



FACULTAD DE CIENCIAS EXACTAS, INGENIERÍA Y  
AGRIMENSURA

DEPARTAMENTO DE FÍSICA

**Exploración automática de la correlación entre  
trazadores de vacíos cósmicos en simulaciones  
cosmológicas**

Franco Schivazappa

Director: Dr. Andrés Nicolás Ruiz

Co-Direcor: Dr. Gabriel Ignacio Perren

Mayo de 2024, Rosario

---

Esta obra está bajo una licencia  
Creative Commons Reconocimiento 4.0 Internacional



## Agradecimientos

En primer lugar quiero agradecer a las caras en mi monitor, que durante el último año dedicaron su tiempo semana tras semana para conectarse a las reuniones virtuales para guiarme en este proyecto: a Andrés por dirigirme, a Juan por ponerme en contacto desde el inicio con este excelente grupo de trabajo, a Carlos y a Dante por su entusiasmo contagioso por el tema. Y lógicamente a los cuatro por el conocimiento que me brindaron, la paciencia que me tuvieron, y todos los consejos y sugerencias que me dieron.

Agradezco a mis padres por todas las oportunidades que me han dado en la vida, por aguantarme todo el tiempo que me tomó llegar hasta este punto y por darme su apoyo cuando opté por seguir este camino.

Agradezco a Marcos, que en los tres años y un poquito que nos conocemos me elevó como persona, me acompañó cada paso del camino celebrando todos mis logros e hizo de ese camino uno de los mejores períodos de mi vida.

Agradezco a Flor, mi bra, compañera de experimental, compañera de crisis. A mi tocayo Franco, que hicimos la carrera prácticamente a la par, por todas las sesiones de estudio y los mates. A Esteban, Pablo y Nahuel, por su valiosa amistad durante el trayecto que fue mi carrera.

Agradezco a Roki, Fati y Lari por estar ahí siempre ofreciendo un respiro de la vorágine académica.

## Resumen

En el presente trabajo se buscó explorar la hipótesis de que existe una relación entre las curvas de abundancia de vacíos cósmicos identificados en halos de galaxias y en materia. Esto permitiría sortear el obstáculo que supone ser la relación de bias a la hora de analizar la distribución de materia en el Universo. Se emplearon relevamientos de vacíos cósmicos obtenidos a partir de simulaciones cosmológicas con  $\Omega = [0,1; 0,5]$ , tomando un paso de  $\Omega = 0,025$ , y densidades de halos  $\delta_h$  que varían entre 0,001, 0,002 y 0,004. Para poner a prueba la hipótesis, se exploró la relación entre las curvas utilizando tres tipos de regresores: un Regresor lineal, un Random Forest Regressor y una Regresión gaussiana. Con esto se buscó establecer si al entrenar un algoritmo con información sobre la abundancia de vacíos identificados con ambos trazadores, se puede predecir el comportamiento de la curva de abundancia en materia al entregarle únicamente datos de la curva de abundancia en halos.

# Índice general

Índice de figuras	5
Índice de tablas	8
Glosario	8
<b>1. Introducción</b>	<b>11</b>
<b>2. Objetivos</b>	<b>14</b>
<b>3. Cosmología</b>	<b>15</b>
3.1. Universos FLRW . . . . .	17
3.2. Modelo cosmológico estándar . . . . .	20
3.3. El problema de la energía oscura . . . . .	26
3.4. Vacíos cósmicos y por qué nos interesan . . . . .	28
3.5. La informática y la cosmología . . . . .	35
3.5.1. Simulaciones cosmológicas . . . . .	36
<b>4. Machine learning</b>	<b>38</b>
4.1. Muchas formas de aprender . . . . .	39
4.1.1. El aprendizaje supervisado . . . . .	40
4.2. Algoritmos de regresión . . . . .	41

4.2.1. Regresión Lineal . . . . .	41
4.2.2. Gaussian Process Regressor . . . . .	42
4.2.3. Random Forest Regressor . . . . .	46
4.3. Métricas para validación de modelos . . . . .	50
4.4. Validación cruzada . . . . .	53
4.5. Búsqueda de Hiperparámetros . . . . .	54
<b>5. Datos y diseño experimental</b>	<b>56</b>
5.1. Conjuntos de datos utilizados . . . . .	56
5.2. Entrenamiento del algoritmo . . . . .	59
<b>6. Resultados</b>	<b>63</b>
6.1. Regresión Lineal . . . . .	65
6.2. Gaussian Process Regressor . . . . .	68
6.3. Random Forest Regressor . . . . .	70
6.4. Comparación del Mean Squared Error (MSE) acorde al parámetro de densidad . . . . .	73
<b>7. Discusión y conclusiones</b>	<b>75</b>
<b>A. Abundancia</b>	<b>78</b>
<b>B. Regresión lineal</b>	<b>81</b>
<b>C. Gaussian Process Regressor</b>	<b>91</b>
<b>D. Random Forest Regressor</b>	<b>101</b>
<b>Bibliografía</b>	<b>111</b>

# Índice de figuras

3.1.	Esquema de la evolución del Universo, desde el Big Bang, seguido inmediatamente por el período de inflación inicial hasta el presente. Créditos: NASA / WMAP Science Team. . . . .	23
3.2.	Red cósmica. En esta imagen, cada punto representa una galaxia, con el color denotando la densidad. Crédito: Sloan Digital Sky Survey. . . . .	25
4.1.	Proceso genérico de un Gaussian Process Regressor (GPR). A la izquierda se aprecian funciones aleatorias previas a la introducción de los datos. A la derecha algunas de las funciones luego de condicionar con 5 observaciones [Murphy, 2012]. . . . .	45
4.2.	Esquema de un árbol de decisión aplicado al análisis de ilustraciones [Criminisi et al., 2011]. . . . .	47
4.3.	Ejemplo genérico de datos ajustados por una Regresión Lineal (RL). Al evaluar este ejemplo, la gráfica naranja obtiene $R^2 = 0$ , es decir que la predicción no tiene ninguna similitud con los datos, mientras que la gráfica verde obtiene $R^2 = 0,81$ . . . . .	52

4.4. Ejemplo genérico de datos ajustados por una RL. Al evaluar este ejemplo, la gráfica naranja obtiene $MSE = 15,13$ , es decir que la predicción no tiene ninguna similitud con los datos, mientras que la gráfica verde obtiene $MSE = 0,65$ .	52
4.5. Representación esquemática del proceso de validación cruzada, para un conjunto $\mathcal{D}$ dividido en 10 subconjuntos. En la primera etapa, el primer subconjunto sirve como conjunto de prueba. En la segunda etapa, este rol lo ocupa el subconjunto siguiente, y así se continúa hasta llegar al décimo subconjunto [Berrar, 2019].	54
5.1. Cortes de dos de las simulaciones empleadas en el proyecto, para los valores extremos de $\Omega_m$ . Queda en evidencia como a mayores valores de $\Omega_m$ , la estructura a gran escala se vuelve más densa, consecuencia natural del aumento de materia que habita la simulación.	57
5.2. Remoción de los valores espurios en las gráficas.	59
6.1. Función de Tamaño de Vacíos (VSF) en materia y halos para $\Omega_m = 0,3$ . Se ponen en evidencia las diferencias que surgen de las abundancias en distintos trazadores: los vacíos identificados en partículas, son más pequeños que los identificados en halos.	64
6.2. Detalle de las curvas de abundancia identificadas en halos para $\Omega_m = 0,3$ .	64
6.3. Se ilustra como impacta el número de bines en el gráfico de la curva de abundancia. Se usa la simulación con $\Omega_m = 0,300$ , la misma que fue usada como ejemplo en la Fig. 6.1.	65
6.4. VSF para $\Omega_m = 0,475$ y $\delta_h = 0,004$ obtenida a través de una regresión lineal.	66

6.5. Correspondencia entre valores del set de prueba y los valores predichos por el algoritmo de regresión lineal. La función identidad está a modo de ayuda visual. . . . .	67
6.6. Contraste en las predicciones para distintos parámetros de densidad de materia, $\Omega_m = 0,100$ y $\Omega_m = 0,475$ . . . . .	67
6.7. VSF para $\Omega_m = 0,450$ y $\delta_h = 0,001$ obtenida a través de un Gaussian Process Regressor. . . . .	69
6.8. Correspondencia entre los valores simulados y los valores predichos por el algoritmo GPR para $A_v$ y $R_v$ . La función identidad está a modo de ayuda visual. . . . .	70
6.9. VSF para $\Omega_m = 0,200$ y $\delta_h = 0,002$ obtenida con Random Forest Regressor. . . . .	71
6.10. Correspondencia entre valores del conjunto de prueba y los valores predichos por el algoritmo Random Forest Regressor (RF). La función identidad está a modo de ayuda visual. . . . .	72
6.11. Comparación en las predicciones para $\Omega_m = 0,100$ y $\Omega_m = 0,450$ . . . . .	73
6.12. Métrica MSE obtenida para cada predicción de $A_v$ , es decir, variando desde $\Omega_m = 0,1$ a $\Omega_m = 0,5$ , en las tres densidades de halos. . . . .	74
6.13. Contraste entre las métricas MSE en los ajustes $R_v$ , para los tres regresores. En este caso no se evidencia tendencia a disminuir conforme aumenta $\Omega_m$ . . . . .	74

# Índice de cuadros

5.1. Primeras filas de la matriz de entrada con la que se entrena los algoritmos. Las primeras seis columnas corresponden a los features de los regresores, donde: $\Omega_m$ es el parámetro de densidad de materia de la simulación, $\delta_h$ es la densidad de halos, ABH es el Ancho de Bins en Halos, $R_vH$ es la Cota superior de Radios en Halos para cada bin, $A_vH$ es el Conteo de vacíos identificados en Halos y ABP es el Ancho de Bins en Partículas. Finalmente, las ultimas dos columnas corresponden cada una a las variables objetivo que pretenden reproducir los regresores entrenados: $R_vP$ es la Cota superior de Radios en Partículas para cada bin, y $A_vP$ es el Conteo de vacíos identificados en Partículas. . . . .	60
7.1. Valores promedio de las métricas para los ajustes de radio y abundancia para los tres regresores. . . . .	76

# Glosario

<b>ΛCDM</b>	Lambda Cold Dark Matter	8, 14–16, 20, 22, 24, 27, 32, 35, 57, 75
<b>CMB</b>	Radiación de Fondo de Microondas	8, 21, 22, 26, 28, 35
<b>DS</b>	Ciencia de Datos	8, 35
<b>FLRW</b>	Métrica Friedmann-Lemaître-Robertson-Walker	3, 8, 17–19
<b>GPR</b>	Gaussian Process Regressor	4, 5, 7, 8, 41, 42, 44, 45, 61, 68–70, 76, 92–94
<b>GS</b>	Grid Search	8, 54, 55, 61, 68, 70
<b>IA</b>	Inteligencia Artificial	8, 35, 38, 39
<b>ML</b>	Machine Learning	8, 35, 37–39, 50, 75
<b>MSE</b>	Mean Squared Error	4, 7, 8, 48, 51, 53, 62, 65, 73–75
<b>MVN</b>	Distribución Normal Multivariable	8, 43
<b>RBF</b>	Radial Basis Function	8, 46, 68, 69
<b>RF</b>	Random Forest Regressor	4, 7, 8, 41, 46, 48–50, 61, 70–72, 75, 76
<b>RL</b>	Regresión Lineal	4–6, 8, 41, 51, 52, 61, 65, 66, 68, 69, 72, 73, 76
<b>RQ</b>	Rational Quadratic	8, 46, 68, 69

**VSF** Función de Tamaño de Vacíos. 6–8, 12–14, 31, 34, 37, 41, 59, 61–67, 69, 71, 73, 75, 76, 79, 80, 82–84, 92–94, 102–104

# Capítulo 1

## Introducción

En el año 1978, S. Gregory y L. A. Thompson postularon que, contrario al modelo de E. Hubble que planteaba un Universo homogéneo, éste en realidad estaba compuesto a gran escala por distintas estructuras: cúmulos de galaxias, filamentos y vacíos cósmicos. Con el correr de los años, los avances tecnológicos han permitido escrutar regiones cada vez más amplias del cielo, dejando en evidencia un enorme número de estas estructuras. Gracias a esto, a principios de la década de 1980, J. Peebles introdujo el concepto de materia oscura fría (o CDM por sus siglas en inglés) como explicación al fenómeno de la formación de las mismas [Thompson, 2020, Peebles, 1982].

Los vacíos cósmicos son estructuras de gran prominencia al analizar el Universo a gran escala, puesto que ocupan el mayor porcentaje de su volumen. Se trata de regiones subdensas del espacio, que resultan del agrupamiento de la materia en otras regiones aledañas por acción de la gravedad, y esta característica hace que sean de gran utilidad para estudiar la geometría e historia del Universo [Contarini et al., 2019].

Gracias a su estructura y forma relativamente simples, los vacíos representan un ambiente ideal para poner a prueba una variedad de teorías cosmológicas. Sin embargo, para que dichos estudios tengan éxito, es necesario conocer el sesgo en el mapeo de

los vacíos, dado que pueden influenciar la interpretación de los datos llevando así a conclusiones erróneas. Las propiedades estadísticas de los vacíos cósmicos dependen de dos factores: los trazadores de materia usados, y el método para identificarlos a partir de la distribución espacial de los mismos [Contarini et al., 2019, Correa, 2021].

La Función de Tamaño de Vacíos (VSF) es una de las herramientas estadísticas que existen para estudiar los vacíos, y un obstáculo al momento de realizar los análisis es que en la vida real sólo se puede medir su VSF utilizando galaxias como trazadores. La VSF para vacíos definidos en el campo de densidad de materia oscura, que es la que sería de mayor utilidad, resulta observacionalmente inaccesible, debido a la imposibilidad de medir el campo de densidad de materia oscura del Universo.

Este problema que presentan los vacíos posee una naturaleza un tanto abstracta, por lo que quizás sea de utilidad explicarlo con una analogía:

*Supóngase un cartógrafo, con interés en realizar un mapa del relieve de una cadena montañosa, con todos sus picos y valles. Para llevar a cabo tal tarea, el cartógrafo se sube a un helicóptero para obtener una vista de pájaro de la región, con la esperanza de que cada detalle del terreno será visible (para completitud de la analogía, también se debe suponer que el pobre cartógrafo no tiene acceso a imágenes satelitales para realizar su trabajo).*

*Sin embargo, cuando llega al área, el cartógrafo se lleva una sorpresa: todo el terreno se encuentra cubierto por nubes, a excepción de los picos más altos. En este punto, podría tomar la desatinada decisión de asumir que sólo existen montañas en donde sus cimas sobresalen de las nubes y que el resto del terreno es un valle pero, ¿acaso no es posible que existan cerros de menor altura ocultos por el mal tiempo? Este mapa resultaría sesgado en consecuencia.*

Para los vacíos, ocurre lo mismo: identificar vacíos en halos de galaxia presen-

ta una VSF incompleta al compararla con la que sería obtenida identificando a los vacíos usando toda la distribución de materia en el Universo, es decir, incluyendo a la distribución de materia oscura.

Comprender las propiedades de los vacíos resulta de gran importancia para poder utilizar dichas estructuras como laboratorios cosmológicos, permitiendo así el estudio de la formación de galaxias, la naturaleza de la materia y energía oscuras, la evolución a gran escala del Universo, entre otros temas.

Los resultados parciales de este trabajo fueron presentados en la charla *Automatic exploration of the correlation between cosmic void tracers in cosmological simulations*, en la conferencia “XIII Friends-Of-Friends Meeting”<sup>1</sup> en mes de Abril del 2024, en el Observatorio Astronómico de Córdoba, Córdoba, Argentina.

Así este trabajo tiene la siguiente estructura: en el Capítulo 2 se listan los objetivos del proyecto, seguidos de los Capítulos 3 y 4 en donde se introducen los conceptos necesarios para tratar el problema, luego en el Capítulo 5 se expone la metodología utilizada, y los resultados en el Capítulo 6. Finalmente, en el Capítulo 7 se elaboran las conclusiones del trabajo.

---

<sup>1</sup><https://fof.oac.uncor.edu/2024/program/>

# **Capítulo 2**

## **Objetivos**

El objetivo de este trabajo es emplear distintos métodos de aprendizaje automático para inferir el comportamiento de la VSF identificada en materia oscura, a partir de la VSF identificados en halos de galaxias.

La labor se lleva a cabo sobre simulaciones de universos Lambda Cold Dark Matter ( $\Lambda$ CDM) planos creadas para este proyecto, en las cuales se identifican los radios de los vacíos cósmicos utilizando el identificador esférico para los dos trazadores de interés: halos de galaxia y partículas.

Una vez realizado el ajuste de los regresores se evalúa su desempeño al realizar las predicciones.

# Capítulo 3

## Cosmología

Al día de hoy, el modelo cosmológico del Universo más ampliamente aceptado en la comunidad científica se denomina Lambda Cold Dark Matter ( $\Lambda$ CDM), y para entender por qué ha tomado protagonismo este modelo en el presente, es conveniente analizar brevemente los hitos de la cosmología en el último siglo que le dieron tal relevancia.

Se puede considerar que la cosmología como se la conoce hoy en día tuvo sus inicios cuando Albert Einstein postuló la teoría general de la relatividad en 1917 [Einstein, 1917]. Este avance modificó la manera en que los astrónomos estudiaban los objetos distantes, dejando atrás la noción de que el Universo posee una estructura estática.

Más adelante, en 1929, Hubble descubrió que el Universo se encuentra en expansión, basándose en el hecho de que la longitud de onda de la luz emitida por los objetos distantes se estira a causa del efecto Doppler [Hubble, 1929]. Si una fuente emite fotones con una longitud de onda  $\lambda$ , un observador en el presente percibirá una longitud de onda  $\lambda_0$  que varió proporcionalmente al factor de escala de Universo  $a(t)$ . Este fenómeno es conocido popularmente como corrimiento al rojo o *redshift*, pues ese es el extremo del espectro al que  $\lambda$  se suele acercar [Tsujikawa, 2018].

En la década de 1930, Georges Lemaître postula por primera vez una versión rudimentaria de la teoría del Big Bang, en la que sostiene que el Universo tuvo su inicio a partir de una singularidad, un estado de alta densidad que dio origen al espacio y al tiempo, y que evolucionó durante 13000 millones de años hasta llegar al estado actual. La base matemática del modelo, al igual que la base del modelo  $\Lambda$ CDM, deriva de la teoría general de la relatividad<sup>1</sup> [Thompson, 2020, Lemaître, 1931].

También en la década de 1930, estudios llevados a cabo por Fritz Zwicky presentaron evidencia de la existencia de materia oscura, implicando que la materia visible representa sólo una pequeña fracción total de la materia en el Universo [Zwicky, 1933]. Si bien no es la primera mención del concepto en la literatura<sup>2</sup>, y pasarían años hasta que sus hallazgos fueran tenidos en cuenta, fue un importante resultado que tendría eco en investigaciones a futuro.

En los años siguientes, nuevas publicaciones fueron aportando información en el área, a la vez que creaban debate en las posibles interpretaciones de los resultados, y para mediados de la década de 1970 la mayoría de los astrónomos estaban convencidos de la existencia de masa no detectada en cantidades cosmológicamente significativas en el Universo. Tal es así que los primeros modelos computacionales del Universo que se construyeron para simular la formación de estructuras a gran escala no podían explicar la distribución de galaxias, cúmulos y vacíos observada partiendo de materia ordinaria [Thompson, 2020].

A inicios de la década de 1980 se introdujo el término de *materia oscura fría* como respuesta a la pregunta de qué podría ser la materia no detectada [Peebles, 1982, Blumenthal et al., 1984]. La materia oscura sólo interactúa de manera

---

<sup>1</sup>La teoría general de la relatividad, en una forma propuesta por Alexander Friedmann [Friedmann, 1922].

<sup>2</sup>En aquel momento, Zwicky no usó el término “materia oscura” de la misma manera que se usa en la época actual, además de que anteriormente también se usaba el término de con un significado diferente.

gravitatoria, y el término "frío" hace referencia a que se mueve a velocidades relativamente bajas comparadas con la velocidad de la luz. Este modelo es el que domina la especulación en el tema, y al día de hoy la estimación más popular describe a la composición del Universo como  $\sim 70\%$  energía oscura,  $\sim 25\%$  materia oscura y  $\sim 5\%$  materia bariónica [Planck Collaboration, 2020]. La materia oscura cumple un rol importante en el desarrollo de esta tesis.

Fue en 1998 que dos proyectos independientes, el Supernova Cosmology Project y el High-Z Supernova Search Team, llegan a la conclusión de que el Universo tiene en el presente un ritmo de expansión acelerado, gracias al estudio de supernovas tipo Ia. Esto representó en su momento un descubrimiento sorprendente, y una reintroducción de una constante  $\Lambda$  propuesta casi un siglo antes por Einstein al postular sus ecuaciones de campo, que había sido desestimada por décadas.  $\Lambda$  es la responsable de la expansión y tiene un efecto que se puede pensar como una fuerza gravitatoria repulsiva. Esta fue una de las últimas adiciones de mayor importancia al modelo cosmológico actual [Riess et al., 1998, Perlmutter et al., 1998].

### 3.1. Universos FLRW

La teoría de la Relatividad General establece que la geometría del espacio-tiempo está directamente vinculada a fuentes de materia a través de las ecuaciones de campo de Einstein. Para resolverlas, lo que se hace es proponer un tensor energía-momento (asociado a la densidad y al flujo de energía y momento en el espaciotiempo) y una métrica. Históricamente, Einstein propuso un término constante  $\Lambda$  en las soluciones de las ecuaciones de campo buscando justificar un modelo estático de universo, aunque la inestabilidad del mismo frente a pequeñas perturbaciones conduce con mucha facilidad a soluciones contractivas o expansivas[Tsujikawa, 2018].

