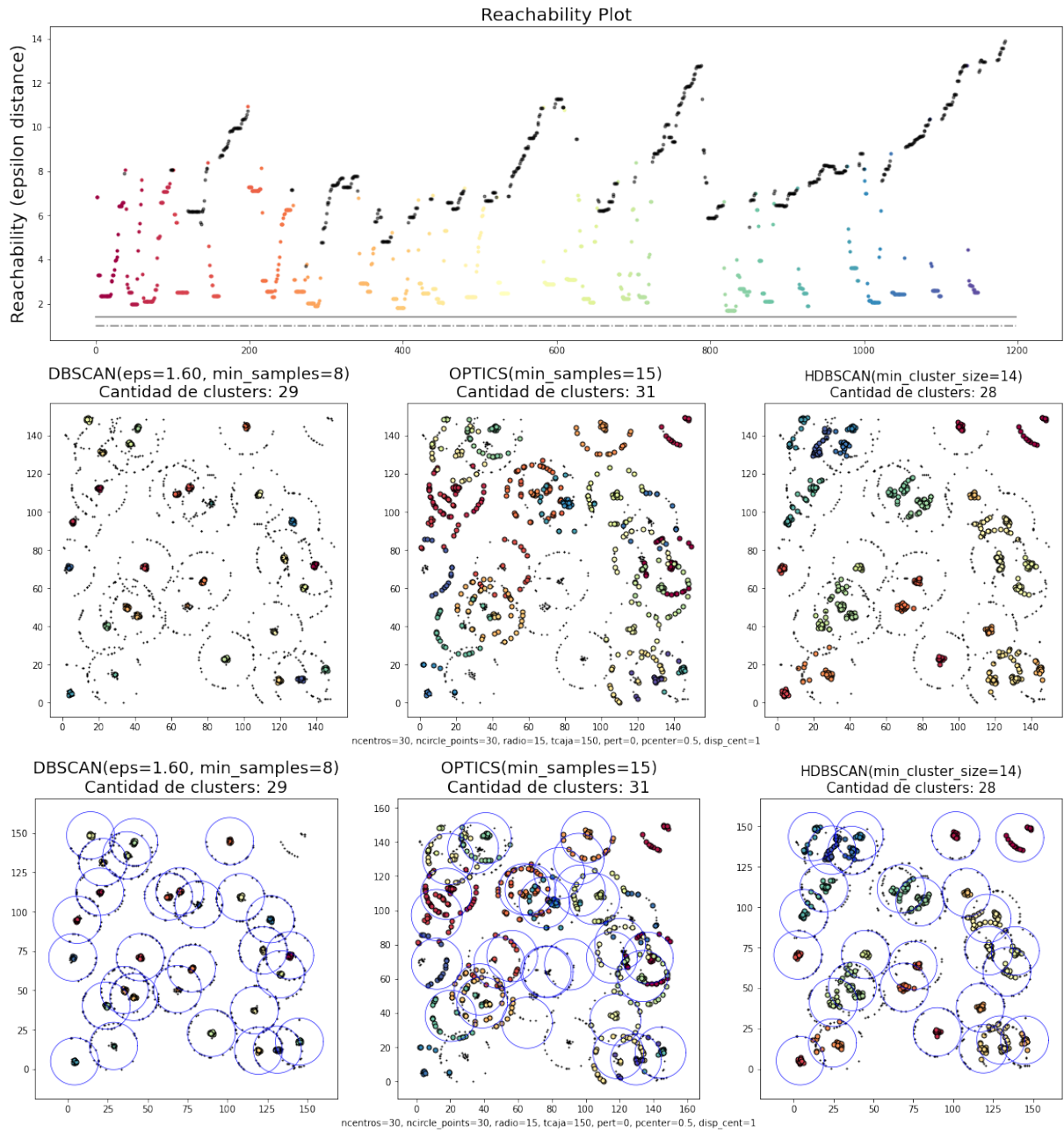


Figura 6.23: Gráfica de alcanzabilidad (Reachability plot) de la distribución de la figura (6.16) (Arriba), así como la detección de agrupamientos con los algoritmos OPTICS y DBSCAN con distintos radios eps (En medio) y su reconstrucción del BAO original (Abajo).

Capítulo 7

Conclusiones y Perspectivas

A lo largo de este trabajo se estudio la escala característica de BAO mediante el uso de diferentes estimadores de la función de correlación de dos puntos. A partir de los resultados obtenidos se exploró la posibilidad de que dichos patrones puedan ser identificados con técnicas de machine learning, en particular con el algoritmo de agrupamiento DBSCAN. A continuación se presentan las principales conclusiones de este estudio.

Al realizar la simulación simplificada de galaxias, notamos que que conforme aumenta el número de anillos de BAO y/o se reduce la cantidad de galaxias sobre cada anillo, este “oculta” visualmente la escala característica, lo que implica que solo puede recuperarse estadísticamente y para ello utilizamos los diferentes estimadores de la función de correlación de dos puntos.

Comenzamos evaluando estos estimadores mediante la utilización de dos distribuciones pseudoaleatorias, con el objetivo de determinar cuál de ellos se adecua mejor a nuestro modelo de Universo plano cuadrado. Tras el análisis, se concluyó que el estimador **Hamilton** muestra un rendimiento sobresaliente en ambas distribuciones pseudoaleatorias en nuestro espacio de estudio. Esta elección se basa en su capacidad de presentar niveles de ruido mínimos y oscila alrededor de cero.

A continuación analizamos diversas distribuciones de galaxias sintéticas, con los distintos estimadores, y se contempló que debido a que para grandes distancias hay pocos conteos, se presenta mayor ruido con todos los estimadores. Luego, se realizó un estudio exhaustivo del pico de BAO, analizando cómo varía en la función de correlación de dos puntos en relación con el radio y la densidad de galaxias en los anillos y centros de BAO. Se observó que, a medida que aumenta la cantidad de galaxias por circunferencia de BAO, mayor será la amplitud de los picos del radio y diámetro y mayor definido será el pico del diámetro. Por otro lado, conforme aumenta el número de anillos de BAO, el pico de disminuirá su amplitud. En adición, podemos ver que entre menor es el porcentaje de galaxias en el centro, menos ancho y alto es el pico del radio mientras que menos ancho y más alto es el pico del diámetro.

Asimismo, la cantidad de perturbación que presenta una distribución de galaxias respecto al BAO original es muy relevante debido a que ésta modifica que tan aleatoria

o no parece ser una distribución. Se concluyó que entre mayor es la perturbación más tenue se muestra el pico de BAO en la función de correlación de dos puntos, es decir, menos predominante es el pico (*se achata*).

Posteriormente, se cotejó la función de correlación de dos puntos obtenida con el estimador **Peebles-Hauser** comparando la diferencia de utilizar una función geométrica para el conteo de pares aleatorios en lugar de hacerlo por el método de integración Monte Carlo, concluimos que para escalas lejanas de cero se obtienen resultados bastante similares por ambos métodos. Por lo que utilizar la función geométrica resulta mucho más eficiente, debido al menor costo computacional y a que es el resultado de tender Monte Carlo a infinito, por lo cual es más precisa.

Finalmente, proponemos como mejor estimador a **Hewett** debido a que, después de **Peebles-Hauser**, fue el segundo que más marcado muestra el pico BAO, aunado a que es junto con **Hamilton** de los que menos ruido presenta para grandes distancias.

Posteriormente comenzamos a trabajar con el algoritmo de agrupamiento DBSCAN con el fin de detectar los centros de los BAO, para todas las distribuciones, a pesar de que consideramos una mayor densidad de datos aleatorios, se logró distinguir entre datos aleatorios y datos de galaxias con DBSCAN. De los resultados de DBSCAN concluimos lo siguiente. Conforme aumenta la cantidad de galaxias sintéticas, aumentará la posibilidad de que el algoritmo detecte como agrupamientos al traslape de galaxias en distintas circunferencias de BAO o que considere varios centros en un mismo agrupamiento. Además, a medida que aumenta la cantidad de galaxias presentes sobre los BAOs, se facilita la correcta detección de los centroides por parte del algoritmo DBSCAN, a pesar de que la cantidad total de galaxias en el espacio aumenta debido al incremento de densidad de galaxias en los centroides. Por otro lado, al aumentar la perturbación se dificulta la detección de centroides por el ojo humano, sin embargo DBSCAN detectó la mayoría de las circunferencias de BAO correctamente. Es importante mencionar que los parámetros tomados son manualmente por lo que es posible que no sean los óptimos, será importante analizar con mayor detalle sus parámetros para obtener mejores resultados.

Podemos concluir que los algoritmos de agrupamiento, en particular con DBSCAN, son bastante prometedores para la detección de centros de BAO, debido a que se detectaron gran parte de éstos, y se logró detectar el radio del BAO original con ayuda de la función de correlación de dos puntos, logrando caracterizar distribuciones sintéticas de galaxias; así mismo, se localizó el radio del BAO original sin la función de correlación utilizando únicamente el algoritmo DBSCAN. El hecho de detectar la escala característica sin utilizar la función de correlación de dos puntos implica evitar el cálculo de bastantes distancia, por lo que el tiempo de computo es menor. Además, resulta interesante considerar que DBSCAN busca agrupamientos con densidades iguales por lo que incluso podría funcionar para clasificación utilizando diferentes parámetros de densidades. Así mismo, no es un algoritmo perfecto, es bastante sensible a los parámetros que le otorgamos manualmente: `min_pts` y `eps`; por lo que para obtener mejores resultados se propone trabajar con mejoras de DBSCAN que sean menos sensibles a estos parámetros, es posible que considere traslape de galaxias de

diferentes circunferencias de BAO como centroides, o que considere a dos centroides cercanos como uno sólo.

Finalmente se realizó un pequeño análisis con OPTICS y HDBSCAN pudiendo determinar que son algoritmos bastante prometedores para detectar agrupamientos de galaxias y se consideraran como propuestas para futuros trabajos, así como DRL-DBSCAN.

Bibliografía

[Bol, 2015] (2015). Introduction: The bolshoi simulation.

[Ade et al., 2014] Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Battaner, E., Benabed, K., Benoît, A., Benoit-Lévy, A., Bernard, J.-P., Bersanelli, M., Bielewicz, P., Bobin, J., Bock, J. J., Bonaldi, A., Bond, J. R., Borrill, J., Bouchet, F. R., Bridges, M., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cappellini, B., Cardoso, J.-F., Catalano, A., Challinor, A., Chamballu, A., Chary, R.-R., Chen, X., Chiang, H. C., Chiang, L.-Y., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Couchot, F., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Delouis, J.-M., Désert, F.-X., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Doré, O., Douspis, M., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Gaier, T. C., Galeotta, S., Galli, S., Ganga, K., Giard, M., Giardino, G., Giraud-Héraud, Y., Gjerløw, E., González-Nuevo, J., Górski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Haissinski, J., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D., Henrot-Versillé, S., Hernández-Monteagudo, C., Herranz, D., Hildebrandt, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hou, Z., Hovest, W., Huffenberger, K. M., Jaffe, A. H., Jaffe, T. R., Jewell, J., Jones, W. C., Juvela, M., Keihänen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lähteenmäki, A., Lamarre, J.-M., Lasenby, A., Lattanzi, M., Laureijs, R. J., Lawrence, C. R., Leach, S., Leahy, J. P., Leonardi, R., León-Tavares, J., Lesgourgues, J., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vørnle, M., López-Cañiego, M., Lubin, P. M., Macías-Pérez, J. F., Maffei, B., Maino, D., Mandolesi, N., Maris, M., Marshall, D. J., Martin, P. G., Martínez-González, E., Masi, S., Massardi, M., Matarrese, S., Matthai, F., Mazzotta, P., Meinhold, P. R., Melchiorri, A., Melin, J.-B., Mendes, L., Menegoni, E., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M.-A., Moneti, A., Montier, L., Morgante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Nørgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., O'Dwyer, I. J., Osborne, S., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge, B., Pasian, F., Patanchon, G., Pearson, D., Pearson, T. J., Peiris, H. V., Perdureau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczyński, S., Platania, P.,