La descripción matemática que sirve como fundamento para el modelo de universo

homogéneo, isotrópico y en concordancia con la teoría de la Relatividad General es la Métrica Friedmann-Lemaître-Robertson-Walker (FLRW). Se trata de una solución particular a las ecuaciones de campo de Einstein donde un elemento de línea en el espacio-tiempo se expresa como

$$\begin{aligned} ds^2 &= g_{\mu\nu} dx^\mu dx^\nu \\ &= -dt^2 + a^2(t) [\gamma_{ij}(x^k) dx^i dx^j] \end{aligned} \quad (3.1.1)$$

siendo  $g_{\mu\nu}$  el tensor métrico,  $\gamma_{ij}$  es un elemento de línea tridimensional independiente del tiempo,  $x^i$  son coordenadas comóviles y  $a(t)$  es un factor de escala dependiente del tiempo, que se encarga de describir la distancia entre dos puntos en el espacio conforme el Universo evoluciona. Utilizando coordenadas esféricas con una curvatura espacial  $K$ , la métrica FLRW se puede escribir

$$\gamma_{ij}(x^k) dx^i dx^j = \frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2 \theta d\varphi^2) \quad (3.1.2)$$

donde el signo de  $K$  caracteriza la geometría del Universo: se tiene un universo cerrado cuando  $K > 0$ , abierto cuando  $K < 0$  y plano cuando  $K = 0$  (lo que implica que posee una geometría euclíadiana). Cuando  $K = 0$  y  $a = 1$ , el elemento de línea representado en la ec. (3.1.1) corresponde a la métrica en el espacio-tiempo de Minkowski [Tsujikawa, 2018, Correa, 2021]

$$ds^2 = -c^2 dt^2 + dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2) \quad (3.1.3)$$

Las ecuaciones de campo de Einstein son las que gobiernan la geometría del espacio-tiempo, y se caracterizan por el tensor métrico  $g_{\mu\nu}$  y sus componentes energéticas, que a su vez son descritas por el tensor de energía-momento  $T_{\mu\nu}$

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} \quad (3.1.4)$$

siendo  $G$  la constante de gravitación y  $G_{\mu\nu}$  el tensor de Einstein, que a su vez es

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R \quad (3.1.5)$$

donde  $R_{\mu\nu}$  es el tensor de Ricci, y  $R = g^{\mu\nu}R_{\mu\nu}$  el escalar de curvatura.

Si se utiliza la métrica FLRW en las ecuaciones de campo de Einstein y se asume que el tensor de energía-momento es el de un fluido perfecto

$$T^{\mu\nu} = (p + \rho)U^\mu U^\nu + pg^{\mu\nu} \quad (3.1.6)$$

surgen ecuaciones para  $a(t)$ . Esta forma para el tensor  $T^{\mu\nu}$  es la más general para la restricción de que debe ser compatible con el principio cosmológico de homogeneidad e isotropía. Allí se encuentran presentes  $p$  y  $\rho$ , la presión y la densidad de energía respectivamente. Con esto, las ecuaciones de campo de Einstein se reducen a dos ecuaciones independientes, las ecuaciones de Friedmann [Correa, 2021]

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \quad (3.1.7)$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) \quad (3.1.8)$$

Estas ecuaciones rigen la manera en que el Universo se expande para modelos homogéneos e isotrópicos gobernados por la teoría de la relatividad de Einstein y en términos del factor de escala  $a(t)$ . Haciendo especial énfasis en el lado izquierdo de la ec. (3.1.7), el cociente elevado al cuadrado es en realidad el parámetro de Hubble, que se usa para describir el ritmo de expansión del Universo y se denota  $H$

$$H = \frac{\dot{a}(t)}{a(t)} \quad (3.1.9)$$

Es posible cuantificar la expansión del Universo en términos del factor de escala  $a(t)$ . Para un Universo homogéneo, isótropo y plano, esto queda dado por

$$H^2 = H_0^2 (\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_\Lambda) \quad (3.1.10)$$

donde  $a = 1/(1+z)$  es el factor de escala normalizado a la unidad en el presente (redshift  $z = 0$ ),  $H$  es el parámetro de Hubble y  $H_0$  es el valor de  $H$  en el tiempo

actual, que normalmente se refiere como constante de Hubble. Las cantidades  $\Omega_r$ ,  $\Omega_m$  y  $\Omega_\Lambda$  son parámetros de densidad adimensionales que se definen como

$$\Omega_m = \frac{\rho_m}{\rho_{crit}}, \quad \Omega_r = \frac{\rho_r}{\rho_{crit}}, \quad \Omega_\Lambda = \frac{\Lambda}{3H_0^2}, \quad (3.1.11)$$

donde  $\rho_m$  representa la densidad media de materia no relativista en el Universo en la actualidad,  $\rho_r$  representa la densidad media de radiación y materia relativista en la actualidad y  $\rho_{crit}$  es la densidad crítica, necesaria para cumplir la condición de un Universo plano ( $K = 0$ ), y se expresa como

$$\rho_{crit} = \frac{3H_0^2}{8\pi G} \quad (3.1.12)$$

La última de las fórmulas en la ec. (3.1.11) incluye un término  $\Lambda$ , que denota la constante cosmológica. Este término constante puede ser añadido a las ecuaciones de campo, y posee un comportamiento afín al de un fluido con presión negativa [Glover et al., 2014].

Al tratarse de cantidades normalizadas relacionadas con la densidad crítica, los parámetros  $\Omega$  cumplen con la siguiente relación [Peebles, 1993]

$$\Omega_m + \Omega_\Lambda + \Omega_r = 1 \quad (3.1.13)$$

Sin embargo, como a fines prácticos la contribución de la radiación al contenido del Universo en la actualidad es muy pequeña ( $\Omega_r \sim 9,2 \times 10^{-5}$  [Planck Collaboration, 2020]), se puede plantear que

$$\Omega_m + \Omega_\Lambda \simeq 1 \quad (3.1.14)$$

por lo que si se habla de la cantidad de materia en el Universo (que a su vez incluye a la materia bariónica y a la materia oscura,  $\Omega_m = \Omega_b + \Omega_{mo}$ ), se está hablando indirectamente de los aportes de  $\Lambda$  al plantear que  $\Omega_\Lambda \simeq 1 - \Omega_m$ .

## 3.2. Modelo cosmológico estándar

El modelo cosmológico estándar, también denominado Lambda Cold Dark Matter ( $\Lambda$ CDM), es la descripción del Universo vigente con mayor aceptación por parte de la

comunidad. Parte de dos hipótesis centrales:

- El Universo a gran escala es marcadamente simple, en el sentido que posee homogeneidad e isotropía del espacio. Al observarlo, se puede ver lo mismo en todas las direcciones, y no se identifica nada que se pueda denominar un “centro” o un “borde” [Peebles, 1993].
- Las interacciones gravitatorias se rigen por la teoría de la relatividad general de Einstein [Correa, 2021].

Y una serie de evidencias observacionales sirven para respaldar estas afirmaciones: [Correa, 2021]

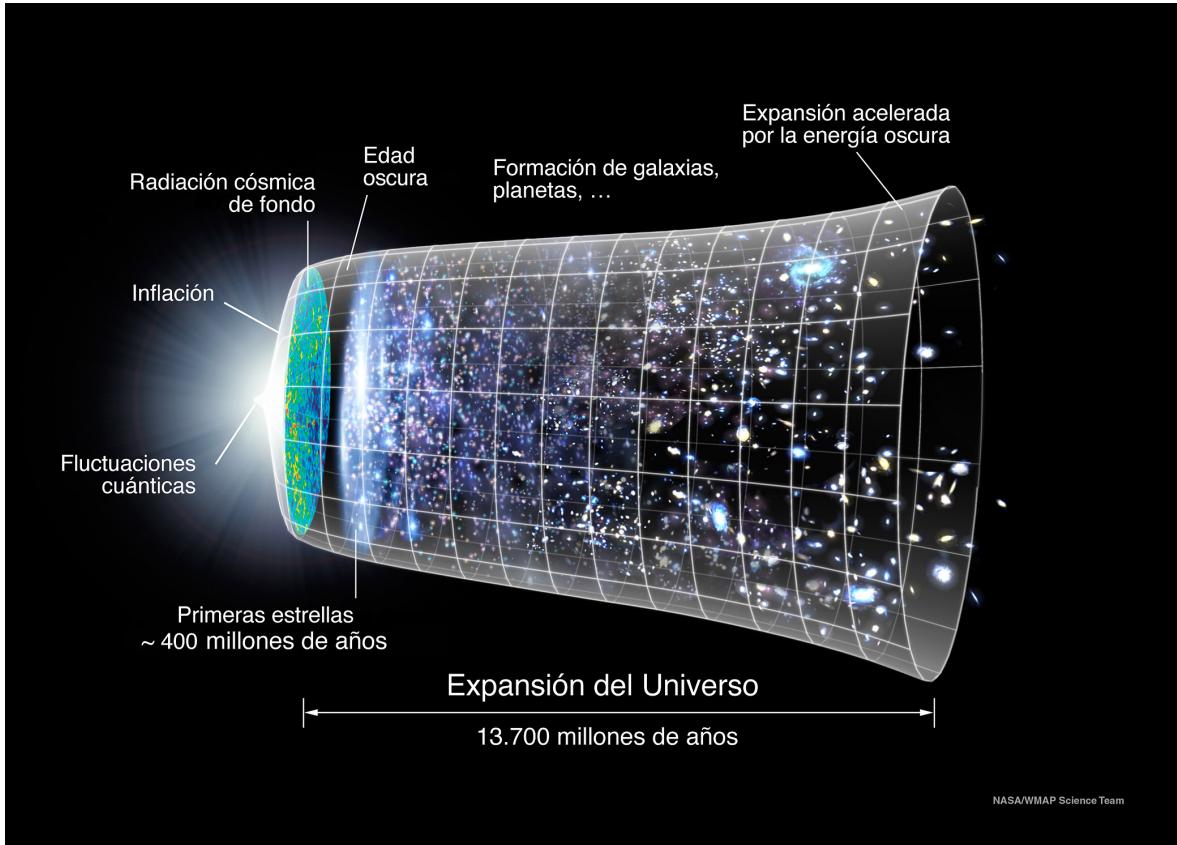
- Las galaxias aparentan alejarse de la Tierra. Más específicamente, el corrimiento de las líneas espectrales de cada una de ellas resulta en una velocidad radial positiva en casi todos los casos. Dicha velocidad tiene una relación de crecimiento lineal con la distancia a través de la ley de Hubble, motivo para interpretar que el Universo se encuentra en una expansión global [Hubble, 1929]. Siguiendo esta línea de pensamiento también se puede rastrear esta expansión cósmica hacia el pasado, manera en la que surge el modelo del Big Bang.
- La detección de la Radiación de Fondo de Microondas (CMB) en la década de 1960, una radiación electromagnética que proviene de todas direcciones y que posee un espectro de cuerpo negro. Esta radiación es una de las principales evidencias de la existencia del Big Bang [Penzias and Wilson, 1965].
- La composición uniforme del universo, con una densidad promedio de un átomo por metro cúbico. Posee una predominancia de hidrógeno del 74 %, seguida de un 25 % de helio, y el resto de los elementos contribuyendo al resto. Esta evidencia concuerda con lo predicho por la teoría de la nucleosíntesis [Thompson, 2020].

- Las observaciones del Universo favorecen un modelo de curvatura plana con 30 % de materia y un 70 % de un componente de naturaleza desconocida hasta el día de hoy, denominado energía oscura. Este componente implica a su vez que el Universo experimenta una expansión acelerada. De esto se desprende que la materia oscura y la energía oscura componen el 95 % del contenido energético del Universo [Correa, 2021].
- Los grandes relevamientos espectroscópicos de galaxias que se han llevado a cabo en los últimos años dejan ver que el Universo posee una estructura a grandes escalas, contrario a la noción de un siglo atrás de que las galaxias se distribuyen aleatoriamente. Debido a la interacción gravitatoria, las galaxias se agrupan en cúmulos, filamentos y paredes.

No existen observaciones de estructuras de un tamaño mayor a  $150h^{-1}Mpc$ , por lo que se puede hablar de la homogeneidad del Universo a partir de dicha escala.

- La distribución de temperatura del CMB es bastante homogénea e isótropa. Sin embargo, se observan pequeñas fluctuaciones relativas en temperatura del orden de  $10^{-5}$ , que sirven como semilla para la formación de estructuras.

La naturaleza evolutiva del modelo  $\Lambda$ CDM establece un vínculo entre las condiciones iniciales del Universo, es decir, la naturaleza física de la materia y energía en esa época, con las características del mismo que se pueden observar hoy en día.  $\Lambda$ CDM describe al inicio del Universo empleando la cosmología del Big Bang, con el respaldo de la abundante evidencia provista por el descubrimiento del CMB. En esta teoría, la primera etapa de evolución consistió en la expansión espacial del Universo en varios órdenes de magnitud en sólo una fracción de segundo, en un periodo que recibe el nombre de inflación. A lo largo de la existencia del Universo, su evolución estuvo dominada por distintos componentes. El modelo  $\Lambda$ CDM plantea que inicialmente estuvo dominado por la radiación, seguido de un dominio de la materia, y recientemente se ha



**Figura 3.1:** Esquema de la evolución del Universo, desde el Big Bang, seguido inmediatamente por el período de inflación inicial hasta el presente. Créditos: NASA / WMAP Science Team.

ingresado en una etapa dominada por la energía oscura. De manera esquemática, esta evolución se puede apreciar en la Fig. 3.1 [Frieman et al., 2008].

Es durante la etapa inflacionaria que aparecen fluctuaciones en la densidad de materia, que la inestabilidad gravitacional hizo crecer en el tiempo, para eventualmente originar las estructuras a gran escala en el Universo. Lo que en un principio fue una mezcla de gas y materia oscura fue evolucionando a una configuración en donde el gas colapsó al centro de halos de materia oscura. Se entiende por halo de materia oscura a las unidades básicas hacia donde la materia es atraída, y se los puede pensar como regiones de materia ligadas gravitacionalmente que se han desacoplado de la expansión

de Hubble y colapsaron [Wechsler and Tinker, 2018].

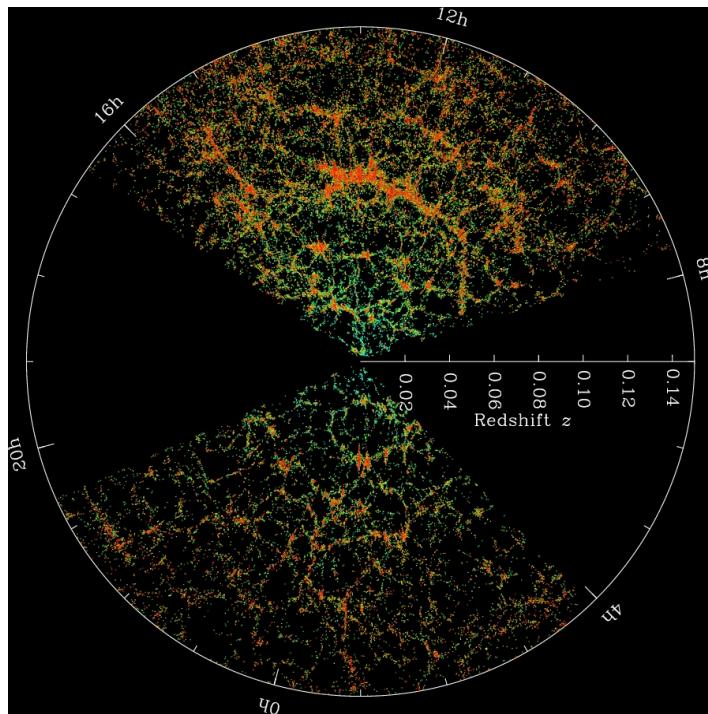
Para halos lo suficientemente grandes, se comenzaron a formar protogalaxias gracias a que el gas se enfrió y pudo formar estrellas. Con el pasar del tiempo, más estrellas se fueron formando en esas regiones, a la vez que los halos se fusionaban con otros halos, así también como sus respectivas protogalaxias. Estas configuraciones de materia iban creciendo hasta formar galaxias. El modelo  $\Lambda$ CDM contempla que las galaxias habitan el centro de sus respectivos halos de materia oscura, que poseen una estructura aproximadamente esférica que se extiende más allá del disco galáctico [Wechsler and Tinker, 2018].

Al observar la proyección de las galaxias distantes sobre la esfera celeste, se puede determinar la distancia a la que se encuentran de la Tierra gracias al corrimiento al rojo del espectro que emiten. Gracias a este procedimiento, es posible obtener un mapa tridimensional del Universo. En este mapa se pueden identificar regiones con densidades menores o mayores de galaxias. Estas últimas regiones con una mayor población de galaxias, formadas por acción de la gravedad, se denominan cúmulos [Schneider, 2006].

A esta altura se vuelve pertinente introducir el concepto de jerarquía en la cosmología, indicando con esto que se puede reconocer un nivel de organización de la materia en estructuras en un gran rango de órdenes de magnitud a nivel espacial. En la escala más pequeña, por ejemplo, están las galaxias.

Continuando en la jerarquía de tamaño, la estructura que le sigue a los cúmulos son los supercúmulos, y consisten en cúmulos de cúmulos, grupos y galaxias individuales. Estas estructuras llegan a medir hasta 100 Mpc. Cuando los supercúmulos se agrupan, forman las paredes, que probablemente son las estructuras de materia a gran escala de mayor tamaño en el Universo observable. Llegan a tamaños en el orden de los cientos de megaparsecs [NASA, 2024].

Al final de la jerarquía de las estructuras a gran escala en el Universo se encuentra



**Figura 3.2:** Red cósmica. En esta imagen, cada punto representa una galaxia, con el color denotando la densidad. Crédito: Sloan Digital Sky Survey.

la red cósmica, que consiste en el entramado total de todas las estructuras mencionadas previamente y se puede apreciar en la Fig. 3.2. Al igual que las estructuras de menor escala, su formación y evolución también se encuentra gobernada por la atracción gravitatoria de los objetos del Universo [NASA, 2024].

Es en esta escala de la jerarquía que se pueden identificar los vacíos cósmicos, los objetos de interés de este proyecto. Al agruparse la materia en las diversas estructuras que componen la red cósmica, se va concentrando en volúmenes cada vez más pequeños. En contraparte, el espacio que la materia “abandona” se convierte en una región subdensa que se va agrandando con el pasar del tiempo.

### 3.3. El problema de la energía oscura

Hasta donde se sabe, la materia y radiación ordinaria, como los bariones, los leptones y fotones, sólo conducen a modelos del Universo que debido a la acción de la gravedad y en concordancia con la teoría de la relatividad general, poseen una expansión desacelerada. Sin embargo, contrario a las expectativas, abundantes observaciones cosmológicas que se llevaron a cabo en los últimos años, particularmente sobre el comportamiento observado de las supernovas Ia, dan evidencia para pensar que el Universo se encuentra en una etapa de expansión acelerada, es decir, que  $\ddot{a} > 0$ . Esto hace necesaria la introducción de nuevos conceptos físicos para explicar estos fenómenos [Correa, 2021, Tsujikawa, 2018, Frieman et al., 2008].

La evidencia observational de la aceleración cósmica fue rápidamente aceptada por la comunidad científica, ya que proveía un elemento necesario para completar el modelo cosmológico actual. El origen físico de dicha expansión acelerada, sin embargo, todavía es un misterio. De acuerdo a lo observado, quedan dos rumbos a seguir para dar explicación al fenómeno. Por un lado, se introduce un nuevo componente al Universo denominado energía oscura, una forma energética con inusuales propiedades que ocupa el  $\sim 70\%$  de la densidad energética del Universo. La otra posibilidad es que la relatividad general deje de funcionar en ciertas escalas cosmológicas y se deba completar con otros modelos de gravedad. No se harán más comentarios sobre esta segunda posibilidad en este proyecto [Frieman et al., 2008].

En 2001 se lanzó al espacio el satélite WMAP con el fin de medir con precisión las anisotropías de temperatura del CMB, y con sus observaciones demostró que al día de hoy el Universo está compuesto aproximadamente por un 70% de energía oscura, un 25% de materia oscura, un 5% de átomos y un 0,01% radiación [Spergel et al., 2007]. Mientras que la materia oscura presenta efectos debido a la interacción gravitatoria, la energía oscura posee una presión efectiva negativa que la vuelve responsable de la expansión acelerada del Universo [Tsujikawa, 2018].

El candidato más simple para explicar la energía oscura es la constante cosmológica  $\Lambda$ , que posee propiedades cualitativamente similares a la energía del vacío<sup>3</sup>. Sin embargo, esta afirmación tampoco hace demasiado para aclarar el problema, que todavía está en vías de resolverse: la energía oscura posee propiedades muy llamativas y controvertidas. Se trata de un componente del Universo cuyos efectos puramente gravitacionales se manifiestan sólo a grandes escalas, y no pueden ser atribuidos a la materia ordinaria observada. Resulta curioso e irónico que la constante cosmológica haya surgido como una herramienta matemática que Einstein introdujo en las ecuaciones de campo intentando justificar un modelo de universo estático, que fue temporalmente desestimada cuando apareció la evidencia del Universo en expansión, y volvió a cobrar importancia con el problema de la expansión acelerada muchas décadas más tarde.

Con la introducción de  $\Lambda$ , la ec. (3.1.4) se convierte en

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} \quad (3.3.1)$$

Y teniendo en cuenta esta modificación, también sufren cambios las ecuaciones de Friedmann (3.1.7) y (3.1.8)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3} \quad (3.3.2)$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3} \quad (3.3.3)$$

El modelo  $\Lambda$ CDM presenta una gran cantidad de preguntas abiertas, desde el origen de la expansión en sí y la naturaleza de la energía oscura hasta la creación de la materia bariónica. El descubrimiento de la expansión acelerada del Universo es parte de esas cuestiones, y tiene importantes vínculos con otras interrogantes de la cosmología: el destino del Universo está ligado al entendimiento de la energía oscura; el período inflacionario podría estar relacionado con la expansión observada hoy en día;

---

<sup>3</sup>Cuantitativamente, sin embargo, los aportes de la energía del vacío difieren entre 50 y 120 órdenes de magnitud [Adler et al., 1995]

la energía oscura y la materia oscura podrían estar relacionadas; la aceleración cósmica podría ayudar a obtener un sucesor para la teoría de la gravedad de Einstein, entre otros [Frieman et al., 2008].

En el presente se utilizan distintos tests cosmológicos independientes entre sí para aportar información sobre la naturaleza de la energía oscura y sus propiedades: el uso de oscilaciones acústicas bariónicas (BAO) como reglas estándar [Eisenstein, 2005], el uso de supernovas de tipo Ia como candelas estándar para estudiar la historia de la expansión del Universo [Goliath et al., 2001], y también la reconstrucción de señales del CMB para analizar fluctuaciones primordiales [Ichiki and Sumiya, 2022]. El correcto funcionamiento del método propuesto en este proyecto y su posterior complejización para que refleje mejor las características observables del Universo harían posible darle un lugar como otro test independiente de la energía oscura.