Pointecouteau, E., Polenta, G., Ponthieu, N., Popa, L., Poutanen, T., Pratt, G. W., Prézeau, G., Prunet, S., Puget, J.-L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Ricciardi, S., Riller, T., Ristorcelli, I., Rocha, G., Rosset, C., Roudier, G., Rowan-Robinson, M., Rubiño-Martín, J. A., Rusholme, B., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Shellard, E. P. S., Spencer, L. D., Starck, J.-L., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sureau, F., Sutton, D., Suur-Uski, A.-S., Sygnet, J.-F., Tauber, J. A., Tavagnacco, D., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Tucci, M., Tuovinen, J., Türler, M., Umama, G., Valenziano, L., Valiviita, J., Tent, B. V., Vielva, P., Villa, F., Vittorio, N., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. (2014). iplanck/i2013 results. XVI. cosmological parameters. *Astronomy & Astrophysics*, 571:A16.

[Aghanim et al., 2020] Aghanim, N., Akrami, Y., Arroja, F., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J.-P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J.-F., Carron, J., Casaponsa, B., Challinor, A., Chiang, H. C., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J.-M., Désert, F.-X., Valentino, E. D., Dickinson, C., Diego, J. M., Donzelli, S., Doré, O., Douspis, M., Ducout, A., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Falgarone, E., Fantaye, Y., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Helou, G., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J.-M., Langer, M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Jeune, M. L., Leahy, J. P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniago, M., Lubin, P. M., Ma, Y.-Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meerburg, P. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M.-A., Molinari, D., Moneti, A., Montier, L., Morgante, G., Moss, A., Mottet, S., Münchmeyer, M., Natoli, P., Nørgaard-Nielsen, H. U., Oxborrow, C. A., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Pearson, T. J., Peel, M., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J.-L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renault, C., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Shiraishi, M., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A.-S., Tauber, J. A., Tavagnacco, D., Tenti, M., Terenzi, L., Toffolatti, L., Tomasi, M., Trombetti, T., Valiviita, J., Tent, B. V., Vibert, L., Vielva, P.,

- Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. (2020). iplanck/i2018 results. *Astronomy & Astrophysics*, 641:A1.
- [Alam et al., 2021] Alam, S., Aubert, M., Avila, S., Balland, C., Bautista, J. E., Bershad, M. A., Bizyaev, D., Blanton, M. R., Bolton, A. S., Bovy, J., Brinkmann, J., Brownstein, J. R., Burtin, E., Chabanier, S., Chapman, M. J., Choi, P. D., Chuang, C.-H., Comparat, J., Cousinou, M.-C., Cuceu, A., Dawson, K. S., de la Torre, S., de Mattia, A., de Sainte Agathe, V., du Mas des Bourboux, H., Escoffier, S., Etourneau, T., Farr, J., Font-Ribera, A., Frinchaboy, P. M., Fromenteau, S., Gil-Marín, H., Goff, J.-M. L., Gonzalez-Morales, A. X., Gonzalez-Perez, V., Grabowski, K., Guy, J., Hawken, A. J., Hou, J., Kong, H., Parker, J., Klaene, M., Kneib, J.-P., Lin, S., Long, D., Lyke, B. W., de la Macorra, A., Martini, P., Masters, K., Mohammad, F. G., Moon, J., Mueller, E.-M., Muñoz-Gutiérrez, A., Myers, A. D., Nadathur, S., Neveux, R., Newman, J. A., Noterdaeme, P., Oravetz, A., Oravetz, D., Palanque-Delabrouille, N., Pan, K., Paviot, R., Percival, W. J., Pérez-Ràfols, I., Petitjean, P., Pieri, M. M., Prakash, A., Raichoor, A., Ravoux, C., Rezaie, M., Rich, J., Ross, A. J., Rossi, G., Ruggeri, R., Ruhlmann-Kleider, V., Sánchez, A. G., Sánchez, F. J., Sánchez-Gallego, J. R., Sayres, C., Schneider, D. P., Seo, H.-J., Shafieloo, A., Slosar, A., Smith, A., Stermer, J., Tamone, A., Tinker, J. L., Tojeiro, R., Vargas-Magaña, M., Variu, A., Wang, Y., Weaver, B. A., Weijmans, A.-M., Yèche, C., Zarrouk, P., Zhao, C., Zhao, G.-B., and Zheng, Z. (2021). Completed SDSS-IV extended baryon oscillation spectroscopic survey: Cosmological implications from two decades of spectroscopic surveys at the apache point observatory. *Physical Review D*, 103(8).
- [Alonso, 2013] Alonso, D. (2013). Cute solutions for two-point correlation functions from large cosmological datasets.
- [Alonso, 2020] Alonso, V. R. F. (2020). *Función de correlación de materia de dos puntos en el formalismo combinado de perturbaciones-halo*. [tesis para optar al grado de magister en astronomía, facultad de ciencias], Universidad Nacional de Colombia. Repositorio de la Universidad Nacional de Colombia. <https://repositorio.unal.edu.co/bitstream/handle/unal/78331/93410836.2019.pdf?sequence=4&isAllowed=y>.
- [Angulo, 2011] Angulo, R. W. S. D. M. (2011). The millennium-xxl project: Simulating the galaxy population in dark energy universes.
- [Ankerst et al., 1999] Ankerst, M., Breunig, M., Kröger, P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. volume 28, pages 49–60.
- [Anónimo, 2014] Anónimo (2014). La medida más precisa de las galaxias lejanas.
- [Anónimo, 2020] Anónimo (2020). No need to mind the gap: Astrophysicists fill in 11 billion years of our universe’s expansion history.
- [Anónimo, 2021] Anónimo (2021). The dark energy spectroscopic instrument (desi).
- [Anónimo, 2022a] Anónimo (2022a). Ley de la inversa del cuadrado.

- [Anónimo, 2022b] Anónimo (2022b). What is clustering in machine learning: Types and methods.
- [Armendariz-Picon and Neelakanta, 2014] Armendariz-Picon, C. and Neelakanta, J. T. (2014). How cold is cold dark matter? *Journal of Cosmology and Astroparticle Physics*, 2014(03):049–049.
- [Ashley, 2020] Ashley, B. (2020). Nasa’s roman space telescope to uncover echoes of the universe’s creation.
- [Bassett and Hlozek, 2009] Bassett, B. A. and Hlozek, R. (2009). Baryon acoustic oscillations.
- [Baumann, 2015] Baumann, D. (2015). University of Cambridge.
- [Behroozi et al., 2012] Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. (2012). THE ROCKSTAR PHASE-SPACE TEMPORAL HALO FINDER AND THE VELOCITY OFFSETS OF CLUSTER CORES. *The Astrophysical Journal*, 762(2):109.
- [Berba, 2020] Berba, P. (2020). Understanding hdbscan and density-based clustering.
- [Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers Geosciences*, 10(2):191–203.
- [Biermann, 2019] Biermann, M. (2019). *A Detailed Look at Estimators for the Two-Point Correlation Function, their Variance and a Verification of the TreeCorr-Algorithm with LoTSS Data*. [master thesis], Bielefeld University. Repository of Bielefeld University. https://www2.physik.uni-bielefeld.de/fileadmin/user_upload/theory_e6/Master_Theses/MA_MarianBiermann.pdf.
- [Bode et al., 2001] Bode, P., Ostriker, J. P., and Turok, N. (2001). Halo formation in warm dark matter models. *The Astrophysical Journal*, 556(1):93–107.
- [Boylan-Kolchin et al., 2009] Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A., and Lemson, G. (2009). Resolving cosmic structure formation with the millennium-II simulation. *Monthly Notices of the Royal Astronomical Society*, 398(3):1150–1164.
- [Burdick et al., 2005] Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., and Yiu, T. (2005). Mafia: A maximal frequent itemset algorithm. *IEEE Trans. on Knowl. and Data Eng.*, 17(11):1490–1504.
- [Campello, 2013] Campello, R., M. D. S. J. (2013). Density-based clustering based on hierarchical density estimates. pages 160–172.
- [Carlson et al., 1992] Carlson, E. D., Machacek, M. E., and Hall, L. J. (1992). Self-interacting Dark Matter. , 398:43.