### **3.4. Vacíos cósmicos y por qué nos interesan**

La acción de la fuerza de gravedad empuja de manera incesante la materia desde las regiones subdensas del Universo hacia las regiones más densas: las paredes, los filamentos y los cúmulos que componen la red cósmica. En los espacios que la materia deja atrás, se forma una red tridimensional de vacíos cósmicos, como proceso complementario. Estas enormes estructuras son las más grandes en el Universo, ocupando entre el 60 % y el 70 % del volumen total del mismo, y el pasar del tiempo sólo hará crecer esta proporción. A su vez, los vacíos sólo poseen entre el 10 % y el 20 % de la densidad promedio de trazadores que se utilizan para su identificación.

En un principio, los vacíos cósmicos y demás estructuras a gran escala surgieron de irregularidades en la distribución de materia oscura en la edad temprana del Universo. Probablemente surgieron durante, o al término, del periodo de inflación, y podían ser irregularidades de densidad positivas o negativas. Las primeras dieron lu-

gar al crecimiento de galaxias, y las segundas, que tenían una densidad inicial menor, evolucionaron en los vacíos cósmicos [Thompson, 2020].

Si bien las propiedades de los vacíos varían de acuerdo a la manera en que se los identifica, también existen propiedades básicas sobre las cuales hay consenso. Se puede generalizar al decir que son regiones subdensas con una densidad global del orden del 10 % al 20 % de la densidad promedio del Universo. Además, a la hora de identificarlos en los relevamientos cosmológicos, todos los métodos de identificación se ven afectados por distorsiones en el espacio de redshift y por el efecto Alcock-Paczynski, que impactan en el número, el tamaño y la posición del centro de los vacíos que se reporta. Esto a su vez influye en las estadísticas realizadas sobre los vacíos [Correa, 2021]. Si bien es importante tener en cuenta estas distorsiones para la completitud del modelo, en este proyecto no serán consideradas y se trabajará exclusivamente en el espacio real.

Como se mencionó antes, los vacíos crecen como proceso complementario al colapso de materia en las regiones sobredensas del Universo. En este trabajo se define a los vacíos como volúmenes esféricos, y de manera resumida la evolución que experimentan se puede describir en 8 puntos[Sheth and van de Weygaert, 2004]:

**Expansión:** los vacíos se expanden, contrario a lo que sucede con las regiones sobredensas, que colapsan.

**Evacuación:** a medida que los vacíos se expanden, su densidad interior decrece continuamente (en primer orden, debido a la reorganización de la masa dentro del vacío, y en segundo orden debido a la pérdida de masa en los bordes del vacío).

**Forma esférica:** la expansión hacia afuera hace que los vacíos tengan una geometría aproximadamente esférica.

**Perfil de densidad escalonado:** la repulsión “efectiva” de la materia en el interior del vacío decrece respecto a la distancia con su centro, de manera que el perfil de densidad presenta la forma de una función escalón.

**Campo de velocidad *Super-Hubble*:** De manera consistente con su distribución de densidad interior uniforme (final), el campo de velocidad de los vacíos tiene una divergencia de la velocidad interior constante, por lo que los vacíos evolucionan a “super burbujas-Hubble”.

**Crecimiento de estructuras suprimido:** las inhomogeneidades de densidad en el interior del vacío están suprimidas, y a medida que el objeto comienza a parecerse a un universo subdenso, la formación de estructuras allí se detiene en el tiempo.

**Formación de una cresta en las paredes:** A medida que la materia en el interior del vacío se acumula cerca del borde, se forma una cresta a su alrededor.

**Cruzamiento de cáusticas:** La transición del régimen semi-lineal a uno no lineal ocurre cuando las cáusticas<sup>4</sup> internas se cruzan con las cáusticas externas.

De acuerdo al modelo esférico, la evolución de los vacíos dependerá únicamente de la amplitud inicial de la fluctuación de densidad, y no de algún radio inicial  $R$  o cantidad de masa encerrada  $M$ . Es por esto que estas magnitudes se pueden tratar sin inconvenientes a través de su relación con el volumen de la región.

Los vacíos no existen como estructuras independientes, sino que se forman y evolucionan bajo una importante influencia del entorno que los rodea. En consecuencia, su estructura y dinámica resultan más complejas de lo que serían en caso de que se expandiesen aisladamente. De acuerdo a cómo evolucionan e interactúan las estructuras que rodean a los vacíos, estos últimos se pueden clasificar. Por un lado, siguiendo el modelo evolutivo *void-in-void* (vacío-en-vacío), están los vacíos cuyo entorno posee una densidad integrada menor a la de la media global del Universo y se expanden como regiones subdensas. De acuerdo a este modelo, vacíos grandes pueden surgir de la fusión de vacíos pequeños en épocas más tempranas. Por otro lado, si los vacíos siguen el modelo evolutivo *void-in-cloud* (vacío-en-nube), entonces lo que se tiene es que los vacíos se

---

<sup>4</sup>Con el término cáustica se hace referencia a los isocontornos de densidad de los vacíos.

encuentran rodeados de una cáscara sobredensa que se contrae paulatinamente por la interacción gravitatoria, para eventualmente colapsar y desaparecer. Este suele ser el caso de los vacíos pequeños [Sheth and van de Weygaert, 2004].

Dos factores condicionan las características estadísticas de los vacíos: los trazadores de materia que se usan al mapear la estructura a gran escala (galaxias, halos de materia oscura, etc) y el método de identificación usado a partir de la distribución espacial de los trazadores [Correa, 2021].

Para el estudio de vacíos predominan dos herramientas estadísticas: la Función de Tamaño de Vacíos (VSF) y la Función de Correlación Cruzada Vacío-Galaxia. La primera de estas, en la que se hará hincapié en esta tesina, describe la abundancia de vacíos en función de su tamaño, y se modela como una combinación de la teoría de excursión probabilística junto con la expansión esférica de las subdensidades de materia que derivan de la teoría cosmológica de perturbaciones. Es importante destacar que ambas herramientas pueden ser utilizadas para diseñar test cosmológicos, dado que es posible modelar su comportamiento teórico para luego contrastarlo con las observaciones reales. De esta manera se pueden hacer estudios sobre diversos parámetros cosmológicos, a través de los cuales se accede a información sobre, por ejemplo, la materia y la energía oscura. Esto convierte a los vacíos en buenos laboratorios cosmológicos [Correa, 2021].

Para poder hacer un análisis empleando la VSF, es necesario especificar el modelo de vacío a utilizar, y como se mencionó anteriormente, en este trabajo se usa el modelo esférico. De esta manera, los vacíos se definen como esferas subdensas de radio  $R_v$ , haciendo que la VSF sea entonces una distribución de radios. A través de esto se puede acceder al observable más importante para medir la abundancia, que es el volumen.

La teoría de excursión probabilística usada para modelar la VSF deriva del modelo usado para modelar la formación de cúmulos, en donde es posible obtener la fracción de irregularidades capaces de superar un cierto nivel crítico  $\Delta_c$  de densidad de materia

que resulta en la formación de sobredensidades, dentro de un intervalo de escalas de densidad  $d \ln \sigma$ . Dicho de otra forma, la teoría permite obtener cuántos cúmulos se van a formar a partir de las irregularidades en densidad que se tiene en edades tempranas del Universo. Formalmente, las trayectorias que cruzan el nivel crítico son

$$f_{\ln \sigma}(\sigma) = \frac{df}{d \ln \sigma} = \sqrt{\frac{2}{\pi}} \frac{\Delta_c}{\sigma} \exp \left[ -\frac{\Delta_c^2}{2\sigma^2} \right] \quad (3.4.1)$$

Siendo una trayectoria una secuencia de sobredensidades determinadas a partir de incrementos sucesivos en una escala de suavizado  $\mathcal{R}$ , escala que a su vez es común relacionar con la correspondiente varianza del campo de densidad lineal [Correa, 2021]

$$\sigma^2(\mathcal{R}) = \int \frac{k^2}{2\pi^2} P_m(k) |W(k, \mathcal{R})|^2 dk \quad (3.4.2)$$

Donde  $|W(k, \mathcal{R})|$  es una función filtro y  $P_m(k)$  es el espectro de potencias de la materia. Este factor es de gran importancia, ya que contiene toda la información que caracteriza la manera en que se generan las estructuras en el Universo, y hace explícita la dependencia cosmológica del cálculo de la abundancia. Además, como la masa y el número de partículas se conservan al colapsar, la función de masas se puede expresar a partir de la densidad numérica comóvil de halos [Correa, 2021]

$$\frac{dn}{d \ln M} = \frac{\rho_m}{M} f_{\ln \sigma}(\sigma) \frac{d \ln \sigma^{-1}}{d \ln M} \quad (3.4.3)$$

Es posible extender el modelo para trabajar con regiones inicialmente subdensas, que posteriormente resultarán en vacíos al evolucionar el Universo. De manera análoga, se define un valor crítico  $\Delta_v$  que deberá ser superado para que se considere que la región puede evolucionar en un vacío, teniendo en cuenta el modelo evolutivo *void-in-void*. En particular, se define  $\Delta_v$  en la instancia en donde las cáusticas se cruzan. Para un universo EdS (esto es, un modelo previo al  $\Lambda$ CDM con  $\Lambda = 0$ ), se tiene  $\Delta_v = -2,71$ , y se ha demostrado que este valor también es aplicable a universos  $\Lambda$ CDM planos [Correa, 2021, Jennings et al., 2013].

Se debe destacar que no se puede asumir que los vacíos se expanden esféricamente en el aislamiento, debido a que la evolución de una región subdensa no es la misma que

en el caso del colapso de materia, por lo cual las predicciones no serán del todo precisas. La evolución de un vacío es más compleja, y por esto se debe tener en cuenta también el modelo evolutivo *void-in-cloud*. Para esto, Sheth & van de Weygaert proponen que una segunda barrera sobredensa  $\Delta_c$  sea incorporada. De esta manera, se contarán los vacíos que hayan cruzado una barrera  $\Delta_v$  de subdensidad, en tanto no hayan cruzado la barrera  $\Delta_c$  para una escala espacial mayor. La densidad numérica de los vacíos en el régimen lineal quedará descrita como [Sheth and van de Weygaert, 2004]

$$A_V = \frac{dn_v}{d \ln R_v} = \frac{f_{\ln \sigma}(\sigma)}{V(R_v)} \frac{d \ln \sigma^{-1}}{d \ln R_v} \quad (3.4.4)$$

Donde el lado derecho de la igualdad constituye el modelo teórico para la abundancia. El volumen de un vacío de radio  $R_v$  se representa como  $V(R_v) = (4\pi R_v^3)/3$ , y además

$$f_{\ln \sigma}(\sigma) = \sum_{j=1}^{\infty} \exp \left[ -\frac{(j\pi x)^2}{2} \right] j\pi x^2 \sin(j\pi \mathcal{D}) \quad (3.4.5)$$

Y se definen a  $\mathcal{D}$  (*void-and-cloud parameter*) y  $x$  como

$$\mathcal{D} = \frac{|\Delta_v|}{\Delta_c + |\Delta_v|} \quad (3.4.6)$$

$$x = \frac{\mathcal{D}\sigma}{\Delta_v} \quad (3.4.7)$$

La ec. (3.4.4) se denomina modelo lineal.

En el modelo de Sheth & van de Weygaert, la hipótesis principal es que se conserva la densidad numérica comóvil de los vacíos durante su evolución, alterando así únicamente sus tamaños. Sin embargo, las predicciones resultantes de este modelo carecen de sentido físico. Por este motivo, Jennings, Lu & Hu (2013) proponen un modelo en donde en lugar de asumir la conservación de la densidad numérica, asume la conservación de la fracción de volumen comóvil [Jennings et al., 2013]. De esta manera, se contempla la fusión de los vacíos con sus vecinos durante su expansión, conservando su volumen y no su número. La ec. (3.4.4), que daba la abundancia de vacíos en el modelo de Sheth & van de Weygaert, ahora es

$$A_V = \frac{dn_v}{d \ln R_v} = \frac{f_{\ln \sigma}(\sigma)}{V(R_v)} \frac{d \ln \sigma^{-1}}{d \ln R_v^L} \frac{d \ln R_v^L}{d \ln R_v} \quad (3.4.8)$$

Aquí,  $L$  hace referencia al radio lineal. La ecuación se conoce como modelo de conservación de volumen, y sirve para obtener la abundancia de vacíos identificados en materia. Es importante hacer énfasis en esta característica de la ec. (3.4.8), pues esta formulación resulta inservible cuando los vacíos están identificados en halos de galaxias [Correa, 2021].

El principal obstáculo para poder utilizar los vacíos y la VSF como test de energía oscura es que, al día de la fecha, no existe detección directa de materia oscura, por lo tanto identificar la verdadera VSF con este trazador es imposible. A lo que sí se tiene acceso es a la distribución de galaxias, a través de los relevamientos de redshift de gran tamaño. Se sabe que las galaxias trazan las regiones de alta densidad de materia oscura (ya que estas se forman en los interiores de los halos de materia oscura), por lo que sería posible usarlas para inferir la distribución subyacente de la misma. El problema es que no ocurre lo mismo en las regiones de baja densidad, pues estas no son trazadas de la misma manera. Como se mencionó respecto a la ec. (3.4.8), los modelos teóricos para la abundancia de vacíos sólo funcionan usando la distribución de materia. La evidencia sugiere que las galaxias presentan un sesgo o *bias* a la hora de trazar las regiones de materia oscura, y estudios sugieren que un bias lineal no es suficiente para inferir una correcta distribución de materia oscura. Esto hace necesario agregar complejidad al modelo [Pollina et al., 2017, Sheth and Lemson, 1999].

Existen varias contribuciones al estudio de la conexión entre vacíos identificados en galaxias y materia oscura desde el punto de vista teórico [Sutter et al., 2014]. Se encuentran entre ellas: ajustes a los parámetros de la teoría de excursión probabilística [Furlanetto and Piran, 2006, Jennings et al., 2013] propuesto por Sheth & van de Weygaert [Sheth and van de Weygaert, 2004] para explicar el fenómeno de vacíos encontrados en poblaciones de galaxias; la comparación de vacíos en el espacio real y en el espacio de redshift [Ryden and Melott, 1996]; el impacto del sesgo en diferentes poblaciones de galaxias [Tinker and Conroy, 2009]; evaluación de los efectos de distintos niveles de dispersión en la reconstrucción de vacíos. [Schmidt et al., 2001, Colberg et al., 2005].

Cuando el modelo  $\Lambda$ CDM es sometido a pruebas, una gran cantidad de factores debe ser tenida en cuenta, tales como la abundancia de elementos livianos, el ritmo al que las galaxias se alejan entre sí con su correspondiente aceleración positiva, y las pequeñas y detalladas irregularidades en el CMB. Thompson (2020) plantea que es pertinente y necesario añadir a esta lista las características observadas de los vacíos cósmicos, para que se vuelvan partes integrales de un modelo optimizado de  $\Lambda$ CDM [Thompson, 2020].

### 3.5. La informática y la cosmología

El rápido crecimiento en los volúmenes de datos y poder computacional está empujando a la ciencia a un nuevo paradigma basado en la integración de la teoría, la experimentación y la simulación, con un énfasis en el análisis de grandes cantidades de datos para descubrir patrones y generar nuevos conocimientos. Este contexto ha recibido el nombre de “Cuarto paradigma”, donde establece que cada vez en mayor medida los avances científicos serán potenciados por capacidades computacionales avanzadas que ayudarán a los investigadores a manipular y explorar grandes conjuntos de datos [Hey et al., 2009]. Este paradigma es también llamado Ciencia de Datos (DS), y ha generado tanto entusiasmo que incluso se afirma que las cantidades cada vez mayores de datos hacen posible construir modelos accionables sin utilizar teorías científicas [Karpatne et al., 2016].

Dentro de las tecnologías emergentes aprovechadas, la Inteligencia Artificial (IA) permite a las computadoras realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de patrones, el aprendizaje y la toma de decisiones. En particular, la subrama del Machine Learning (ML) es la que se enfoca en el desarrollo de algoritmos y modelos estadísticos que permiten a las computadoras aprender y mejorar su rendimiento en una tarea específica sin ser programadas explícitamente. Esto está

permitiendo a los investigadores descubrir patrones ocultos, realizar predicciones y tomar decisiones basadas en datos, lo que acelera el proceso de descubrimiento científico y abre nuevas posibilidades para abordar problemas complejos.

Uno de los grandes beneficiarios de estos desarrollos es, sin duda, el campo de la astronomía, aprovechándolos para procesar relevamientos y simulaciones de gran tamaño y calidad [Mickaelian, 2017]. Así los trabajos que otrora se realizaban de manera manual han sido reemplazados por metodologías automáticas y computacionales, generando un advenimiento de nuevas subdisciplinas como la astro-informática y la astro-estadística [Cabral, 2019].

En este trabajo resultan de particular interés las simulaciones cosmológicas, las cuales se utilizan para estudiar la formación y evolución de estructuras a gran escala en el universo, como galaxias, cúmulos de galaxias y la distribución de la materia oscura. Las simulaciones modernas se benefician enormemente de las capacidades computacionales avanzadas y generan un gran volúmenes de datos [Sun et al., 2018].

### 3.5.1. Simulaciones cosmológicas

Las simulaciones cosmológicas son poderosas herramientas que acompañan los astrónomos desde hace alrededor de 50 años, en el sentido estricto del término cosmológico, ya que desde antes ya existían simulaciones de menor tamaño. Se trata de simulaciones numéricas de  $n$ -cuerpos que permiten acompañar e interpretar las observaciones cosmológicas, e incluso pueden conducir a las mismas. Tienen una gran cantidad de usos, incluyendo la calibración de métodos usados para medir parámetros cosmológicos, proveer información en el área de *clustering* gravitacional no lineal y turbulencia hidrodinámica, mejorar el entendimiento de las limitaciones de los modelos físicos de formación de galaxias, entre tantos otros [Bertschinger, 1998].

La idea detrás de los modelos de formación de estructura es reducir los sistemas

cosmológicos a un problema de valores iniciales. Una vez dadas las condiciones iniciales (composición de materia, radiación, constantes cosmológicas, etc), se computa la evolución del sistema siguiendo leyes físicas ya establecidas, desde el Big Bang hasta el día de hoy. La ventaja que presentan estas herramientas es su nivel de abstracción respecto al verdadero Universo, que es un sistema excesivamente complejo, al permitir enfocarse en ciertos aspectos de interés para los investigadores [Bertschinger, 1998].

En el presente trabajo se tiene particular interés en utilizar ML para intentar hallar el vínculo entre la VSF computadas para vacíos identificados en materia y vacíos identificados en halos, de manera que se puede evitar el análisis del problema desde el punto de vista más teórico.

Finalmente, con este proyecto no se pretende desestimar los esfuerzos de índole más formal, pues en caso de que este método sea exitoso y en su desarrollo se pueda continuar agregando complejidad al modelo, este será de mayor utilidad como complemento de los modelos teóricos.

# Capítulo 4

## Machine learning

Hacer que las computadoras piensen como los seres humanos para realizar tareas complejas y repetitivas de manera rápida y eficiente es una inquietud que se remonta a los comienzos de la computación [Ashby, 1949] y ha recibido el nombre de IA. Diferentes subdisciplinas han surgido para encarar este problema, siendo la que incumbe a este trabajo el ML.

El ML puede entenderse de varias formas, y una de ellas es verlo como un problema de programación en donde se obtiene como resultado un programa determinista, que significa que siempre producirá los mismos resultados dados los mismos inputs. Alternativamente, otro enfoque del ML se basa en dejar que los programas aprendan y mejoren a partir de los datos que se le entregan, sin diseñarlos explícitamente. Esto se logra acondicionando los datos, seleccionando un meta-algoritmo de aprendizaje automático y entrenando un modelo/programa utilizando esos datos. El modelo/programa aprende patrones y relaciones en los mismos, con lo que luego puede hacer predicciones o tomar decisiones basadas en esos patrones.

Esta forma de “programar” es adecuada para problemas donde las reglas no son claras o son difíciles de codificar, y donde hay una gran cantidad de datos disponibles.

En contraste, un programa tradicional funciona mejor para problemas bien definidos con requisitos claros y soluciones especificados [Mitchell, 1997].

Así ML puede definirse de manera formal como:

*Se dice que un programa de computadora aprende de una experiencia E con respecto a alguna clase de tarea T y métrica de desempeño P, si su desempeño en las tareas T, medidas usando P, mejora con la experiencia E [Mitchell, 1997].*

## 4.1. Muchas formas de aprender

Así como el IA tiene muchas subramas, el ML también posee varias formas de “aprender”, las cuales se pueden englobar en tres grandes categorías: aprendizaje supervisado, aprendizaje por refuerzo y aprendizaje no supervisado.

**Aprendizaje no-supervisado** En este caso el modelo resultante se construye a partir de datos no etiquetados, es decir, sin conocer previamente las categorías o valores esperados. Busca descubrir patrones, estructuras y otras relaciones en los datos. Ejemplos de este tipo de modelos incluyen clustering (agrupamiento) o reducción de dimensionalidad [Raschka and Mirjalili, 2017].

**Aprendizaje por refuerzo** Se basa en la interacción entre un agente y su entorno. El agente aprende a tomar acciones que maximicen una recompensa largo plazo, a medida que explora el entorno mediante prueba y error, recibiendo recompensas o penalizaciones para ajustar el comportamiento [Raschka and Mirjalili, 2017].

**Aprendizaje supervisado** En el aprendizaje supervisado, el algoritmo aprende a partir de datos etiquetados, donde se proporcionan tanto las entradas como las salidas deseadas. Este es el tipo de aprendizaje de interés para este proyecto, que será desarrollado a continuación.

#### 4.1.1. El aprendizaje supervisado

El aprendizaje supervisado tiene como meta aprender un modelo a partir de datos con etiquetas o valores que permita hacer predicciones sobre observaciones futuras o nunca antes vistas. El término “supervisado” se refiere a que se conoce la correspondencia entre la entrada y la salida a la hora de entrenar el algoritmo [Raschka and Mirjalili, 2017].

En términos matemáticos, en el aprendizaje supervisado la meta es obtener una relación entre las variables de entrada  $X$  y las de salida  $y$ , dado un conjunto de pares de datos de entrada y salida  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ . En este caso  $\mathcal{D}$  es el conjunto de entrenamiento, y  $N$  es el número de datos en el mismo. Los elementos  $x_i$  que componen cada uno de esos  $N$  datos son denominados atributos, y pueden variar su naturaleza desde vectores a estructuras más complejas, como imágenes, series temporales, etc. De manera análoga, la salida  $y_i$  también puede tomar las mismas formas, aunque la mayoría de los métodos asumen que es una variable categórica o nominal [Murphy, 2012].

A su vez, los métodos de aprendizaje supervisado se pueden dividir de acuerdo al tipo de problema que resuelven, siendo estos clasificación o regresión: el primero de estos corresponde a las situaciones donde  $y_i$  es una variable categórica, es decir, el algoritmo aprende a asignar una etiqueta cuando se le presentan datos nunca antes vistos, mientras que el segundo caso es cuando el algoritmo asigna un valor perteneciente a un continuo [Murphy, 2012].

En este trabajo se hace hincapié en los modelos de regresión, cuyo objetivo es formular una función que represente datos observados de manera precisa. Esto es, que el algoritmo obtenga un mapeo o relación de la entrada  $X$ , donde los  $x_i$  son valores numéricos, a la salida  $y$ , donde  $y$  es una variable numérica continua. Una vez obtenida la función, esta puede ser utilizada para realizar predicciones sobre datos no observados [Wang, 2023]. Los algoritmos utilizados serán descriptos en la siguiente sección.

## 4.2. Algoritmos de regresión

Como se mencionó anteriormente, este trabajo está centrado en buscar la correlaciones entre las VSF de vacíos identificados en halos de materia oscura y en partículas, con el objetivo de efectuar predicciones sobre estos últimos al darle datos al algoritmo sobre los primeros. Por estos motivos se utilizarán regresiones de distintos tipos para encontrar esa relación.

Específicamente se trabajarán los algoritmos de Regresión Lineal (RL), Gaussian Process Regressor (GPR) y Random Forest Regressor (RF).

### 4.2.1. Regresión Lineal

La regresión lineal por mínimos cuadrados es el método más elemental de los utilizados en el presente trabajo. Funciona bajo el supuesto de que la relación que une a los datos de entrada con su objetivo es de forma lineal, como su nombre indica

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i \quad (4.2.1)$$

Aquí, los valores de  $\beta_0$  y  $\beta_1$  son los parámetros de la recta que ajusta a los datos, y  $\epsilon_i$  es la desviación que cada uno de esos datos posee respecto a la recta. Todos estos parámetros son inicialmente desconocidos, siendo el objetivo ajustar  $\beta_0$  y  $\beta_1$ .

Si a la desviación vertical del punto  $(x_i, y_i)$  de la línea  $y = b_0 + b_1x$  es

$$\text{altura del punto} - \text{altura de línea} = y_i - (b_0 + b_1x_i) \quad (4.2.2)$$

Entonces la suma del cuadrado de las desviaciones de los puntos  $(x_1, y_1), \dots, (x_n, y_n)$  es

$$f(b_0, b_1) = \sum_{i=1}^n [y_i - (b_0 + b_1x_i)]^2 \quad (4.2.3)$$

Teniendo esto en cuenta, los valores  $\hat{\beta}_0$  y  $\hat{\beta}_1$ , llamados estimadores de mínimos cuadrados, son los valores que minimizan la función  $f(b_0, b_1)$ . Esto es,  $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1) \forall (b_0, b_1)$ , y la recta que mejor ajusta a los datos es  $y = \hat{\beta}_0 + \hat{\beta}_1 x$  [Devore, 2011].

Para el proyecto, se usa un modelo de regresión n-dimensional, donde la variación respecto al modelo que se acaba de presentar es la dimensión de la variable de entrada.

#### 4.2.2. Gaussian Process Regressor

Los GPR son modelos probabilísticos de aprendizaje automático supervisado, que realizan predicciones al incorporar conocimiento previo sobre la naturaleza de los datos y entregan una medida de la incertezza en sus predicciones [Wang, 2023]. Formalmente se los puede definir como una colección de variables aleatorias, donde cualquier número finito de las mismas poseen una distribución gaussiana multivariante [Rasmussen and Williams, 2006].

Los procesos gaussianos se encuentran completamente definidos si se especifica su función media  $m(x)$  y su covarianza  $k(x, x')$  para un proceso real  $f(x)$

$$m(x) = \mathbb{E}[f(x)] \quad (4.2.4)$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \quad (4.2.5)$$

Y el proceso gaussiano se escribe como

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \quad (4.2.6)$$

Por costumbre, se toma que  $m(x)$  es una función nula.

Si se considera un conjunto de datos reales, una cantidad infinita de potenciales funciones pueden ajustarlos. En el método, los procesos gaussianos realizan una regresión al definir una distribución probabilística sobre este número infinito de funciones para obtener un mejor estimado de aquellas que ajustan las observaciones [Wang, 2023].

Se dice que una variable  $X$  es gaussiana o que posee una distribución normal con una media  $\mu$  y una varianza  $\sigma^2$  si su función de densidad de probabilidad es

$$P_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (4.2.7)$$

En esta expresión,  $X$  representa a alguna variable y  $x$  es el argumento real.

Cuando un sistema es descrito por más de una variable,  $(X_1, X_2, \dots, X_D)$ , que estén correlacionadas, es necesario utilizar una Distribución Normal Multivariable (MVN) para modelarlo como un solo modelo gaussiano. La función de densidad de probabilidad de una MVN  $D$ -dimensional se define como

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right] \quad (4.2.8)$$

donde  $D$  es el número de dimensiones,  $x$  es la variable,  $\mu = \mathbb{E}[x] \in \mathbb{R}^D$  es el vector medio y  $\Sigma = cov[x]$  es la matriz covarianza de tamaño  $D \times D$ .  $\Sigma$  es una matriz simétrica que almacena información sobre la covarianza entre cada par de variables, es decir,  $\Sigma_{ij} = cov(y_i, y_j)$  [Wang, 2023].

La función de regresión modelada por una gaussiana es entonces

$$P(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mu, \mathbf{K}) \quad (4.2.9)$$

donde  $\mathbf{X} = [x_1, \dots, x_n]$  son los datos observados,  $\mathbf{f} = [f(x_1), \dots, f(x_n)]$ ,  $\mu = [m(x_1), \dots, m(x_n)]$  siendo  $m$  la función promedio y  $K_{ij} = k(x_i, x_j)$  son componentes del kernel. El modelo de procesos gaussianos es una distribución sobre funciones cuya forma está determinada por  $\mathbf{K}$ . Si el kernel determina que dos puntos  $x_i$  y  $x_j$  son similares, entonces las salidas  $f(x_i)$  y  $f(x_j)$  también se espera que sean similares [Wang, 2023].

Una vez creada la MVN, se le da a conocer los datos reales. Lo que esto hace es imponer condiciones que reducen el número de posibles funciones que modelicen el sistema, acorde a su probabilidad de ajustar los datos. El algoritmo evalúa la probabilidad de obtener datos nuevos, teniendo en cuenta la manera en que el kernel dice

que los datos provistos se relacionan entre sí. Es importante realizar esta distinción: el kernel no establece la forma que la función tiene, sólo establece la manera en que los datos se conectan.

Luego, sobre esta serie de funciones que ajustan los datos, se realiza un promedio que es el que efectivamente se usa para realizar predicciones para datos no vistos [Wang, 2023].

De esta manera, supóngase que se tiene un conjunto de datos de entrenamiento  $\mathcal{D} = \{(\mathbf{x}_i, f_i), i = 1 : N\}$ , donde  $f_i = f(\mathbf{x}_i)$  es el valor de la función a modelar en  $\mathbf{x}_i$ . Si ahora se introduce un conjunto de datos de prueba  $\mathbf{X}_*$ , de tamaño  $N_* \times D$ , lo que se busca predecir es la función  $\mathbf{f}_*$ .

A la hora de implementar el método, el GPR debe cumplir dos funciones: primero, para valores  $\mathbf{x} \in \mathcal{D}$  deberá predecir exactamente  $f(\mathbf{x})$ , es decir, interpolar los datos de entrenamiento. En segundo lugar, debe efectuar la predicción. Por definición, el proceso gaussiano tiene la forma [Murphy, 2012]

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right) \quad (4.2.10)$$

donde  $\mathbf{K} = k(\mathbf{X}, \mathbf{X})$  tiene tamaño  $N \times N$ ,  $\mathbf{K}_* = k(\mathbf{X}, \mathbf{X}_*)$  tiene tamaño  $N \times N_*$  y  $\mathbf{K}_{**} = k(\mathbf{X}_*, \mathbf{X}_*)$  tiene tamaño  $N_* \times N_*$ . Teniendo esto en cuenta, la distribución posterior toma la forma

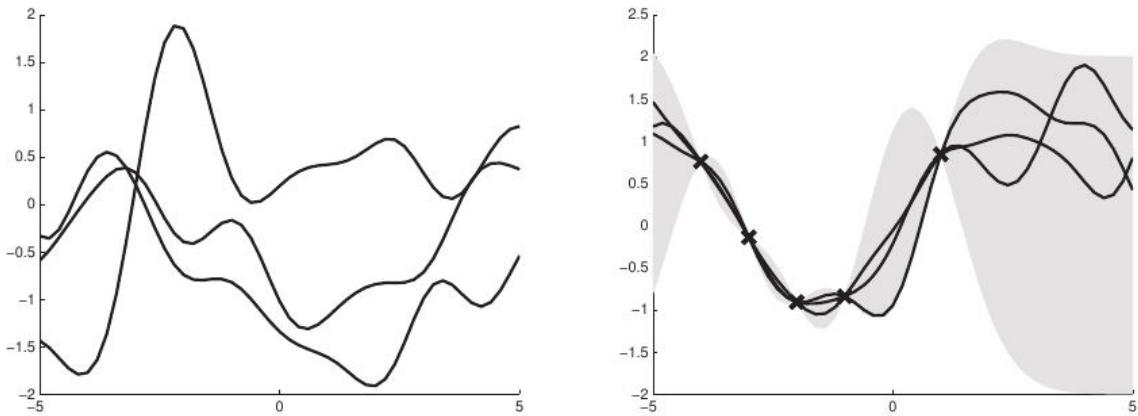
$$P(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(\mathbf{f}_* | \mu_*, \Sigma_*) \quad (4.2.11)$$

donde  $\mu_*$  y  $\Sigma_*$  se definen como

$$\mu_* = \mu(\mathbf{X}_*) + \mathbf{K}_*^T \mathbf{K}^{-1} (\mathbf{f} - \mu(\mathbf{X})) \quad (4.2.12)$$

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_* \quad (4.2.13)$$

Este proceso queda ilustrado en la Fig. 4.1.



**Figura 4.1:** Proceso genérico de un GPR. A la izquierda se aprecian funciones aleatorias previas a la introducción de los datos. A la derecha algunas de las funciones luego de condicionar con 5 observaciones [Murphy, 2012].

El GPR es un modelo de regresión no paramétrico, y contrario a lo que su nombre indica, posee un número infinito de parámetros, derivados directamente de la información provista por los datos. Quizás resulte útil contrastarlo con la regresión lineal, la cual es efectivamente un modelo paramétrico: si se ajustan datos por una curva  $y = \theta_0 + \theta_1 x$ , entonces se presume que toda la información está encapsulada en un número finito de parámetros, que en este caso son  $\theta_0$  y  $\theta_1$  [Wang, 2023].

Entre las ventajas que presenta este método se cuentan [Scikit-Learn, 2024b]:

- La capacidad de interpolar los datos observados (para kernels regulares).
- El carácter probabilístico de las predicciones, que permite obtener intervalos de confianza empíricos acorde a los cuales se puede decidir si se debe adaptar la curva predicha en algún intervalo de interés.
- La variedad a la hora de elegir kernels. Existen kernels ya predefinidos, pero también está la posibilidad de definir un kernel desde cero.

En este trabajo se entrenaron dos GPR utilizando la implementación Scikit-Learn

[Scikit-Learn, 2024b] de manera independiente: uno para hacer un ajuste del radio de los vacíos y otro para las abundancias, ajustando únicamente el hiperparámetro  $\alpha$  para el algoritmo:  $\alpha$  se agrega a la diagonal de la matriz del kernel para prevenir problemas al momento de la ejecución, garantizando que los valores calculados formen una matriz definida positiva. Además, se utilizó un kernel compuesto con los siguientes términos:

- Un kernel constante, que puede escalar la magnitud de algún otro factor o modificar el valor medio del kernel, dependiendo de si se usa multiplicando o sumando respectivamente.
- Un Radial Basis Function (RBF), caracterizado por un parámetro de escala  $l > 0$ , que establece qué tan similares son los datos. Informalmente se puede pensar como la distancia que se debe recorrer antes de que la función cambie significativamente [Rasmussen and Williams, 2006]. El RBF se comporta de la siguiente manera

$$k(x_i, x_j) = \exp \left[ -\frac{d(x_i, x_j)^2}{2l^2} \right] \quad (4.2.14)$$

donde  $d(x_i, x_j)$  es una distancia euclíadiana.

- Un kernel Rational Quadratic (RQ), caracterizado por un parámetro de escala  $l > 0$  y un parámetro de escala de mezcla  $\alpha_k > 0$ , donde el subíndice  $k$  denota la pertenencia al kernel. El comportamiento que sigue es de la forma

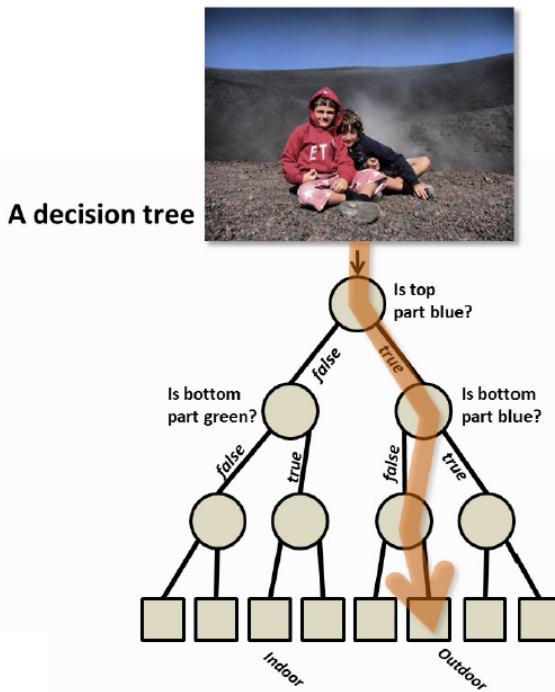
$$k(x_i, x_j) = \exp \left[ 1 + \frac{d(x_i, x_j)^2}{2\alpha l^2} \right]^{-\alpha} \quad (4.2.15)$$

donde  $d(x_i, x_j)$  es una distancia euclíadiana.

### 4.2.3. Random Forest Regressor

Este algoritmo tiene como bloque fundacional a los árboles de decisiones, y si bien su popularidad surge de su uso como un método de clasificación, su utilidad se puede extender a varios tipos de problemas, siendo este caso una regresión.

Un árbol de decisión es un conjunto de preguntas organizadas de manera jerárquica, que pueden ser representadas gráficamente como un árbol. Así, para una dada entrada, el árbol de decisión estima una propiedad desconocida del objeto haciendo sucesivas preguntas sobre las propiedades que sí conoce del mismo, siendo la pregunta a realizar dependiente de la respuesta de la pregunta inmediata anterior. Esta relación entre preguntas y respuestas queda determinada gráficamente como un camino entre la primera pregunta (o *raíz*) y la conclusión que se encuentra en un nodo denominado *hoja*. Un ejemplo de esto se puede ver en la Fig. 4.2 [Criminisi et al., 2011].



**Figura 4.2:** Esquema de un árbol de decisión aplicado al análisis de ilustraciones [Criminisi et al., 2011].

Cuantas más preguntas tiene un árbol, mayor confianza hay en la respuesta. Cada una de estas preguntas está asociada un nodo del árbol.

En la teoría de los árboles de decisión, un dato queda denotado por  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathcal{D}^d$ , donde cada uno de los  $x_i$  representa un atributo (en inglés, *feature*), y  $\mathcal{D}^d$  es el espacio  $d$ -dimensional de los atributos. La dimensión  $d$  depende del

sistema que se esté analizando, y puede llegar incluso a ser infinito. En la práctica, sin embargo, no siempre es posible ni necesario usar las  $d$  dimensiones de  $\mathbf{x}$ , sino que se puede seleccionar un subconjunto  $\mathcal{D}^{d'}$  con  $d' \ll d$  en donde operar con mayor eficiencia [Criminisi et al., 2011].

Como se dijo antes, los nodos del árbol se organizan acorde a una cierta jerarquía, y cada uno de ellos tiene asociada una función de prueba que entrega un valor binario y se encarga de tomar la decisión

$$h(\mathbf{x}, \theta_j) : \mathcal{D} \times \mathcal{T} \rightarrow \{0, 1\} \quad (4.2.16)$$

Donde 0 y 1 se pueden interpretar como *falso* o *verdadero* respectivamente, y  $\theta_j \in \mathcal{T}$  denota los parámetros de la función de prueba en el  $j$ -ésimo nodo. El dato  $\mathbf{x}$  que llega a dicho nodo, luego es enviado al nodo hijo, a la izquierda o a la derecha acorde al resultado [Criminisi et al., 2011].

Al momento de crear un nodo, debe existir un criterio estadístico que determine la mejor manera de dividir los datos. Para el caso de los RF aplicados a problemas de regresión, lo más común es usar el MSE para determinar la impureza de un conjunto (esto es, qué tan heterogéneo es un conjunto de datos)

$$MSE(t) = \frac{1}{N_t} \sum_i \in \mathcal{D} (y_i - y_i^t)^2 \quad (4.2.17)$$

donde  $t$  es el nodo donde se lleva a cabo la operación,  $N_t$  son los datos que llegan a dicho nodo,  $y_i$  es el valor real e  $y_i^t$  es el valor de la predicción.

A la hora de definir un nodo en el árbol, el algoritmo evalúa distintos puntos en los que se pueden dividir los datos, y evalúa la disminución de la varianza en la variable objetivo provista por cada uno de ellos. La división que presente la mayor disminución (es decir, el menor valor para el MSE) será el más relevante para dividir los datos, por lo cual será el elegido para operar en el nodo. Este proceso se lleva a cabo para el nodo raíz y para todos los nodos subsiguientes, en este caso utilizando sólo los

datos que pertenezcan a estos nodos, excluyendo aquellos que estén más arriba en el árbol. El proceso continúa hasta alguno de dos criterios se cumplan: hasta que todos los atributos hayan sido usados en el camino del árbol o hasta que todos los datos en un nodo posean la misma variable objetivo [Mitchell, 1997].

Habiendo establecido todo esto, la manera en que un RF opera quedará definida al quedar establecidos una serie de hiperparámetros: la función utilizada para evaluar el desempeño de un nodo (`criterion` en la implementación usada), el número de árboles en el bosque (`n_estimators`), el número mínimo de datos que el árbol considera necesario para dividir los datos en un nodo (`min_sample_split`) y el número de features a considerar al determinar la función en un nodo (`max_features`).

#### 4.2.3.1. El problema de sesgo y varianza

Los métodos de ajuste de datos poseen dos características intrínsecas que influyen en la calidad de sus predicciones: el sesgo y la varianza. Estas características impactan directamente en los errores que el método comete.

Por un lado, cuando se habla del sesgo de un método, a lo que se hace referencia es a la diferencia entre el valor promedio de la predicción y el valor correcto que el método está intentando predecir. La varianza, por otro lado, hace referencia a qué tan variable es la predicción de un método para un dado punto.

El RF es un algoritmo que resulta de gran utilidad para reducir el sesgo y la varianza en una predicción. Los árboles de decisión individuales poseen una alta varianza que puede causar una sobreestimación de los datos, pero cuando se los combina en el método de RF, esta tiende a disminuir. Esto se lleva a cabo promediando los resultados de los árboles individuales [Friedman et al., 2009, Cabral, 2019].

#### 4.2.3.2. Ensemble de árboles

El *bagging* (del inglés *Bootstrap aggregating*) es una técnica utilizada para reducir la varianza de un método predictivo, con particular eficacia en procedimientos de alta varianza y bajo sesgo, y funciona realizando un promedio sobre dichos métodos. Los ensambles de árboles son una modificación a esta técnica, que construye una gran colección de árboles no correlacionados y luego efectúa un promedio sobre ellos [Breiman, 2001].

En el proceso de *bagging* los árboles son buenos candidatos debido a que si se construyen con la suficiente profundidad, su nivel de sesgo es bajo. Como un conjunto de árboles dará predicciones con mucho ruido (alta varianza), promediar sus predicciones mejora el resultado. Además, como cada árbol generado en el *bagging* posee una distribución idéntica, el valor esperado de uno solo de ellos es idéntico al del conjunto de ellos. Como consecuencia, el sesgo de un árbol independiente es igual al del conjunto de árboles, por lo que la única manera de mejorar la predicción es reducir la varianza.

El cambio que el RF presenta es reducir la correlación entre los árboles. Esto se logra utilizando una selección aleatoria de variables en el proceso de creación del árbol. Aquí cada árbol se especializa en ciertos patrones de datos, resultando en una mejor predicción al operar en conjunto [Friedman et al., 2009, Cabral, 2019].

### 4.3. Métricas para validación de modelos

Ya definidos los métodos de ML que se utilizarán en este trabajo, surge la necesidad de medir su rendimiento y eficacia. Es así que las métricas de evaluación cumplen el papel de cuantificar la calidad de las predicciones realizadas por los algoritmos y comparar objetivamente diferentes enfoques. La elección de una determinada métrica dependerá de la herramienta estadística que se esté utilizando y el tipo de problema

que se analice. De esta manera, aquellas que sirvan para un algoritmo de clasificación, no podrán ser usadas cuando se efectúa una regresión.