- [Cen, 2001] Cen, R. (2001). Decaying cold dark matter model and small-scale power. *The Astrophysical Journal*, 546(2):L77–L80.
- [Chacón Lavanderos, 2018] Chacón Lavanderos, J. (2018). *Modelos de Materia Oscura: Una Perspectiva Numérica*. [tesis que para obtener el título de licenciatura en física y matemáticas], Instituto Politécnico Nacional. Repositorio del Dr. Tonatiuh Matos del Cinvestav. http://pelusa.fis.cinvestav.mx/tmatos/CV/3_RecursosH/Lic/Jazhiel_ESFM.pdf.
- [Coble et al., 2018] Coble, K., Conlon, M., and Bailey, J. (2018). Investigating undergraduate students’ ideas about the curvature of the universe. *Physical Review Physics Education Research*, 14.
- [Colin et al., 2000] Colin, P., Avila-Reese, V., and Valenzuela, O. (2000). Substructure and halo density profiles in a warm dark matter cosmology. *The Astrophysical Journal*, 542(2):622–630.
- [Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.
- [DESI_Collaboration et al., 2016] DESI_Collaboration, Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Prieto, C. A., Annis, J., Bailey, S., Balland, C., Ballester, O., Baltay, C., Beaufore, L., Bebek, C., Beers, T. C., Bell, E. F., Bernal, J. L., Besuner, R., Beutler, F., Blake, C., Bleuler, H., Blomqvist, M., Blum, R., Bolton, A. S., Briceno, C., Brooks, D., Brownstein, J. R., Buckley-Geer, E., Burden, A., Burtin, E., Busca, N. G., Cahn, R. N., Cai, Y.-C., Cardiel-Sas, L., Carlberg, R. G., Carton, P.-H., Casas, R., Castander, F. J., Cervantes-Cota, J. L., Claybaugh, T. M., Close, M., Coker, C. T., Cole, S., Comparat, J., Cooper, A. P., Cousinou, M. C., Crocce, M., Cuby, J.-G., Cunningham, D. P., Davis, T. M., Dawson, K. S., de la Macorra, A., Vicente, J. D., Delubac, T., Derwent, M., Dey, A., Dhungana, G., Ding, Z., Doel, P., Duan, Y. T., Ealet, A., Edelstein, J., Eftekhari, S., Eisenstein, D. J., Elliott, A., Escoffier, S., Evatt, M., Fagrellius, P., Fan, X., Fanning, K., Farahi, A., Farihi, J., Favole, G., Feng, Y., Fernandez, E., Findlay, J. R., Finkbeiner, D. P., Fitzpatrick, M. J., Flaugher, B., Flender, S., Font-Ribera, A., Forero-Romero, J. E., Fosalba, P., Frenk, C. S., Fumagalli, M., Gaensicke, B. T., Gallo, G., Garcia-Bellido, J., Gaztanaga, E., Fusillo, N. P. G., Gerard, T., Gershkovich, I., Giannantonio, T., Gillet, D., de Rivera, G. G., Gonzalez-Perez, V., Gott, S., Graur, O., Gutierrez, G., Guy, J., Habib, S., Heetderks, H., Heetderks, I., Heitmann, K., Hellwing, W. A., Herrera, D. A., Ho, S., Holland, S., Honscheid, K., Huff, E., Hutchinson, T. A., Huterer, D., Hwang, H. S., Laguna, J. M. I., Ishikawa, Y., Jacobs, D., Jeffrey, N., Jelinsky, P., Jennings, E., Jiang, L., Jimenez, J., Johnson, J., Joyce, R., Jullo, E., Juneau, S., Kama, S., Karcher, A., Karkar, S., Kehoe, R., Kennamer, N., Kent, S., Kilbinger, M., Kim, A. G., Kirkby, D., Kisner, T., Kitanidis, E., Kneib, J.-P., Koposov, S., Kovacs, E., Koyama, K., Kremin, A., Kron, R., Kronig, L., Kueter-Young, A., Lacey, C. G., Lafever, R., Lahav, O., Lambert, A., Lampton, M., Landriau, M.,

Lang, D., Lauer, T. R., Goff, J.-M. L., Guillou, L. L., Suu, A. L. V., Lee, J. H., Lee, S.-J., Leitner, D., Lesser, M., Levi, M. E., L’Huillier, B., Li, B., Liang, M., Lin, H., Linder, E., Loebman, S. R., Lukić, Z., Ma, J., MacCrann, N., Magneville, C., Makarem, L., Manera, M., Manser, C. J., Marshall, R., Martini, P., Massey, R., Matheson, T., McCauley, J., McDonald, P., McGreer, I. D., Meisner, A., Metcalfe, N., Miller, T. N., Miquel, R., Moustakas, J., Myers, A., Naik, M., Newman, J. A., Nichol, R. C., Nicola, A., da Costa, L. N., Nie, J., Niz, G., Norberg, P., Nord, B., Norman, D., Nugent, P., O’Brien, T., Oh, M., Olsen, K. A. G., Padilla, C., Padmanabhan, H., Padmanabhan, N., Palanque-Delabrouille, N., Palmese, A., Pappalardo, D., Pâris, I., Park, C., Patej, A., Peacock, J. A., Peiris, H. V., Peng, X., Percival, W. J., Perruchot, S., Pieri, M. M., Pogge, R., Pollack, J. E., Poppett, C., Prada, F., Prakash, A., Probst, R. G., Rabinowitz, D., Raichoor, A., Ree, C. H., Refregier, A., Regal, X., Reid, B., Reil, K., Rezaie, M., Rockosi, C. M., Roe, N., Ronayette, S., Roodman, A., Ross, A. J., Ross, N. P., Rossi, G., Rozo, E., Ruhlmann-Kleider, V., Rykoff, E. S., Sabiu, C., Samushia, L., Sanchez, E., Sanchez, J., Schlegel, D. J., Schneider, M., Schubnell, M., Secroun, A., Seljak, U., Seo, H.-J., Serrano, S., Shafieloo, A., Shan, H., Sharples, R., Sholl, M. J., Shourt, W. V., Silber, J. H., Silva, D. R., Sirk, M. M., Slosar, A., Smith, A., Smoot, G. F., Som, D., Song, Y.-S., Sprayberry, D., Staten, R., Stefanik, A., Tarle, G., Tie, S. S., Tinker, J. L., Tojeiro, R., Valdes, F., Valenzuela, O., Valluri, M., Vargas-Magana, M., Verde, L., Walker, A. R., Wang, J., Wang, Y., Weaver, B. A., Weaverdyck, C., Wechsler, R. H., Weinberg, D. H., White, M., Yang, Q., Yeche, C., Zhang, T., Zhao, G.-B., Zheng, Y., Zhou, X., Zhou, Z., Zhu, Y., Zou, H., and Zu, Y. (2016). The desi experiment part i: Science, targeting, and survey design.

[DiFrancesco et al., 2020] DiFrancesco, P.-M., Bonneau, D., and Hutchinson, D. J. (2020). The implications of m3c2 projection diameter on 3d semi-automated rockfall extraction from sequential terrestrial laser scanning point clouds. *Remote Sensing*, 12(11).

[Díaz, 2013] Díaz, J. (2013). ¿qué es la radiación de fondo de microondas?