Una de las métricas usadas en este proyecto fue la  $R^2$ , cuyo nombre hace referencia a la manera en que se denota al coeficiente de determinación en estadística. Dicho coeficiente se define como la proporción de variación observada en una variable  $y$  que puede ser explicada por las variables independientes del modelo. Da una medida de la bondad del ajuste, y en consecuencia, de qué tan bien las nuevas observaciones pueden ser predichas por el modelo. Si para la  $i$ -ésima muestra  $\hat{y}_i$  es el valor predicho e  $y_i$  es el valor verdadero para  $n$  muestras, entonces  $R^2$  se puede calcular como

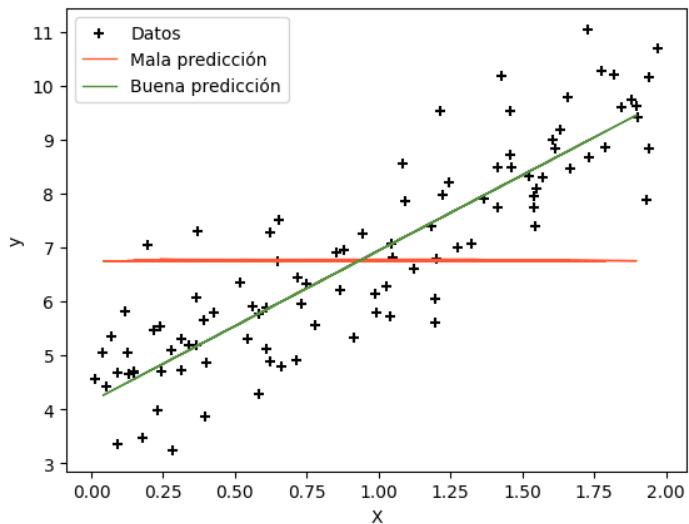
$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (4.3.1)$$

Donde  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  y también  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$ , siendo este último término la suma del cuadrado de los residuos. El resultado de la ec. (4.3.1) es un valor que se encuentra dentro del intervalo  $[0, 1]$ : cuanto mayor es este número, mejor es la calidad de la regresión, entonces  $R^2 = 1$  es una predicción perfecta y  $R^2 = 0$  es imperfecta [Scikit-Learn, 2024a]. Un ejemplo de esto se puede apreciar en la Fig. 4.3, en donde dos RL ajustan una serie de datos con desempeños dispares. El gráfico naranja tiene un valor nulo al evaluarlo con la métrica  $R^2$ , mientras que el gráfico verde ajusta considerablemente bien.

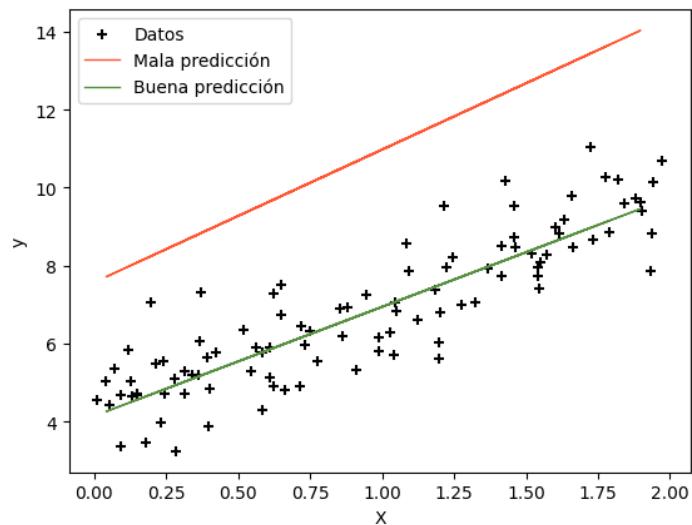
La otra métrica utilizada en este trabajo es el MSE, que se asocia al valor esperado del cuadrado del error. Si  $\hat{y}_i$  es el valor predicho para la  $i$ -ésima muestra e  $y_i$  es el valor real, entonces el MSE estimado sobre  $n$  muestras se define como [Scikit-Learn, 2024a]

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 \quad (4.3.2)$$

A diferencia de la métrica  $R^2$ , el MSE no posee un rango que defina si es un buen o mal valor. Al considerar que es una suma de las diferencias entre el valor real y el valor predicho de una variable, en general se puede decir que cuanto menor sea el valor



**Figura 4.3:** Ejemplo genérico de datos ajustados por una RL. Al evaluar este ejemplo, la gráfica naranja obtiene  $R^2 = 0$ , es decir que la predicción no tiene ninguna similitud con los datos, mientras que la gráfica verde obtiene  $R^2 = 0,81$ .



**Figura 4.4:** Ejemplo genérico de datos ajustados por una RL. Al evaluar este ejemplo, la gráfica naranja obtiene  $MSE = 15,13$ , es decir que la predicción no tiene ninguna similitud con los datos, mientras que la gráfica verde obtiene  $MSE = 0,65$ .

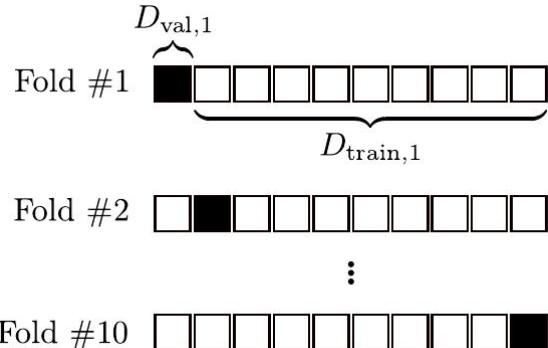
de MSE obtenido (teniendo en cuenta el rango de los datos con los que se trabaja), entonces mejor será el ajuste. Un ejemplo de esto se puede ver en la Fig. 4.4

## 4.4. Validación cruzada

Dado un conjunto  $\mathcal{D}$  de datos con etiquetas,  $x_i, i = 1 \dots n$ , los algoritmos de aprendizaje supervisado operan dividiendo al mismo en dos: un conjunto de entrenamiento  $\mathcal{D}_e$ , que contiene usualmente entre el 70 – 90 % de los datos, y un conjunto de prueba  $\mathcal{D}_p$  que incluye el 30 – 10 % restante. Así, cuando el proceso se encuentra en la fase de entrenamiento, se le aplica un algoritmo  $M$  a  $\mathcal{D}_e$  para que resulte en un modelo estimado  $\hat{f}(x, \mathcal{D}_e)$ , y en la fase siguiente se evalúa el desempeño del modelo acorde a una métrica  $p$  al aplicarlo al conjunto  $\mathcal{D}_p$ .

Sin embargo, un riesgo al que son susceptibles los algoritmos es el sobreajuste de los datos: es una posibilidad que el modelo estimado  $\hat{f}(x, \mathcal{D}_e)$  no sólo refleje la relación entre los datos, sino también el ruido presente en el conjunto. También es posible el caso opuesto, en donde el modelo  $\hat{f}(x, \mathcal{D}_e)$  no esté afectado por el ruido en los datos, pero que tampoco logre establecer la relación entre las variables.

Por este motivo surge la validación cruzada como un método de remuestreo utilizado para evaluar la capacidad de generalización de un modelo, así como también prevenir la sobreestimación de un ajuste. La idea detrás del método es dividir  $\mathcal{D}$  en  $k$  subconjuntos de tamaño aproximadamente igual. Luego, se aplica  $k$  veces el algoritmo  $M$  de la misma manera que se mencionó antes, para pares  $\mathcal{D}_{e,j}$  y  $\mathcal{D}_{p,j}$  con  $j = 1 \dots k$ , donde el conjunto de entrenamiento está compuesto por  $k - 1$  de los subconjuntos, y el conjunto de prueba es el subconjunto restante. Así, el algoritmo  $M$  opera sobre el  $j$ -ésimo conjunto de entrenamiento y evalúa sobre el  $j$ -ésimo conjunto de prueba, como se puede ver en la Fig. 4.5. Es importante destacar que no hay superposición entre los conjuntos de prueba, es decir  $\mathcal{D}_{p,i} \cap \mathcal{D}_{p,j} = \emptyset$ , con  $i \neq j$ . El proceso se repite hasta que



**Figura 4.5:** Representación esquemática del proceso de validación cruzada, para un conjunto  $\mathcal{D}$  dividido en 10 subconjuntos. En la primera etapa, el primer subconjunto sirve como conjunto de prueba. En la segunda etapa, este rol lo ocupa el subconjunto siguiente, y así se continúa hasta llegar al décimo subconjunto [Berrar, 2019].

cada uno de los subconjuntos haya servido como conjunto de prueba, y el promedio de todos los procesos es el modelo final  $f(x, \mathcal{D})$ , y el desempeño del método es el promedio de los  $k$  valores dados por la métrica  $p$  en cada ciclo [Berrar, 2019].

## 4.5. Búsqueda de Hiperparámetros

El ultimo problema sin solución de los que se han planteado en este trabajo es cómo se ajustan los hiperparámetros de los modelos. Si bien existen varias técnicas para hacerlo, la mas popular (y simple) consiste en realizar un producto cartesiano de valores candidatos en una estrategia conocida como Grid Search (GS).

El GS es un algoritmo diseñado para la optimización de hiperparámetros, el cual funciona definiendo un grilla con los valores candidatos de hiperparámetros que se desea optimizar, que constituye el espacio de búsqueda. Sobre esta grilla, el algoritmo explora cada combinación posible de hiperparámetros, para la cual efectúa un ajuste seguido de una evaluación del modelo acorde a una métrica preestablecida. Al finalizar, el GS entrega la combinación de hiperparámetros que haya obtenido la mejor clasificación

[Bergstra and Bengio, 2012].

Por ejemplo si se desea buscar cuál es la mejor combinación de hiperparámetros  $\alpha$  con valores posibles  $\{\alpha_1, \alpha_2, \alpha_3\}$  y  $\beta$  con valores posibles  $\{\beta_1, \beta_2\}$  para un modelo  $M$  dada una métrica  $p$ , GS realizará un proceso exhaustivo de evaluación. Esto implica entrenar y evaluar el modelo  $M$  con todas las combinaciones posibles de  $\alpha$  y  $\beta$ , es decir, se probarán las siguientes configuraciones:  $(\alpha_1, \beta_1), (\alpha_1, \beta_2), (\alpha_2, \beta_1), (\alpha_2, \beta_2), (\alpha_3, \beta_1)$  y  $(\alpha_3, \beta_2)$ . Para cada combinación, se entrenará el modelo y se evaluará su rendimiento utilizando la métrica  $p$ . Finalmente, se seleccionará la combinación de hiperparámetros que obtenga el mejor resultado según la métrica elegida.

# **Capítulo 5**

## **Datos y diseño experimental**

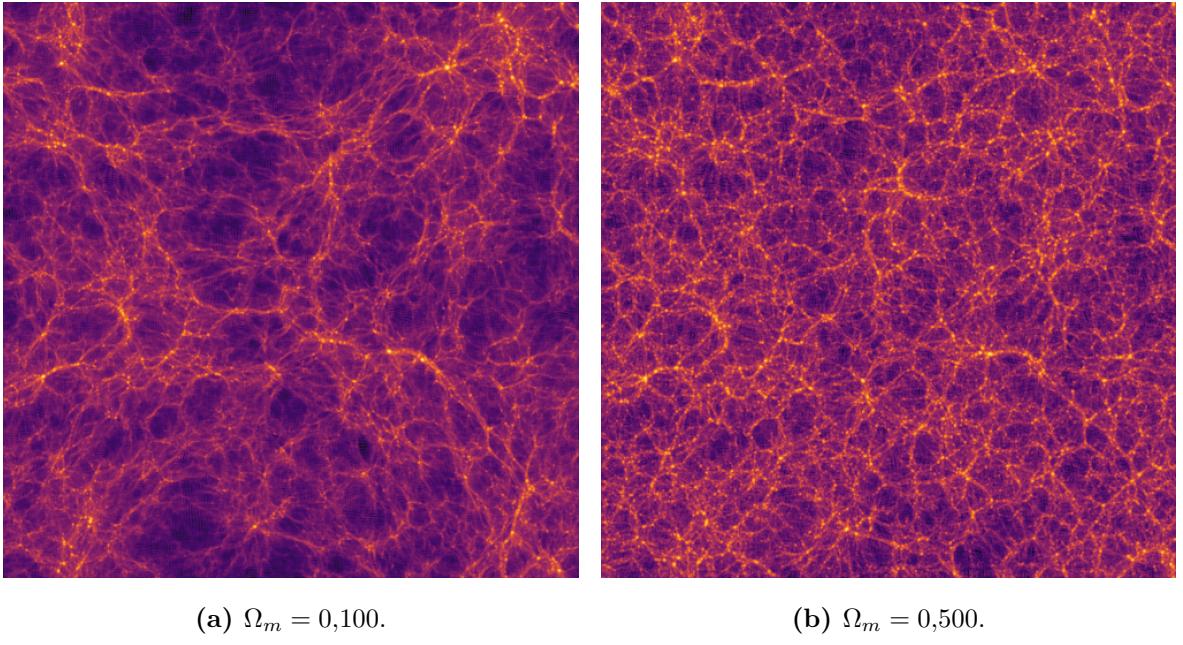
Se describen a continuación la estructura de los conjuntos de datos y los métodos empleados en la tesina.

### **5.1. Conjuntos de datos utilizados**

Al intentar comprobar la hipótesis de que existe una relación entre las curvas de abundancia de vacíos identificados en halos y en materia, hay un obstáculo que aparece desde el primer momento: no existen datos observacionales que permitan obtener las curvas de abundancia para vacíos identificados en el campo de densidad de materia. Esto es lógico, dado que para poder construir dicha curva sería necesario poseer información de la distribución real de materia oscura subyacente.

Por este motivo es que en este proyecto se trabaja sobre simulaciones cosmológicas. La ventaja de esto es que permite acceso a la información de las curvas de abundancia de vacíos usando distintos trazadores, partiendo siempre de la misma distribución de materia subyacente. Es decir, se pueden identificar los vacíos en halos y en partículas,

con la certeza de que corresponden al mismo sistema, y esto permite explorar la relación entre ellos.



**Figura 5.1:** Cortes de dos de las simulaciones empleadas en el proyecto, para los valores extremos de  $\Omega_m$ . Queda en evidencia como a mayores valores de  $\Omega_m$ , la estructura a gran escala se vuelve más densa, consecuencia natural del aumento de materia que habita la simulación.

En el trabajo se emplearon 68 conjuntos de datos, cada uno de ellos conteniendo una lista con los radios de los vacíos identificados con los dos trazadores. Dichos conjuntos fueron identificados en simulaciones cosmológicas de un universo  $\Lambda$ CDM plano realizadas específicamente para este proyecto, que constan de  $N = 512^3$  partículas, en un volumen cúbico de  $500h^{-1}$  Mpc de lado, usando 17 valores distintos para el parámetro de densidad de materia  $\Omega_m$ , en un rango que varía de  $\Omega_m = 0,100$  a  $\Omega_m = 0,500$  con un paso de  $\Omega_m = 0,025$ . A modo de ilustración, en la Fig. 5.1 se pueden ver cortes de dos de las simulaciones, particularmente las de los valores extremos del parámetro de densidad:  $\Omega_m = 0,100$  y  $\Omega_m = 0,500$ . Allí se puede observar fácilmente que cuanto mayor es la cantidad de materia en la simulación, los vacíos poseen un radio menor.

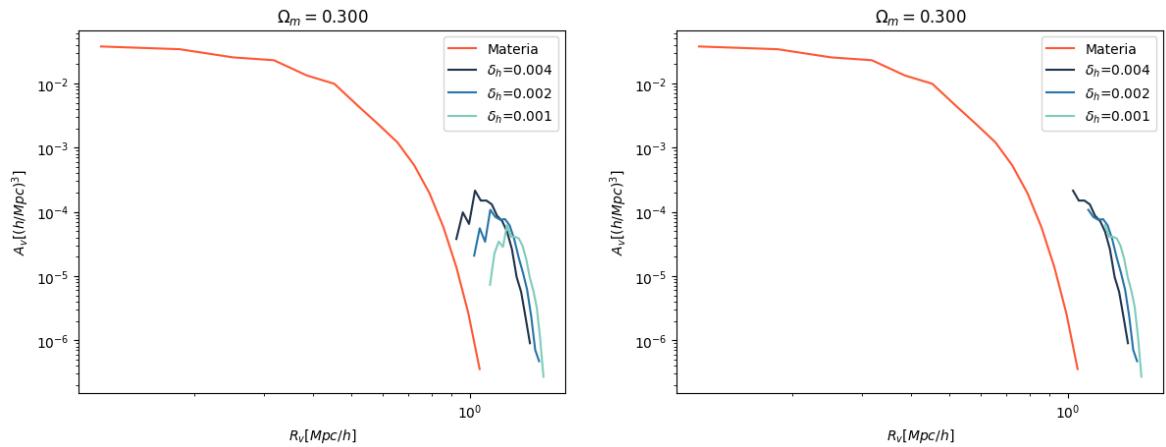
A la hora de identificar los vacíos se emplearon dos trazadores distintos: partículas y halos de materia oscura, tomando 3 valores distintos para la densidad de halos,  $\delta_h = 0,001$ ,  $\delta_h = 0,002$  y  $\delta_h = 0,004$ . Esto resulta en 4 simulaciones por cada valor de  $\Omega_m$ : una de ellas contiene el radio de los vacíos identificados en materia, es decir, considerando todas las partículas de materia bariónica y oscura, mientras que las otras tres simulaciones contienen listas de radios de vacíos identificados en halos en el espacio real, con las respectivas densidades de halo. Estos últimos conjuntos de datos contienen los radios de mayor magnitud, debido a que la distribución de la densidad de galaxias se encuentra diluida con respecto a la distribución de la densidad de materia. A manera ilustrativa, los vacíos identificados en materia poseen un radio que varía entre 2 y  $13h^{-1}$  Mpc, mientras que para aquellos identificados en halos el radio varía entre 10 y  $60h^{-1}$  Mpc aproximadamente.

El identificador utilizado para obtener los radios de los vacíos es el identificador esférico. Este método se basa en el contraste integrado de densidad de regiones subdensas asumiendo que los vacíos poseen simetría esférica [Ruiz et al., 2015, Padilla et al., 2005].

Para poder extraer información útil de los conjuntos de datos, primero se los convirtió a la escala logarítmica, como se acostumbra en el área. A estos nuevos valores para los radios de los vacíos se los separa en bins, que surgen de dividir al rango total de los radios en subintervalos. Para cada subintervalo, se contabiliza el número de elementos en él. Este bineado que se construyó para cada uno de los conjuntos posee un número fijo de bins,  $n = 15$ , es decir, 15 subintervalos de radios. Se optó por este número luego de probar con distintos valores de  $n$ , menores y mayores, para visualizar como variaba el comportamiento de la curva de abundancia. Lo que se vio es que con valores menores, la curva perdía resolución, y en el caso de valores mayores de  $n$  resultaba en una curva ruidosa con muchos picos. Además, se efectúa una normalización en el conteo al dividir al número de vacíos en un bin por un factor relacionado al tamaño de la simulación y el ancho del bin.

$$\text{conteo de vacíos normalizado} = \frac{\text{conteo de vacíos}}{\text{tamaño de la sim.} \times \text{ancho del bin}} \quad (5.1.1)$$

En esta instancia ya es posible graficar las curvas de abundancia. Al hacerlo se aprecia un pico en los valores que corresponden a los vacíos de radio menor, como se muestra en la Fig 5.2a. Los datos a la izquierda de este pico se consideran valores espurios, por lo que sólo se tienen en cuenta los radios a la derecha del pico. En las gráficas que aparecen en el Capítulo 6 no figuran los valores a la izquierda del pico, que ya habrán sido removidos, como se ve en la Fig. 5.2b.



- (a) Antes: VSF simuladas en halos y partículas para  $\Omega_m = 0,300$ , previo a la eliminación de los valores espurios. (b) Despues: VSF simuladas en halos y partículas para  $\Omega_m = 0,300$ , luego de la eliminación de los picos. Esto es notorio en las VSF para halos.

**Figura 5.2:** Remoción de los valores espurios en las gráficas.

## 5.2. Entrenamiento del algoritmo

Una vez obtenidas todas las curvas de abundancia para los 68 conjuntos de datos, se construyó con ellas una matriz que incluye la totalidad de la información, que también sirve para simplificar el código más adelante.

La matriz de datos posee columnas con la siguiente información: la cantidad de materia en el Universo ( $\Omega_m$ ), la densidad de halos presentes en la simulación ( $\delta_h$ ), el ancho del bin de la curva de abundancia en halos (esto es, el ancho del intervalo de radios en halos tenido en cuenta a la hora de hacer el conteo de vacíos), la cota superior de radios  $R_v$  en halos cuyos vacíos están incluidos, el conteo de vacíos  $A_v$  en halos, el ancho del bin de la curva de abundancia en partículas, la cota superior de radios  $R_v$  en partículas cuyos vacíos están incluidos y el conteo de vacíos  $A_v$  en partículas.

A modo ilustrativo, se muestran las primeras filas de la matriz de entrada en el Cuadro 5.1.

Features						Variable objetivo	
$\Omega_M$	$\delta_h$	ABH	$R_vH$	$A_vH$	ABP	$R_vP$	$A_vP$
0.1	0.004	0.049688	1.031968	0.000096	0.063577	0.115670	0.049480
0.1	0.004	0.049688	1.081655	0.000081	0.063577	0.179247	0.030391
0.1	0.004	0.049688	1.131343	0.000085	0.063577	0.242824	0.024558
0.1	0.004	0.049688	1.181031	0.000060	0.063577	0.306401	0.014410
0.1	0.004	0.049688	1.230719	0.000046	0.063577	0.369978	0.007630
0.1	0.004	0.049688	1.280406	0.000037	0.063577	0.433555	0.004414
:	:	:	:	:	:	:	:

Cuadro 5.1: Primeras filas de la matriz de entrada con la que se entrena los algoritmos. Las primeras seis columnas corresponden a los features de los regresores, donde:  $\Omega_m$  es el parámetro de densidad de materia de la simulación,  $\delta_h$  es la densidad de halos, ABH es el Ancho de Bins en Halos,  $R_vH$  es la Cota superior de Radios en Halos para cada bin,  $A_vH$  es el Conteo de vacíos identificados en Halos y ABP es el Ancho de Bins en Partículas. Finalmente, las ultimas dos columnas corresponden cada una a las variables objetivo que pretenden reproducir los regresores entrenados:  $R_vP$  es la Cota superior de Radios en Partículas para cada bin, y  $A_vP$  es el Conteo de vacíos identificados en Partículas.

Con estos datos se procedió a entrenar los algoritmos que luego realizarán la predicción. Se realizó el mismo procedimiento de entrenamiento para métodos de regresión seleccionados: Regresión Lineal, Gaussian Process Regressor y Random Forest Regressor.

Se abordaron dos estrategias diferentes para verificar si el tamaño del conjunto de datos de entrenamiento impactaría en los resultados predichos:

- Se entrenó el algoritmo una vez tomando como conjunto de entrenamiento a los conjuntos con  $\Omega_m$  entre  $\Omega_m = 0,100$  y  $\Omega_m = 0,500$ , con intervalos de  $\Omega_m = 0,050$ . Luego se evaluó el algoritmo usando como conjunto de prueba cada uno de los conjuntos de datos de vacíos identificados en halos restantes, es decir, aquellos con  $\Omega_m$  en el rango entre  $\Omega_m = 0,125$  y  $\Omega_m = 0,475$ , nuevamente con un intervalo de  $\Omega_m = 0,050$ . Las predicciones realizadas fueron contrastadas con los conjuntos de datos de vacíos identificados en materia.
- Se entrenó el algoritmo tantas veces como conjuntos de datos de vacíos identificados en halos había (51 en total, para los 17 valores de  $\Omega_m$ , con 3 densidades de halos cada uno). En cada ocasión, el conjunto de prueba es el *i-ésimo* conjunto de datos en halos, y el conjunto de entrenamiento está compuesto por los restantes conjuntos. Una vez hecha la predicción, se contrastó este resultado con el *i-ésimo* conjunto de datos en materia.