[Eisenstein et al., 2005] Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castander, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., Frieman, J. A., Glazebrook, K., Gunn, J. E., Hendry, J. S., Hennessy, G., Ivezić, Z., Kent, S., Knapp, G. R., Lin, H., Loh, Y.-S., Lupton, R. H., Margon, B., McKay, T. A., Meiksin, A., Munn, J. A., Pope, A., Richmond, M. W., Schlegel, D., Schneider, D. P., Shimasaku, K., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Tucker, D. L., Yanny, B., and York, D. G. (2005). Detection of the baryon acoustic peak in the large-scale correlation function of SDSS luminous red galaxies. *The Astrophysical Journal*, 633(2):560–574.

[Escamilla Torres, 2018] Escamilla Torres, L. A. (2018). *Reconstrucción de la energía oscura con base en observaciones cosmológicas*. [tesis de maestría], CINVESTAV. Repositorio del

Dr. José Alberto Vázquez. https://www.fis.unam.mx/~javazquez/files/Thesis/LAdrian_thesis.pdf.

- [Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, page 226–231. AAAI Press.
- [Fan et al., 2021] Fan, T., Guo, N., and Ren, Y. (2021). Consumer clusters detection with geo-tagged social network data using dbscan algorithm: a case study of the pearl river delta in china. *GeoJournal*, 86(1):317–337.
- [Fan and Xu, 2019] Fan, Z. and Xu, X. (2019). Application and visualization of typical clustering algorithms in seismic data analysis. In Shakshuki, E. M. and Yasar, A., editors, *The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops, April 29 - May 2, 2019, Leuven, Belgium*, volume 151 of *Procedia Computer Science*, pages 171–178. Elsevier.
- [Forster and Murphy, 2009] Forster, A. and Murphy, A. L. (2009). Clique: Role-free clustering with q-learning for wireless sensor networks. In *2009 29th IEEE International Conference on Distributed Computing Systems*, pages 441–449.
- [Francis et al., 2011] Francis, Z., Villagrasa, C., and Clairand, I. (2011). Simulation of dna damage clustering after proton irradiation using an adapted dbscan algorithm. *Computer Methods and Programs in Biomedicine*, 101(3):265–270.
- [Freedman et al., 2001] Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Robert C. Kennicutt, J., Ford, H. C., Graham, J. A., Huchra, J. P., Hughes, S. M. G., Illingworth, G. D., Macri, L. M., and Stetson, P. B. (2001). Final results from the hubble space telescope/ikey project to measure the hubble constant. *The Astrophysical Journal*, 553(1):47–72.
- [Garcia, 2019] Garcia, F. (2019). Espectroscopía y efecto doppler.
- [García-Lambas, 1984] García-Lambas, D. (1984). Función de correlación de cúmulos de galaxias. *Boletín de la Asociación Argentina de Astronomía*, (28).
- [Garreta and Moncecchi, 2013] Garreta, R. and Moncecchi, G. (2013). *Learning Scikit-Learn: Machine Learning in Python*. Community experience distilled. Packt Publishing.
- [Géron, 2017] Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn and Tensor-Flow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
- [González González, 2013] González González, J. R. (2013). *Funciones de correlación para una fuente coherente con dos y tres detectores*. [tesis que para obtener el título de físico], Facultad de Ciencias, UNAM. <http://132.248.9.195/ptd2014/enero/0707342/0707342.pdf>.

- [Goodman, 2000] Goodman, J. (2000). Repulsive dark matter. *New Astronomy*, 5(2):103–107.
- [Guan et al., 2018] Guan, C., Kevin Kam Fung, Y., and Yue, Y. (2018). Towards a personalized item recommendation approach in social tagging systems using intuitionistic fuzzy dbSCAN. In *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 01, pages 361–364.
- [Guha et al., 1998] Guha, S., Rastogi, R., and Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98*, page 73–84, New York, NY, USA. Association for Computing Machinery.
- [Guha et al., 2000] Guha, S., Rastogi, R., and Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366.
- [Halliday, 2009] Halliday, D. Resnick, R. W. J. (2009). *Fundamentos de física: volumen dos*, volume 2. Grupo Editorial Patria, México.
- [He, 2021a] He, C.-C. (2021a). A fast and accurate analytic method of calculating galaxy two-point correlation functions. *The Astrophysical Journal*, 921(1):59.
- [He, 2021b] He, C.-C. (2021b). A fast and accurate analytic method of calculating galaxy two-point correlation functions. *The Astrophysical Journal*, 921(1):59.
- [Herschel, 1785] Herschel, W. (1785). Xii. on the construction of the heavens. *Philosophical Transactions of the Royal Society of London*, 75:213–266.
- [Hirvonen, 2023] Hirvonen, V. (2023). Einstein field equations: A step-by-step derivation (two ways).
- [Holmberg, 1941] Holmberg, E. (1941). On the Clustering Tendencies among the Nebulae. II. a Study of Encounters Between Laboratory Models of Stellar Systems by a New Integration Procedure. , 94:385.
- [Hu, 2022] Hu, W. (2022). Seing sound.
- [Hu et al., 2000] Hu, W., Barkana, R., and Gruzinov, A. (2000). Fuzzy cold dark matter: The wave properties of ultralight particles. *Physical Review Letters*, 85(6):1158–1161.
- [Huang et al., 2019] Huang, M., Bao, Q., Zhang, Y., and Feng, W. (2019). A hybrid algorithm for forecasting financial time series data based on dbSCAN and svr. *Information*, 10(3).
- [Hubble, 1926] Hubble, E. P. (1926). Extragalactic nebulae. , 64:321–369.