El primero de los caminos fue rápidamente abandonado al notar que entregaba ajustes limitados, particularmente en el sentido de que las VSF predichas eran de longitud considerablemente menor a las simuladas.

Para conseguir los mejores hiperparámetros para cada regresor, se aplicó un algoritmo de GS evaluado en las métricas  $R^2$  luego de obtener el bineado de los conjuntos de datos. Con los resultados, se entrenaron los algoritmos.

En todos los casos se hicieron dos predicciones independientes: una para obtener el radio de los vacíos  $R_v$ , y otra para obtener el conteo  $A_v$ . Esto es, los ejes  $x$  e  $y$  respectivamente de la VSF. Con los datos de las predicciones, se realizaron los gráficos de las curvas, en donde se contrastó la predicción con el valor real, y además se evaluó cada predicción con las métricas  $R^2$  y MSE.

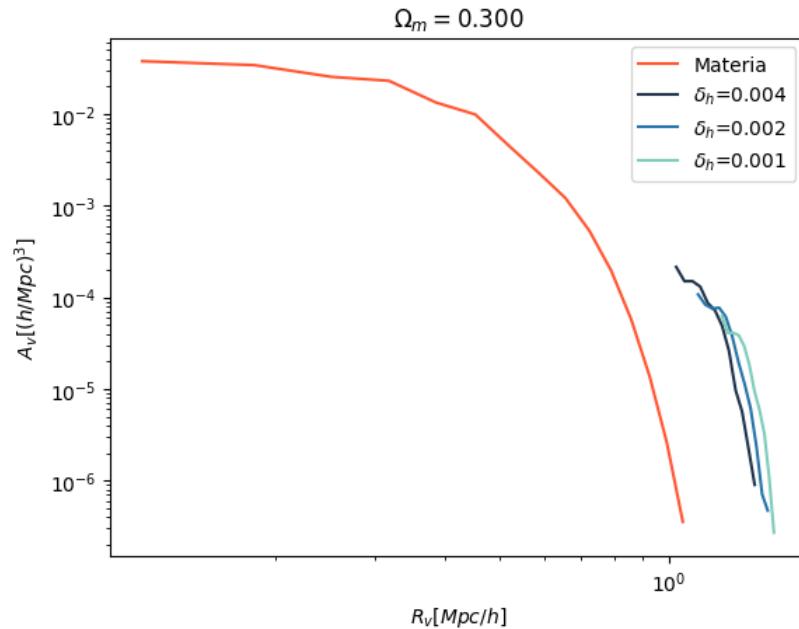
# Capítulo 6

## Resultados

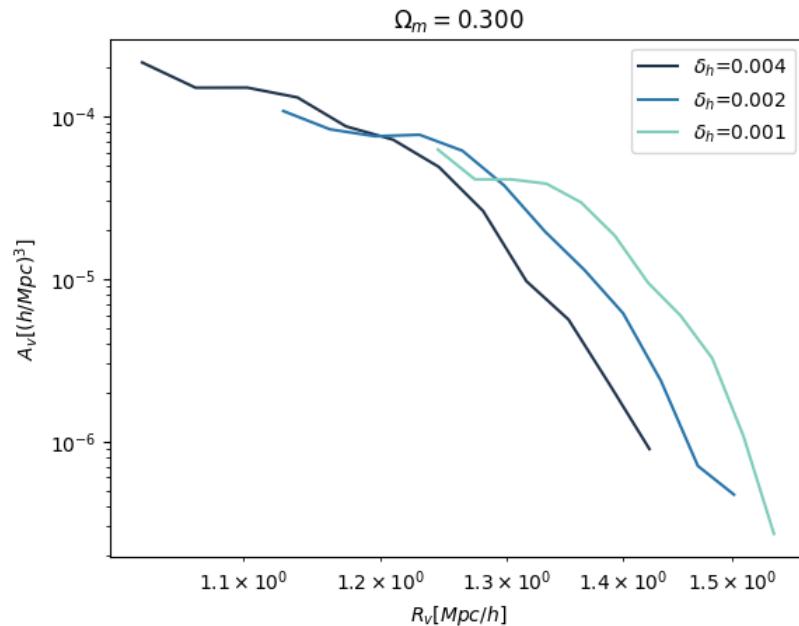
El comportamiento de las VSF para halos y partículas se muestra en la Fig. 6.1, particularmente para  $\Omega_m = 0,3$ , y un bineado de  $n = 15$  en datos simulados (ver Apéndice A). Allí se aprecian las curvas en distintos trazadores: naranja para partículas y tonos de azul para halos. Además, se ve de manera gráfica el objetivo principal de este proyecto: a partir de los datos en azul, se busca predecir los datos en naranja.

Al referirse a partículas como un trazador, lo que se implica es que el vacío ha sido identificado directamente en el campo de densidad de materia oscura. Cuando se identifican los vacíos en halos, lo que se dice es que la determinación de sus radios sólo se basa en la ubicación de halos como trazadores sesgados del campo de densidad. De ahí el hecho de que la curva de abundancia en materia esté acotada para vacíos de menor tamaño, mientras que los vacíos identificados en halos son más grandes. Un detalle de las VSF en halos se puede en la Fig. 6.2.

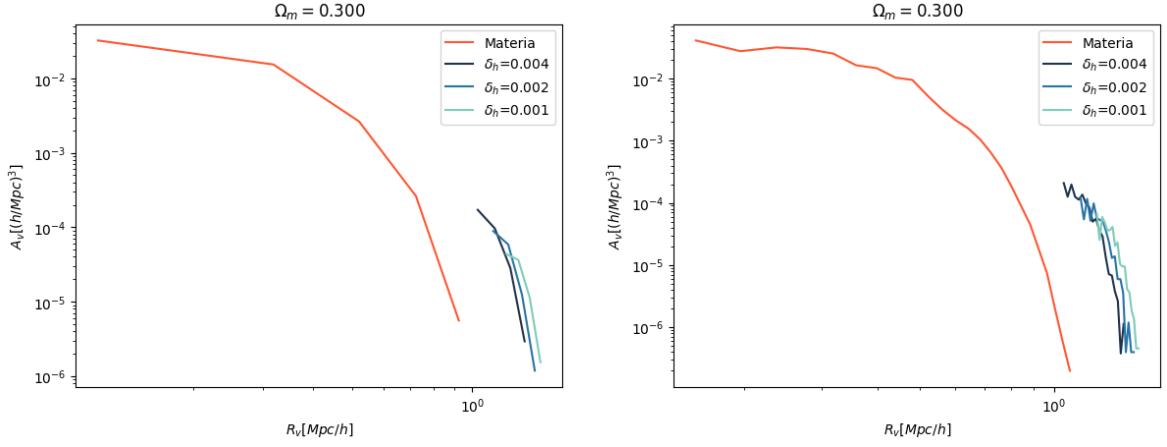
En la Fig. 6.3 se ven dos alternativas al gráfico de las curvas de abundancia de la Fig. 6.1. En ellas se muestran curvas realizadas a partir del mismo set de datos para  $\Omega_m = 0,300$ , pero variando el número de bins, es decir, variando el número de puntos que componen la gráfica.



**Figura 6.1:** VSF en materia y halos para  $\Omega_m = 0,3$ . Se ponen en evidencia las diferencias que surgen de las abundancias en distintos trazadores: los vacíos identificados en partículas, son más pequeños que los identificados en halos.



**Figura 6.2:** Detalle de las curvas de abundancia identificadas en halos para  $\Omega_m = 0,3$ .



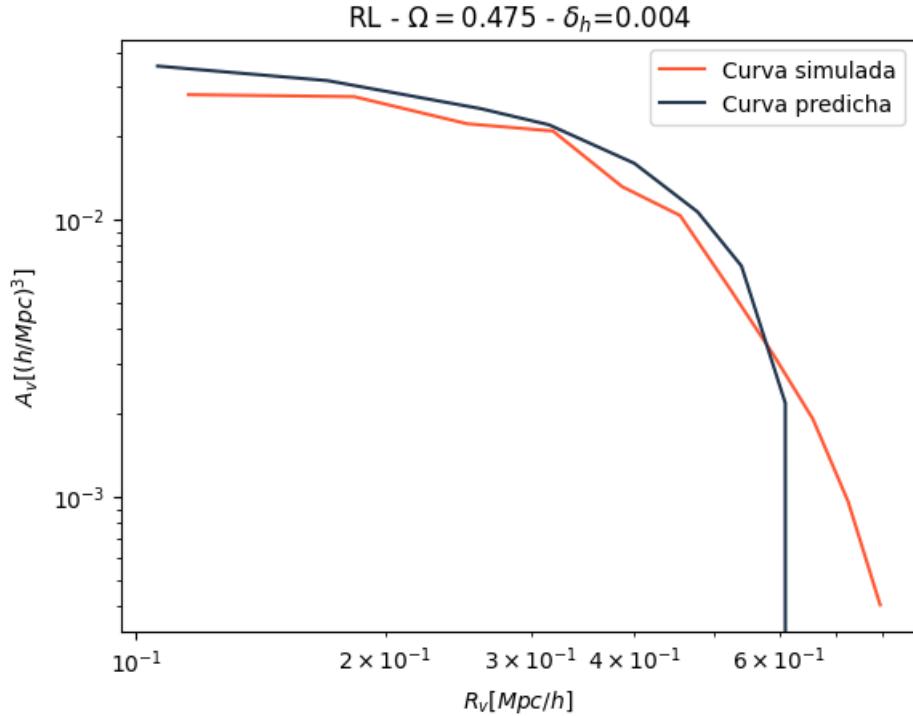
(a) VSF usando 5 bines para agrupar a los radios. La resolución aquí es menor. (b) VSF usando 25 bines para agrupar a los radios. El ruido en esta gráfica es mayor.

**Figura 6.3:** Se ilustra como impacta el número de bines en el gráfico de la curva de abundancia. Se usa la simulación con  $\Omega_m = 0,300$ , la misma que fue usada como ejemplo en la Fig. 6.1.

La Fig. 6.3a fue realizada dividiendo al conjunto de datos en 5 intervalos (bines) de radio, lo que resulta en una curva suave pero con poca resolución. En contraparte, la Fig. 6.3b utiliza 25 intervalos de radio para realizar el conteo de vacíos, lo que resulta es un gráfico más fiel al conjunto de datos, pero con un exceso de ruido que no aporta ningún beneficio al tratamiento del problema. De estos resultados surge la decisión de trabajar con  $n = 15$  bines en el resto del trabajo.

## 6.1. Regresión Lineal

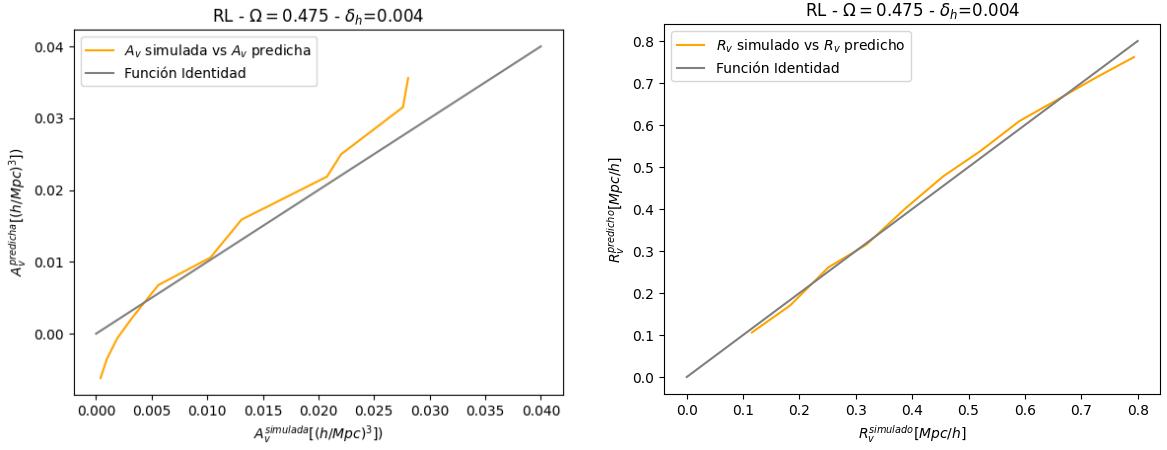
El primer algoritmo a probar fue el RL, debido a que conceptualmente constituye el modelo más simple de los tres utilizados. Como era esperable, fue método de desempeño más pobre, con un coeficiente  $R_r^2 = 0,92$  promedio para el radio de los vacíos, y  $R_a^2 = 0,85$  en el caso de la abundancia. Cuando se probó la métrica MSE arrojó los valores promedios de  $MSE_a = 0,0000298$  y  $MSE_r = 0,0037000$  para la predicción de la



**Figura 6.4:** VSF para  $\Omega_m = 0,475$  y  $\delta_h = 0,004$  obtenida a través de una regresión lineal.

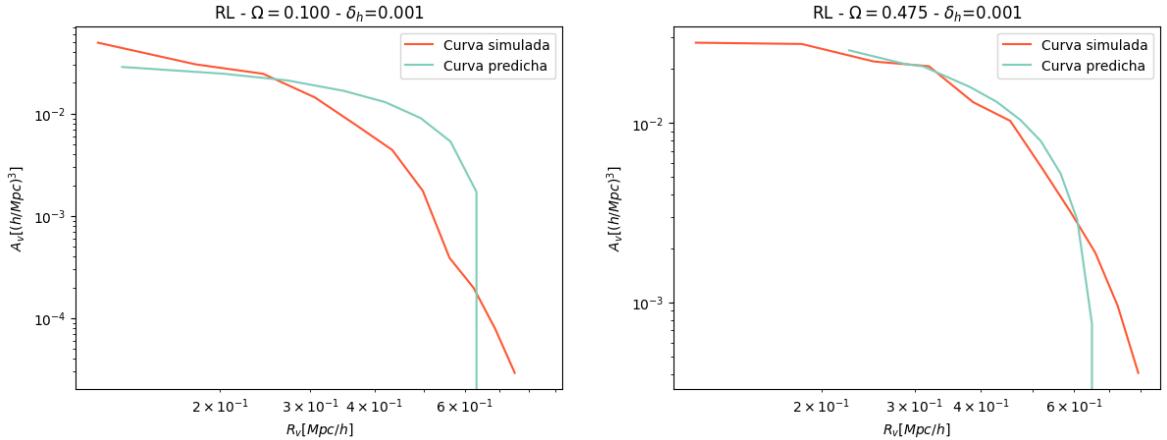
abundancia y el radio respectivamente. En la Fig. 6.4 se puede apreciar que, a pesar de ello, hay una tendencia de la VSF predicha a acercarse a los valores del conjunto de prueba, es decir, la VSF simulada. Sin embargo, conforme aumenta el radio, tienden a separarse (ver Apéndice B).

En la Fig. 6.5 se ven dos gráficos relacionados a los ajustes independientes que forman la curva de abundancia: por un lado, la Fig. 6.5a representa el conteo de vacíos  $A_v$  simulado en el eje x, mientras que en el eje y se ven los valores predichos por el RL. Se contrasta esta curva con una función identidad para apreciar mejor la relación de correspondencia entre los ejes. Por otro lado, la Fig. 6.5b hace lo análogo pero con los radios  $R_V$  de los vacíos. La finalidad de estos gráficos es mostrar que cada predicción independiente no se desvíe demasiado de los datos de prueba, e idealmente tendrían el aspecto de una función identidad.



- (a) Abundancia  $A_v$  simulada contra la abundancia  $A_v$  predicha por el regresor, con  $\Omega_m = 0,475$  y  $\delta_h = 0,004$ .  
(b) Radios  $R_v$  simulados contra el radio  $R_v$  predicho por el regresor, con  $\Omega_m = 0,475$  y  $\delta_h = 0,004$ .

**Figura 6.5:** Correspondencia entre valores del set de prueba y los valores predichos por el algoritmo de regresión lineal. La función identidad está a modo de ayuda visual.



- (a) VSF simulada contra la VSF predicha por el regresor para  $\Omega_m = 0,100$  y  $\delta_h = 0,001$ .  
(b) VSF simulada contra la VSF predicha por el regresor para  $\Omega_m = 0,475$  y  $\delta_h = 0,001$ .

**Figura 6.6:** Contraste en las predicciones para distintos parámetros de densidad de materia,  $\Omega_m = 0,100$  y  $\Omega_m = 0,475$ .

Otro resultado de interés obtenido para el RL es que, en general, las predicciones mejoran visualmente conforme aumenta  $\Omega_m$ . Esto se ilustra a modo de ejemplo en la Fig. 6.6.

## 6.2. Gaussian Process Regressor

De los tres regresores probados, el GPR fue el que más dificultades presentó a la hora de su implementación, y arrojó los valores intermedios de desempeño, con  $R_r^2 = 0,95$  y  $R_a^2 = 0,91$ , y  $MSE_a = 0,0000177$  y  $MSE_r = 0,0023000$ . En los Códigos 2 y 1 se pueden ver los kernels y el parámetro  $\alpha$  empleados para hacer las respectivas predicciones, con sus hiperparámetros obtenidos a través de un algoritmo GS utilizando el criterio  $R^2$ .

---

```

1 GaussianProcessRegressor(
2     alpha=0.0001,
3     kernel=
4         0.447 ** 2 * RBF(length_scale=0.02) +
5         RationalQuadratic(alpha=0.78, length_scale=1.2)
6     )
7 )
```

---

Código 1: Kernel y parámetro  $\alpha = 0,0001$  para el GPR entrenado para predecir  $R_v$ . En el kernel se tiene un término RBF que establece la similitud de los datos a través del parámetro de escala  $l = 0,02$ , escalado por un término constante 0,447, y otro término RQ caracterizado por un parámetro de escala  $l = 1,2$  y un parámetro de escala de mezcla  $\alpha_k = 0,78$ .

La forma de los kernels fue obtenida empíricamente.

En la Fig. 6.7 se contrasta a modo de ejemplo la curva de abundancia simulada contra la predicha a partir del conjunto de datos correspondiente a los vacíos con

---

```

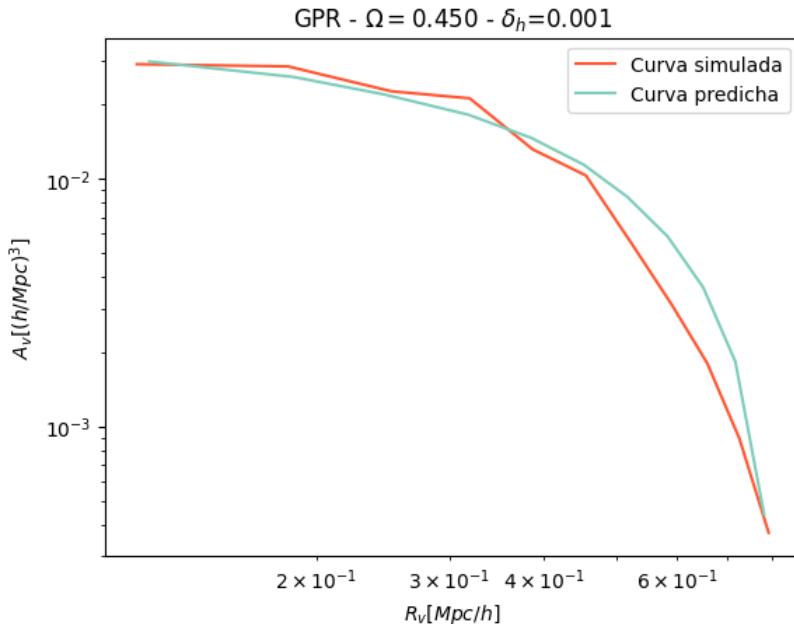
1 GaussianProcessRegressor(
2     alpha=0.0001,
3     kernel=(
4         0.0316 ** 2 + RBF(length_scale=0.001) +
5         RationalQuadratic(alpha=0.0001, length_scale=0.001)
6     )
7 )

```

---

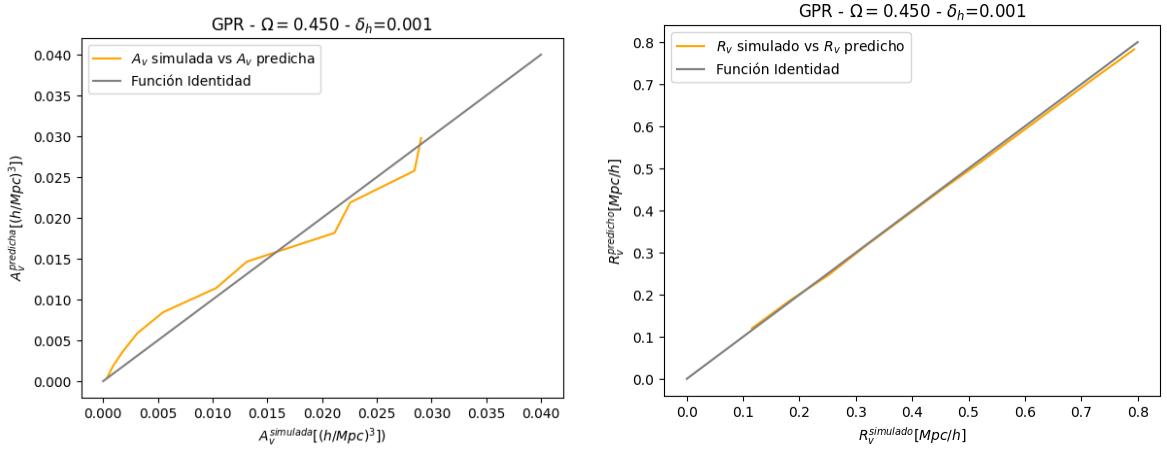
Código 2: Kernel y parámetro  $\alpha = 0,0001$  para el GPR entrenado para predecir  $A_v$ . En el kernel se tiene un término constante 0,0316, un término RBF que establece la similitud de los datos a través del parámetro de escala  $l = 0,001$ , y otro término RQ caracterizado por un parámetro de escala  $l = 0,001$  y un parámetro de escala de mezcla  $\alpha_k = 0,0001$ .

$\Omega_m = 0,450$  y  $\delta_h = 0,001$  (ver Apéndice C).



**Figura 6.7:** VSF para  $\Omega_m = 0,450$  y  $\delta_h = 0,001$  obtenida a través de un Gaussian Process Regressor.

De manera análoga al análisis hecho para la RL, en las Figs. 6.8a y 6.8b se grafica el



(a) Abundancia  $A_v$  simulada contra la abundancia predicha por el regresor, con  $\Omega_m = 0,450$  y  $\delta_h = 0,001$ .

(b) Radios  $R_v$  simulados contra el radio predicho por el regresor, con  $\Omega_m = 0,450$  y  $\delta_h = 0,001$ .

**Figura 6.8:** Correspondencia entre los valores simulados y los valores predichos por el algoritmo GPR para  $A_v$  y  $R_v$ . La función identidad está a modo de ayuda visual.

conteo  $A_v$  simulado contra el predicho, y el radio  $R_V$  simulado contra el radio predicho respectivamente, con una función identidad presente para facilitar la visualización.

A diferencia del regresor lineal (y del RF, como se verá más adelante), el GPR no mostró tanta variación visual en las curvas predichas conforme variaba  $\Omega_m$ .

### 6.3. Random Forest Regressor

El RF fue el algoritmo que obtuvo los mejores puntajes en ambas métricas, con  $R_a^2 = 0,96$  y  $R_r^2 = 0,96$ , y  $MSE_a = 0,0000088$  y  $MSE_r = 0,0018000$ . Los parámetros usados para entrenar este regresor fueron establecidos luego de hacer una exploración a través de un algoritmo GS de las distintas configuraciones posibles, medidas con la métrica  $R^2$ . Los resultados se encuentran en el Listado 3.

En la Fig. 6.9 se aprecia un ejemplo para  $\Omega_m = 0,0200$  y  $\delta_h = 0,002$ , con los

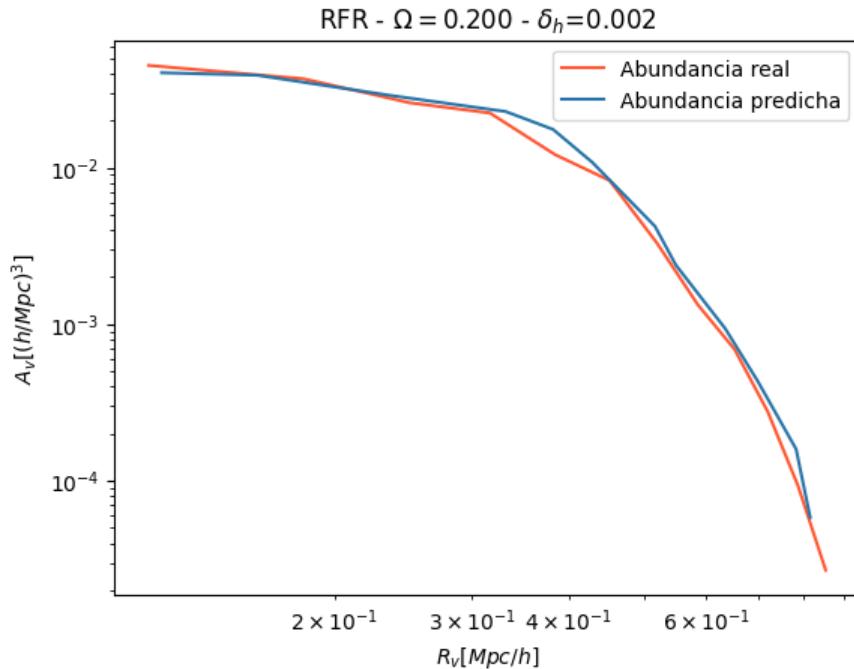
```

1 RandomForestRegressor(
2     criterion='squared_error', max_features=None,
3     min_samples_split=2, n_estimators=500,
4 )

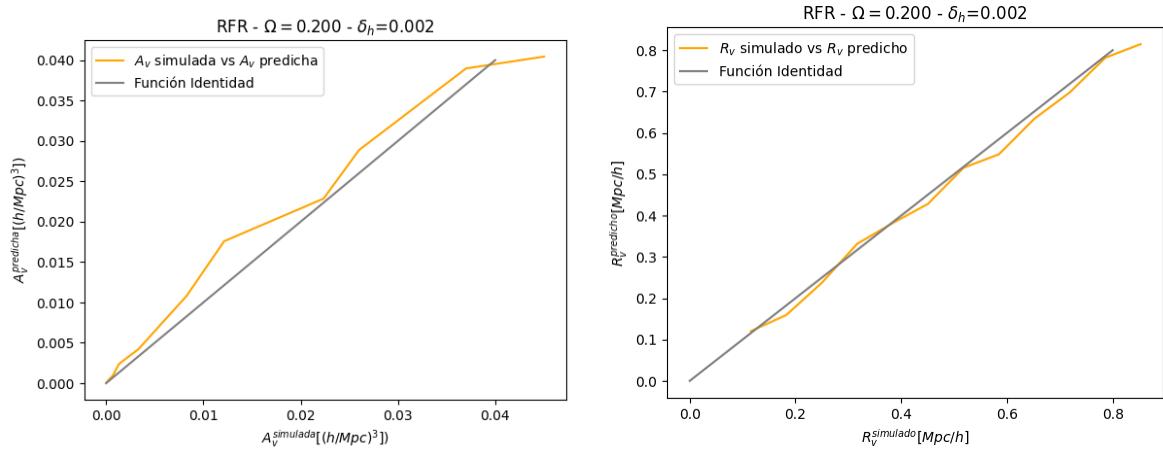
```

Código 3: Parámetros del RF entrenado para predecir el radio y la abundancia de los vacíos. Se establece que el criterio a usar es el de mínimos cuadrados, que se usarán 500 árboles en el algoritmo, que la cantidad mínima de datos para crear un nodo es 2 y que el algoritmo evaluará entre todas las features de los datos al definir la función de los nodos.

correspondientes análisis de radio  $R_v$  y conteo  $A_v$  de vacíos simulados contra predichos en las Figs. 6.10b y 6.10a respectivamente (ver Apéndice D).



**Figura 6.9:** VSF para  $\Omega_m = 0,200$  y  $\delta_h = 0,002$  obtenida con Random Forest Regressor.

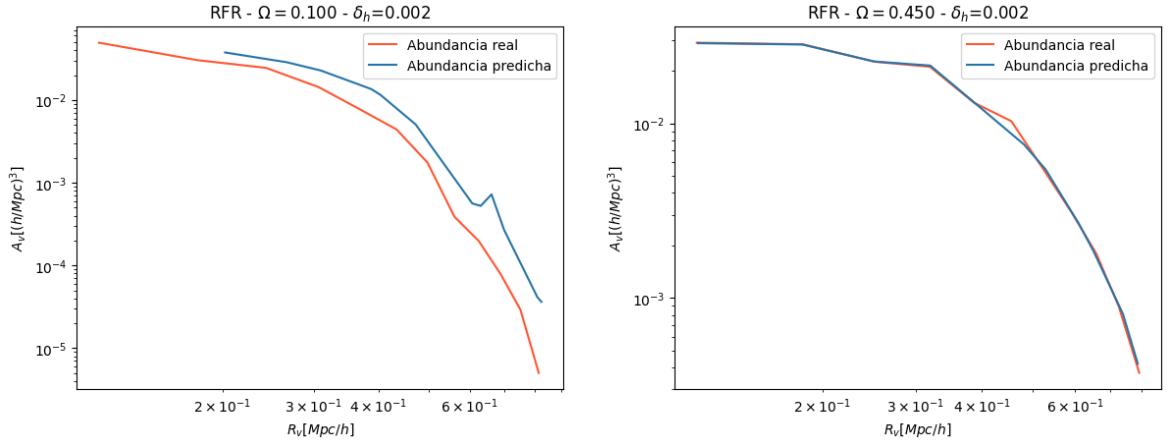


(a) Abundancia  $A_v$  simulada contra la predicha por el regresor, con  $\Omega_m = 0,200$  y  $\delta_h = 0,002$ .

(b) Radio  $R_v$  simulado contra el predicho por el regresor, con  $\Omega_m = 0,200$  y  $\delta_h = 0,002$ .

**Figura 6.10:** Correspondencia entre valores del conjunto de prueba y los valores predichos por el algoritmo RF. La función identidad está a modo de ayuda visual.

Para el caso del RF, también se evidencia que las predicciones mejoran visualmente conforme aumenta  $\Omega_m$ , como se vio en el caso de la RL. Esto se ilustra a modo de ejemplo en la Fig. 6.11.



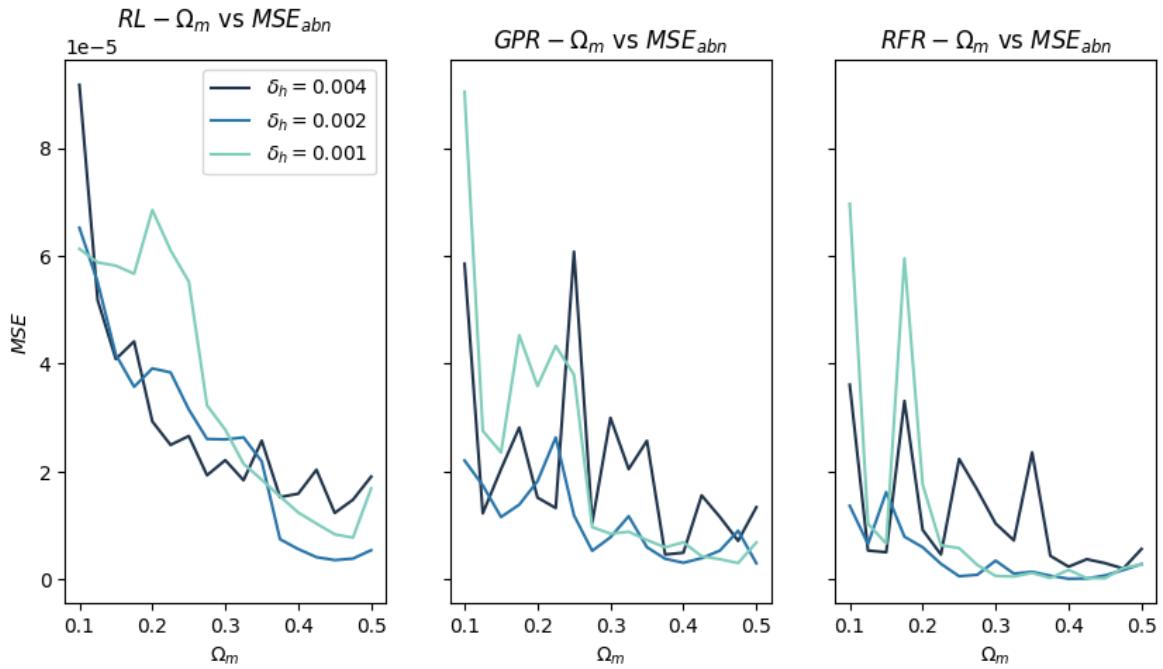
(a) VSF simulada contra la VSF predicha por el regresor, con  $\Omega_m = 0,100$  y  $\delta_h = 0,002$ . (b) VSF simulada contra la VSF predicha por el regresor, con  $\Omega_m = 0,450$  y  $\delta_h = 0,002$ .

**Figura 6.11:** Comparación en las predicciones para  $\Omega_m = 0,100$  y  $\Omega_m = 0,450$ .

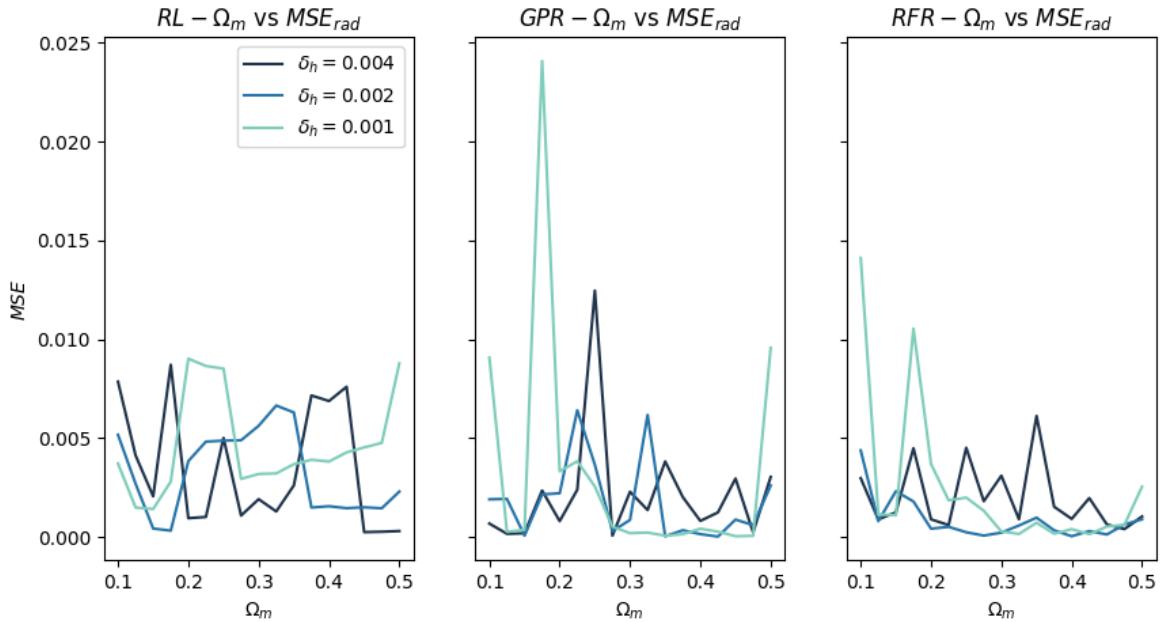
## 6.4. Comparación del MSE acorde al parámetro de densidad

Ya habiendo analizado independientemente cada uno de los regresores, la última prueba que se hizo a nivel global fue comparar la métrica MSE entre ellos, por un lado para el ajuste del conteo de vacíos, y por el otro lado el ajuste de sus radios. Dichos resultados se pueden apreciar en las Figs. 6.12 y 6.13 respectivamente.

Lo que se observó es que no existe mucha variación en la métrica MSE conforme aumenta  $\Omega_m$  para el caso de  $R_v$ . Esto es particularmente evidente para el caso del RL. Sin embargo, no ocurre lo mismo al analizar lo que ocurre en el caso del ajuste del conteo  $A_v$ . Todos los regresores evidencian una reducción en el MSE obtenido en sus predicciones conforme aumenta  $\Omega_m$ .



**Figura 6.12:** Métrica MSE obtenida para cada predicción de  $A_v$ , es decir, variando desde  $\Omega_m = 0,1$  a  $\Omega_m = 0,5$ , en las tres densidades de halos.



**Figura 6.13:** Contraste entre las métricas MSE en los ajustes  $R_v$ , para los tres regresores. En este caso no se evidencia tendencia a disminuir conforme aumenta  $\Omega_m$ .

# Capítulo 7

## Discusión y conclusiones

En el presente trabajo se presenta una nueva prueba de concepto que se aprovecha de las capacidades que hoy en día ofrecen los métodos de ML. El método consiste en entrenar algoritmos de distintos tipos para que sean capaces de predecir la VSF de vacíos identificados en materia, que no es observable, a partir de los datos de vacíos identificados en halos de materia oscura o galaxias, que sí pueden identificarse en catálogos observacionales.

En este trabajo se logró demostrar que se puede establecer una conexión entre las curvas de abundancia de vacíos para distintos trazadores de materia identificados en simulaciones cosmológicas a través de técnicas de ML.

Con variados niveles de correspondencia entre las curvas simuladas y predichas dependiendo del regresor empleado, el método mostró ser efectivo para cosmologías  $\Lambda$ CDM planas con niveles variados de  $\Omega_m$ . Las métricas  $R^2$  y MSE empleadas dieron los valores promedios para la abundancia y el radio de los vacíos enunciados en el Cuadro 7.1. Acorde a las métricas usadas, el algoritmo que mejor desempeño tuvo fue el RF, y este resultado también se pone en evidencia al inspeccionar las gráficas que contrastan las VSF hechas a partir de los datos simulados y a partir de los datos

Métrica	RL	GPR	RF
$R^2_{abn}$	0,8500000	0,9500000	0,9600000
$R^2_{rad}$	0,9200000	0,9100000	0,9600000
$MSE_{abn}$	0,0000298	0,0000177	0,0000088
$MSE_{rad}$	0,0037000	0,0023000	0,0018000

Cuadro 7.1: Valores promedio de las métricas para los ajustes de radio y abundancia para los tres regresores.

predichos, que presentan un ajuste considerablemente mejor al de los otros regresores.

También se debe destacar el no tan buen desempeño del RL, aunque este resultado era en realidad esperable. El hecho de que un ajuste lineal falle en ajustar el comportamiento de la VSF confirma que un bias lineal no es suficiente para trazar la distribución de vacíos identificados en materia oscura a partir de los vacíos identificados en halos o galaxias.

De las pruebas con los métodos RL y RF se debe destacar la mejora en los ajustes conforme aumenta el parámetro de densidad  $\Omega_m$ , y de los tres regresores se concluye que se benefician de conjuntos de entrenamiento mayores para entregar VSF más cercanas a las simuladas inicialmente.

Otro de los resultados obtenidos en este trabajo es que, al comparar con el parámetro de densidad de materia del Universo medido de  $\Omega_m \sim 0,3$  [Planck Collaboration, 2020], se aprecia que el método logró obtener predicciones para valores marcadamente distintos de  $\Omega_m$ . Al día de hoy, es impensado hablar de parámetros de densidad como  $\Omega_m = 0,1$  o  $\Omega_m = 0,5$ , por lo que estos resultados son favorables para la robustez del método.

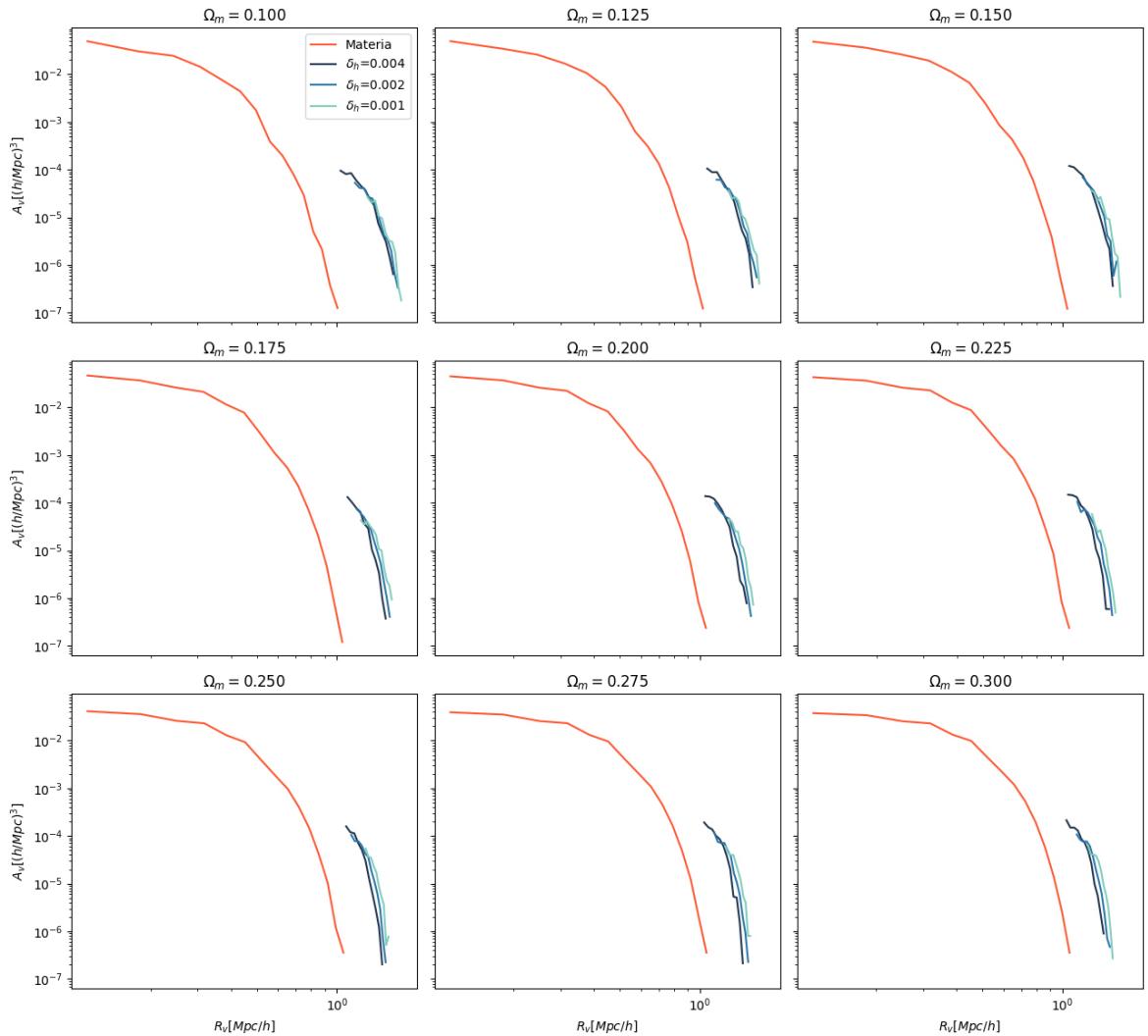
Con respecto al desempeño del GPR, se concluye que a pesar de no haber tenido el mejor desempeño, existe la posibilidad de mejora debido a la flexibilidad que el

algoritmo posee en la elección y diseño de los kernels usados, en contraste con la elección simple del mismo por la que se optó para este proyecto.