- [Ishiyama et al., 2021] Ishiyama, T., Prada, F., Klypin, A. A., Sinha, M., Metcalf, R. B., Jullo, E., Altieri, B., Cora, S. A., Croton, D., de la Torre, S., Millán-Calero, D. E., Oogi, T., Ruedas, J., and Vega-Martínez, C. A. (2021). The uchuu simulations: Data release 1 and dark matter halo concentrations. *Monthly Notices of the Royal Astronomical Society*, 506(3):4210–4231.
- [Jain, 2022] Jain, T. (2022). Spanning tree | minimum spanning tree.
- [Jimenez and Loeb, 2002] Jimenez, R. and Loeb, A. (2002). Constraining cosmological parameters based on relative galaxy ages. *The Astrophysical Journal*, 573(1):37–42.
- [Kaiser, 2007] Kaiser, N. (2007). The Pan-STARRS Survey Telescope Project. In Ryan, S., editor, *Advanced Maui Optical and Space Surveillance Technologies Conference*, page E9.
- [Kaplinghat et al., 2000] Kaplinghat, M., Knox, L., and Turner, M. S. (2000). Annihilating cold dark matter. *Physical Review Letters*, 85(16):3335–3338.
- [Karypis et al., 1999] Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon a hierarchical clustering algorithm using dynamic modeling. *Computer*, 32:68 – 75.
- [Kerscher, 1999] Kerscher, M. (1999). The geometry of second-order statistics - biases in common estimators. , 343:333–347.
- [Kerscher et al., 2000a] Kerscher, M., Szapudi, I., and Szalay, A. S. (2000a). A comparison of estimators for the two-point correlation function. *The Astrophysical Journal*, 535(1):L13–L16.
- [Kerscher et al., 2000b] Kerscher, M., Szapudi, I., and Szalay, A. S. (2000b). A comparison of estimators for the two-point correlation function. *The Astrophysical Journal*, 535(1):L13.
- [Kneib and Natarajan, 2012] Kneib, J.-P. and Natarajan, P. (2012). Cluster lenses. *The Astronomy and Astrophysics Review*, 19.
- [Kumar, 2021] Kumar, V. (2021). Tutorial for dbscan clustering in python sklearn.
- [Kuzelewska and Wichowski, 2015] Kuzelewska, U. and Wichowski, K. (2015). A modified clustering algorithm dbscan used in a collaborative filtering recommender system for music recommendation. In *Theory and Engineering of Complex Systems and Dependability: Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, June 29–July 3 2015, Brunów, Poland*, pages 245–254. Springer.
- [Lara Torralbo, 2010] Lara Torralbo, J. A. (2010). *Marco de Descubrimiento de Conocimiento para Datos Estructuralmente Complejos con Énfasis en el Análisis de Eventos en Series Temporales*. [tesis doctoral de ingeniería informática], Universidad Politécnica de Madrid. Repositorio UPM. https://oa.upm.es/5729/1/JUAN_ALFONSO_LARA_TORRALBO.pdf.

- [Lemson, 2006] Lemson, G. (2006). Halo and galaxy formation histories from the millennium simulation: Public release of a vo-oriented and sql-queryable database for studying the evolution of galaxies in the cdm cosmogony.
- [Li and Li, 2007] Li, X. and Li, D. (2007). Discovery of rules in urban public facility distribution based on DBSCAN clustering algorithm. In Wang, Y., Li, J., Lei, B., and Yang, J., editors, *MIPPR 2007: Remote Sensing and GIS Data Processing and Applications; and Innovative Multispectral Technology and Applications*, volume 6790 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 67902E.
- [Liddle, 2003] Liddle, A. (2003). *An Introduction to Modern Cosmology*. Wiley.
- [Liddle, 2015] Liddle, A. (2015). *An Introduction to Modern Cosmology*. Wiley.
- [M. Manero, 2020] M. Manero, V. (2020). El paralaje, el cálculo matemático para medir la distancia a las estrellas.
- [Malzer and Baum, 2020] Malzer, C. and Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.
- [Martinez Heras, 2020] Martinez Heras, J. (2020). Clustering (agrupamiento), k-means con ejemplos en python.
- [Matt, 2019] Matt, O. (2019). Sound waves from the beginning of time.
- [Mohammed et al., 2018] Mohammed, N. N., Cawthorne, M., and Abdulazeez, A. M. (2018). Detection of genes patterns with an enhanced partitioning-based dbscan algorithm. *Journal of information and communication engineering*, 4(1):188–195.
- [Morissette and Chartier, 2013] Morissette, L. and Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in mathematica. *Tutorials in quantitative methods for psychology*.
- [Nadathur et al., 2019] Nadathur, S., Carter, P. M., Percival, W. J., Winther, H. A., and Bautista, J. E. (2019). Beyond BAO: Improving cosmological constraints from BOSS data with measurement of the void-galaxy cross-correlation. *Physical Review D*, 100(2).
- [NASA, 2014] NASA, W. S. T. . (2014). Geometry of the universe.
- [Newton, 1687] Newton, I. (1687). *Philosophiae naturalis principia mathematica*. W. Dawson.
- [Padilla et al., 2021] Padilla, L. E., Tellez, L. O., Escamilla, L. A., and Vazquez, J. A. (2021). Cosmological parameter inference with bayesian statistics. *Universe*, 7(7):213.

- [Pavlis et al., 2017] Pavlis, M., Dolega, L., and Singleton, A. (2017). A modified DBS-CAN clustering method to estimate retail center extent. *Geographical Analysis*, 50(2):141–161.
- [Perlmutter et al., 1999] Perlmutter, S., Turner, M. S., and White, M. (1999). Constraining dark energy with type Ia supernovae and large-scale structure. *Physical Review Letters*, 83(4):670–673.
- [Ponce et al., 2012] Ponce, R., Cardenas-Montes, M., Rodriguez-Vazquez, J. J., Sanchez, E., and Sevilla, I. (2012). Application of gpus for the calculation of two point correlation functions in cosmology.
- [Pons-Borderia et al., 1999] Pons-Borderia, M.-J., Martinez, V. J., Stoyan, D., Stoyan, H., and Saar, E. (1999). Comparing estimators of the galaxy correlation function. *The Astrophysical Journal*, 523(2):480–491.
- [Rehioui et al., 2016] Rehioui, H., Idrissi, A., Abourezq, M., and Zegrari, F. (2016). Denclue-im: A new approach for big data clustering. *Procedia Computer Science*, 83:560–567. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.
- [Riess et al., 1998] Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., and Tonry, J. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116(3):1009–1038.
- [Riess et al., 2016] Riess, A. G., Macri, L. M., Hoffmann, S. L., Scolnic, D., Casertano, S., Filippenko, A. V., Tucker, B. E., Reid, M. J., Jones, D. O., Silverman, J. M., Chornock, R., Challis, P., Yuan, W., Brown, P. J., and Foley, R. J. (2016). A 2.4% DETERMINATION OF THE LOCAL VALUE OF THE HUBBLE CONSTANT. *The Astrophysical Journal*, 826(1):56.
- [Rodríguez, 2010] Rodríguez, L. (2010). Materia oscura, energía oscura.
- [Rubin and Ford, 1970] Rubin, V. C. and Ford, W. Kent, J. (1970). Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. , 159:379.
- [Schubert and Rousseeuw, 2021] Schubert, E. and Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804.
- [Schutz, 2009] Schutz, B. (2009). *Perfect fluids in special relativity*. Cambridge University Press.

- [SDSS-II_Collaboration, 2014] SDSS-II_Collaboration (2014). Sdss-iii: Four surveys executed simultaneously.
- [SDSS_collaboration, 2022] SDSS_collaboration (2022). The extended baryon oscillation spectroscopic survey (eboss).
- [Sinclair, 2019] Sinclair, C. (2019). Clustering using optics.
- [Sjödahl, 2019] Sjödahl, M. (2019). Gradient correlation functions in digital image correlation. *Applied Sciences*, 9(10).
- [Spergel and Steinhardt, 2000] Spergel, D. N. and Steinhardt, P. J. (2000). Observational evidence for self-interacting cold dark matter. *Phys. Rev. Lett.*, 84:3760–3763.
- [Takada and Jain, 2003] Takada, M. and Jain, B. (2003). The three-point correlation function in cosmology. *Monthly Notices of the Royal Astronomical Society*, 340(2):580–608.
- [Tippens and Ruiz, 2006] Tippens, P. and Ruiz, Á. (2006). *Física: conceptos y aplicaciones*. McGraw-Hill Interamericana.
- [Torres Rudloff, 2017] Torres Rudloff, N. I. (2017). *Clustering y diversidad en sistemas de recomendación top-N*. [tesis para optar al grado académico de magíster en ciencias de la ingeniería informática], Universidad técnica Federico Santa María. Repositorio USM. <https://repositorio.usm.cl/bitstream/handle/11673/22693/3560900231825UTFSM.pdf?sequence=1&isAllowed=y>.
- [Téllez Tovar, 2018] Téllez Tovar, L. O. (2018). *Constricciones de modelos de materia y energía oscura escalar*. [tesis para obtener el grado de maestro en ciencias en la especialidad de física], CINVESTAV. Repositorio del Dr. José Alberto Vázquez. https://www.fis.unam.mx/~javazquez/file_thesis.html.
- [Vargas-Magaña et al., 2013] Vargas-Magaña, M., Bautista, J. E., Hamilton, J.-C., Busca, N. G., Aubourg, É., Labatie, A., Goff, J.-M. L., Escoffier, S., Manera, M., McBride, C. K., Schneider, D. P., and Willmer, C. N. A. (2013). An optimized correlation function estimator for galaxy surveys. *Astronomy & Astrophysics*, 554:A131.
- [Verde, 2021] Verde, L. Gil-Marín, H. (2021). Cartografiando el universo. *Sociedad Española de Astronomía*, 45(6-19).
- [Vázquez González, 2008] Vázquez González, A., . M. T. (2008). La materia oscura del universo: retos y perspectivas. *Revista mexicana de física E*, 2(54):193–202.
- [Wei and Sun, 2019] Wei, J. and Sun, S. (2019). Commercial activity cluster recognition with modified dbscan algorithm: A case study of milan. *2019 IEEE International Smart Cities Conference (ISC2)*, pages 228–234.
- [Wendy,] Wendy, H. . Recitation – soft k-means clustering.

- [Yang et al., 2014] Yang, Y., Lian, B., Li, L., Chen, C., and Li, P. (2014). Dbscan clustering algorithm applied to identify suspicious financial transactions. In *Proceedings of the 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CYBERC '14*, page 60–65, USA. IEEE Computer Society.
- [Zhang, 2019] Zhang, M. (2019). Use density-based spatial clustering of applications with noise (dbscan) algorithm to identify galaxy cluster members. *IOP Conference Series: Earth and Environmental Science*, 252(4):042033.
- [Zhang et al., 2022a] Zhang, R., Peng, H., Dou, Y., Wu, J., Sun, Q., Zhang, J., and Yu, P. S. (2022a). Automating dbscan via deep reinforcement learning.
- [Zhang et al., 2022b] Zhang, R., Peng, H., Dou, Y., Wu, J., Sun, Q., Zhang, J., and Yu, P. S. (2022b). Automating dbscan via deep reinforcement learning.
- [Zhang et al., 1996] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD '96*, page 103–114, New York, NY, USA. Association for Computing Machinery.
- [Zwicky, 1937] Zwicky, F. (1937). On the Masses of Nebulae and of Clusters of Nebulae. , 86:217.