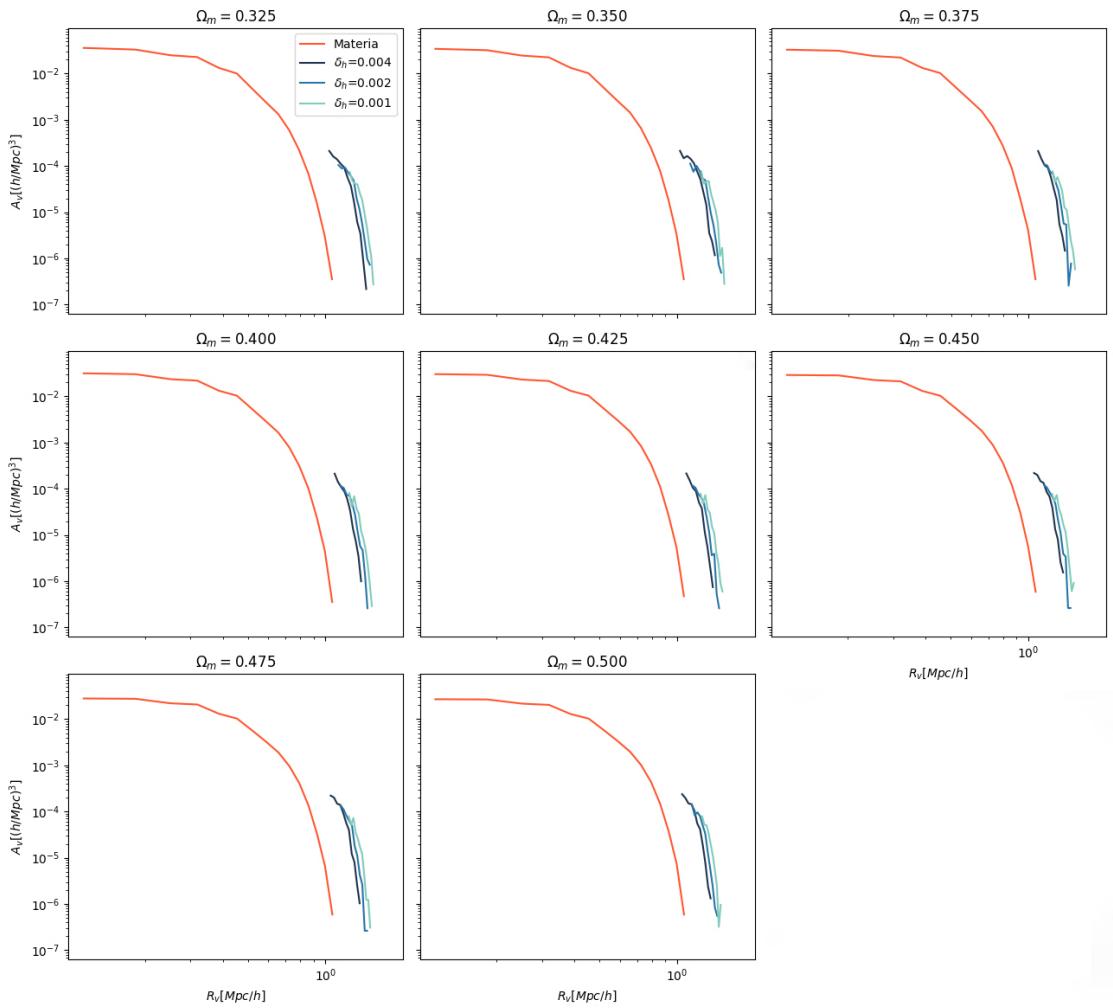
Todavía es necesario mucho trabajo para convertir este método en un test cosmológico robusto capaz de hacer aportes significativos al problema de la energía oscura, pero un buen punto inicial podría ser reemplazar el identificador esférico de vacíos empleado aquí por el *popcorn void finder*, un identificador de vacíos que refleja más fielmente la forma de las estructuras. [Paz et al., 2023]. Por otro lado, incluir en la identificación de los vacíos las distorsiones y efectos que pueden ser observados en los catálogos reales, que están más allá del alcance de este proyecto.

# **Apéndice A**

## **Abundancia**



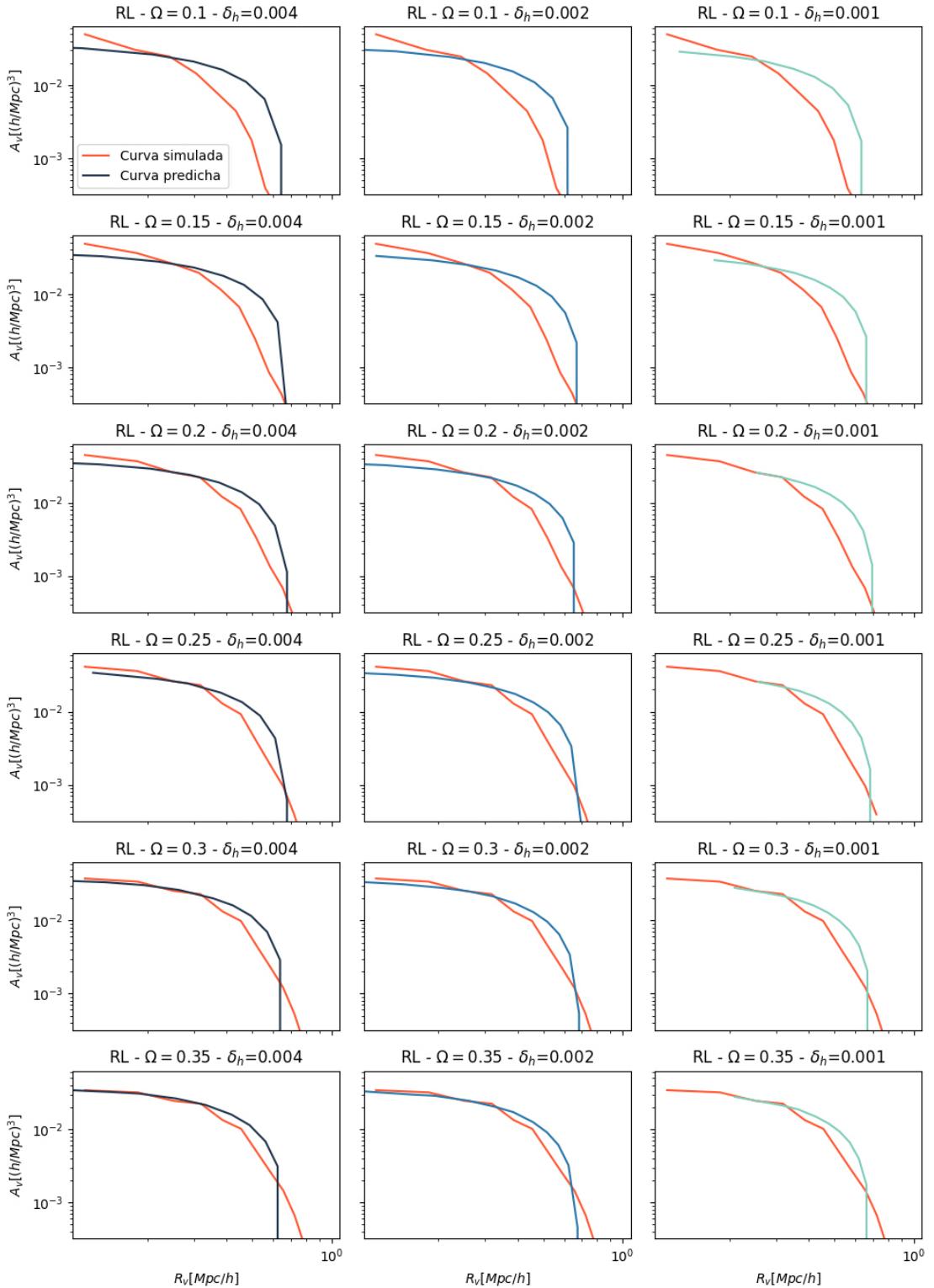
**Figura A.1:** VSF simuladas en materia y halos. Se ponen en evidencia las diferencias que surgen en las abundancias para distintos trazadores.



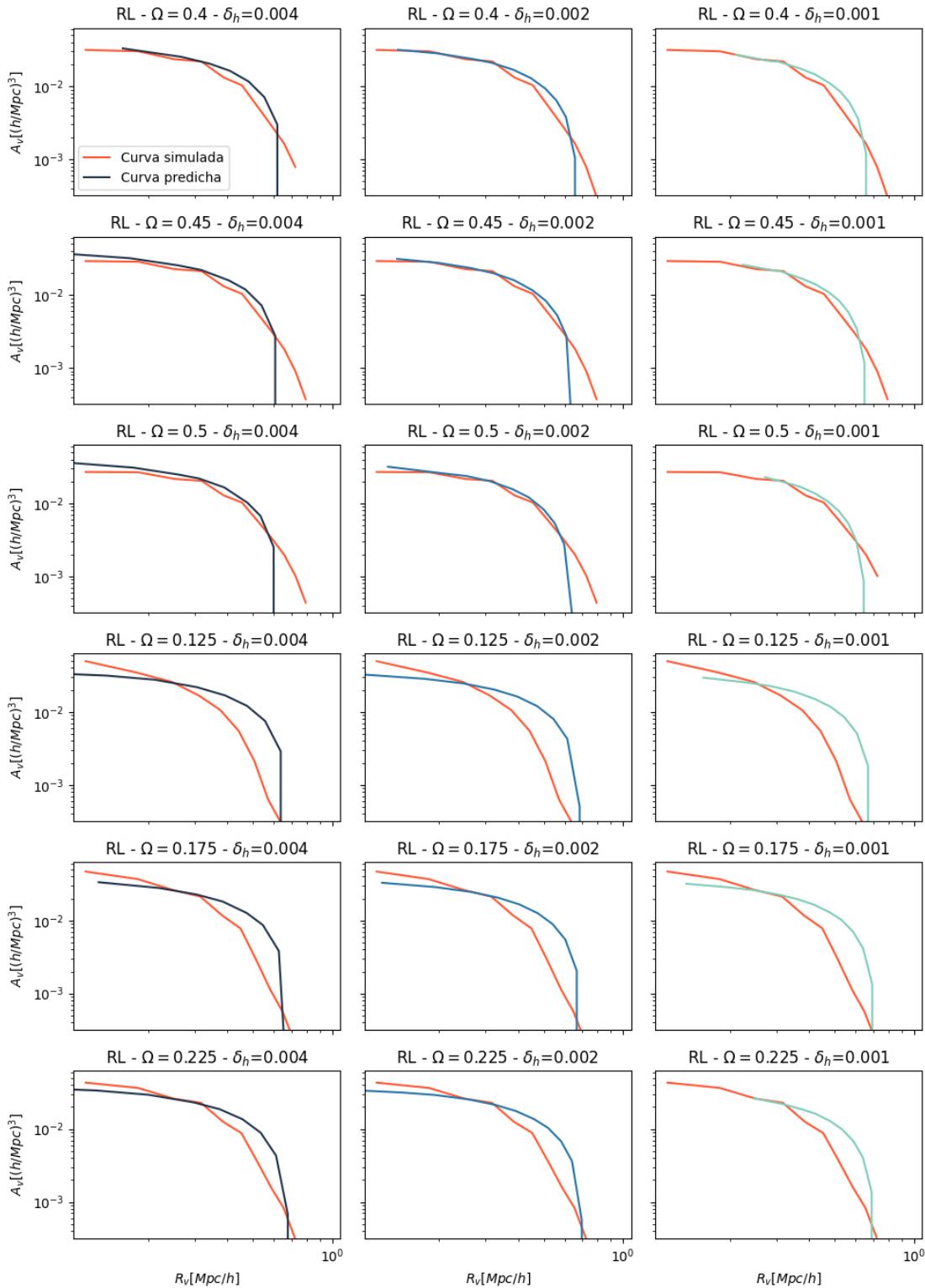
**Figura A.2:** (cont.) VSF simuladas en materia y halos. Se ponen en evidencia las diferencias que surgen en las abundancias para distintos trazadores.

# **Apéndice B**

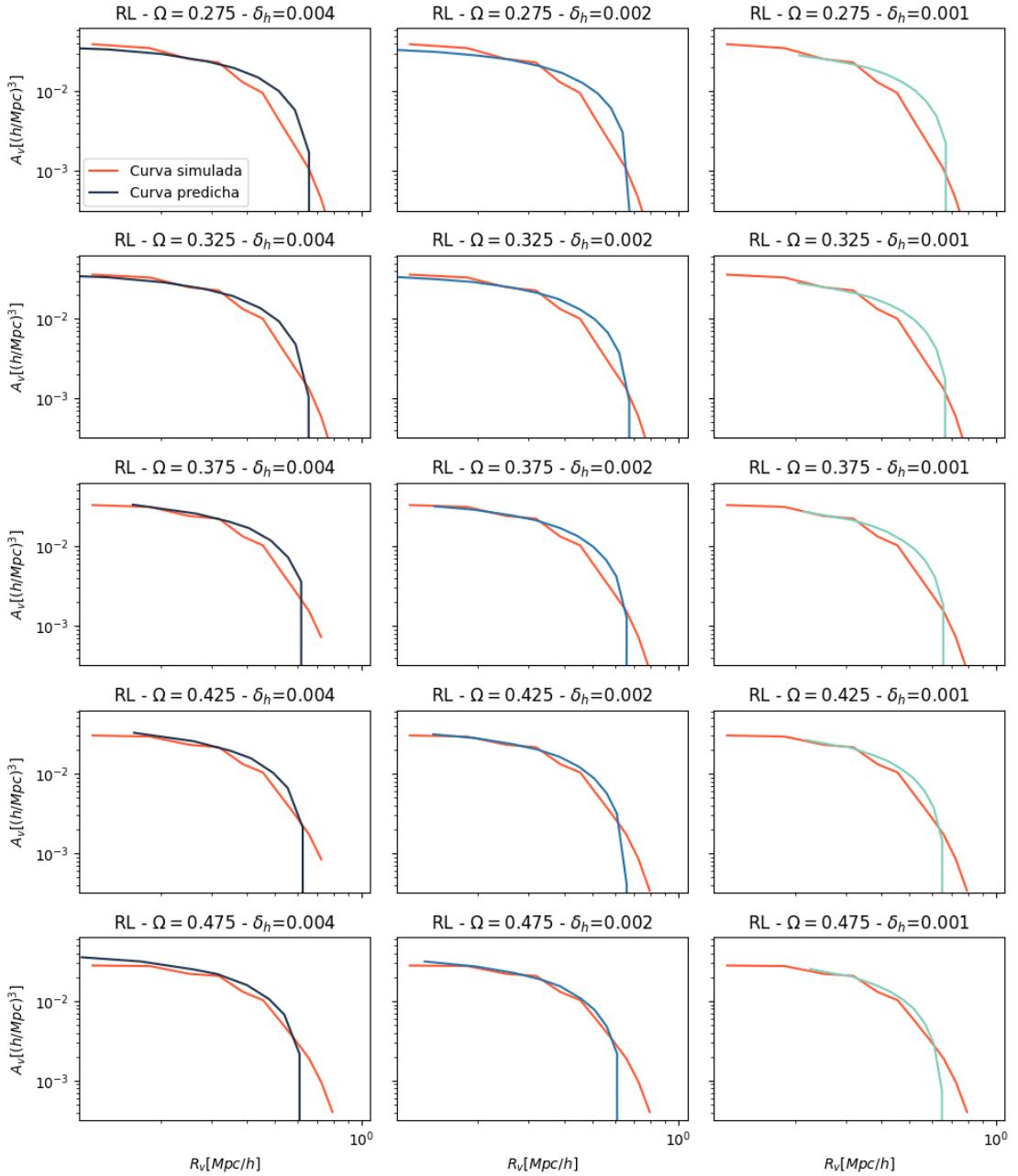
## **Regresión lineal**



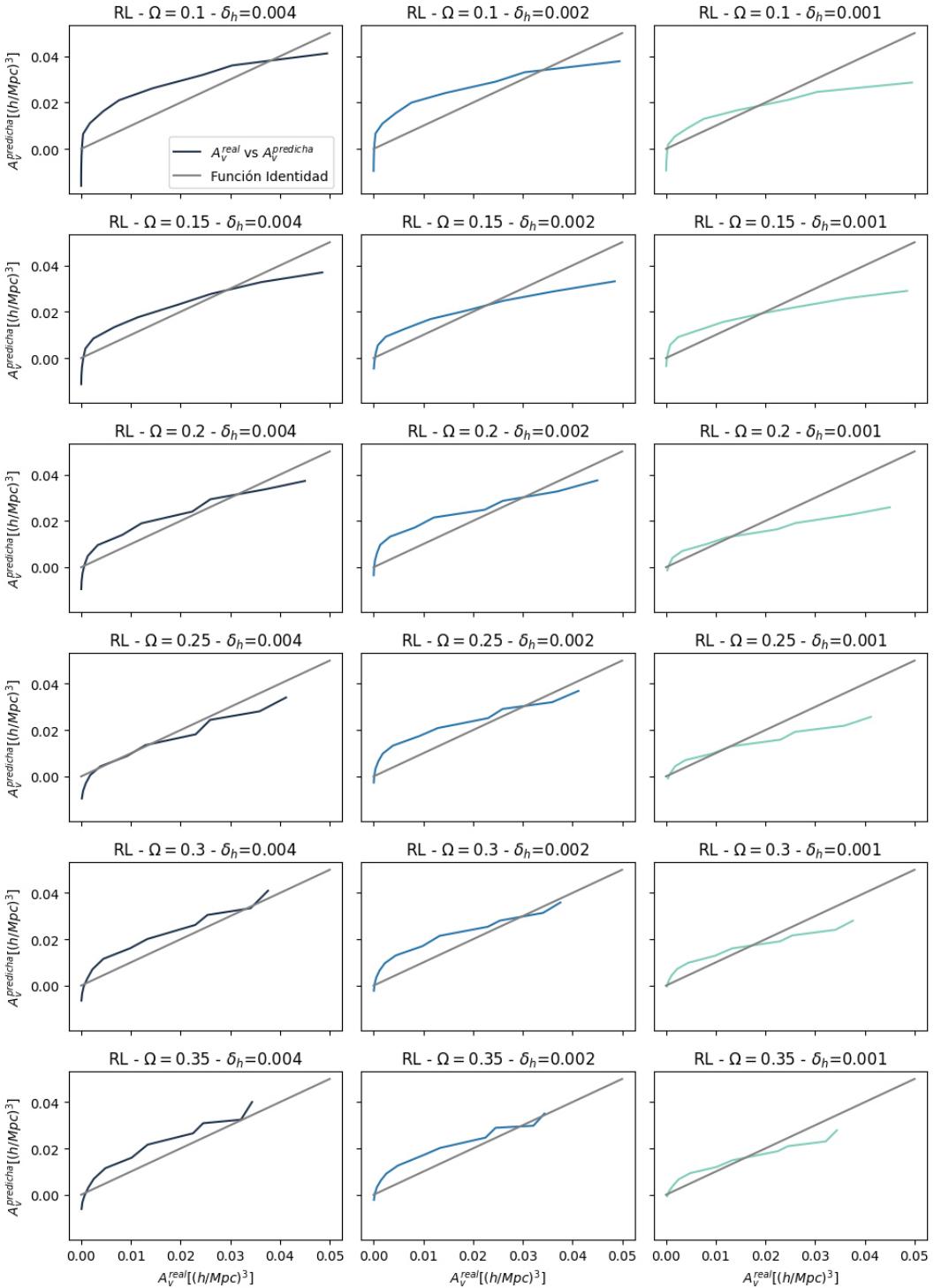
**Figura B.1:** VSF en materia simuladas vs VSF en materia predichas con un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



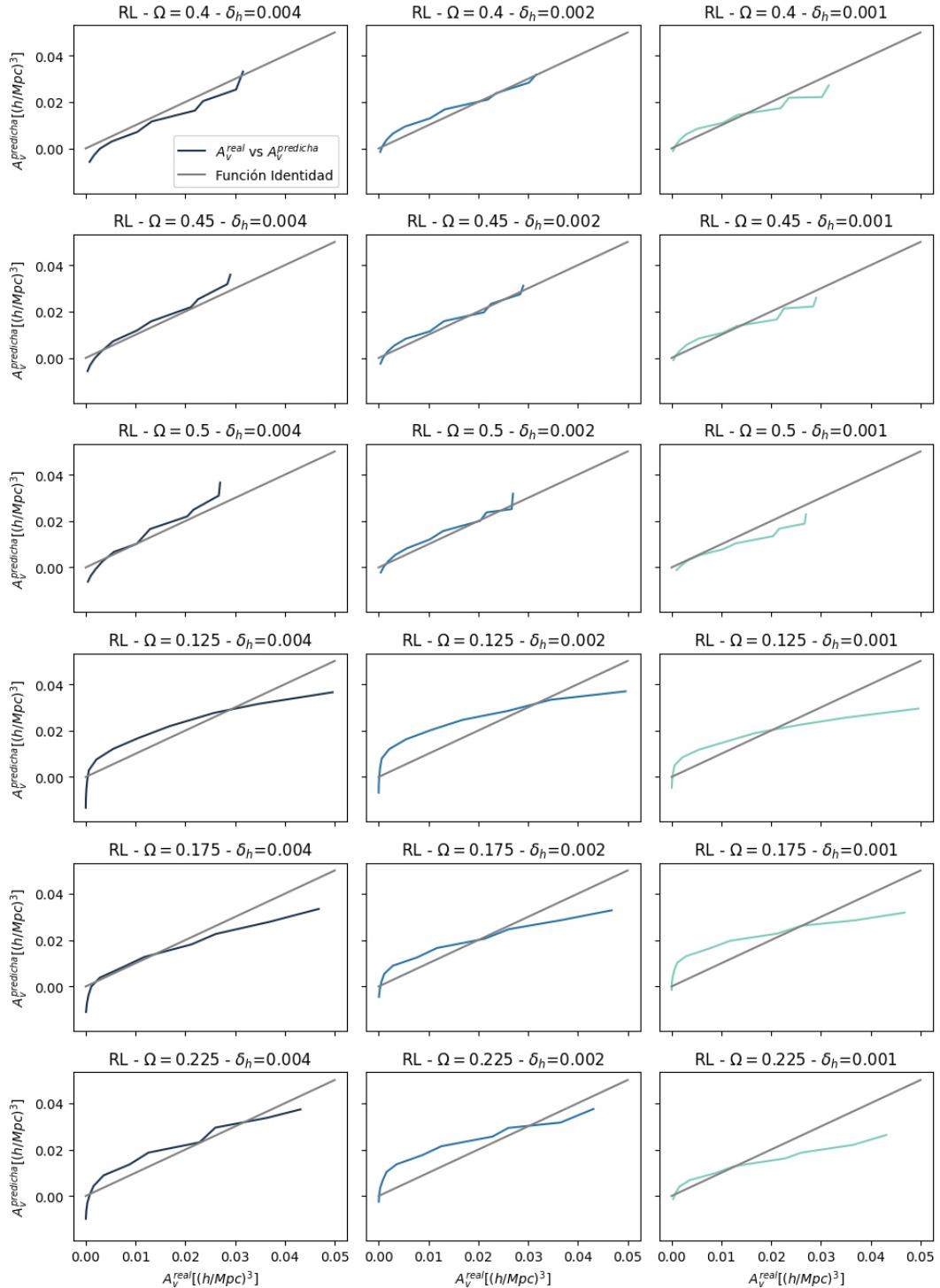
**Figura B.2:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



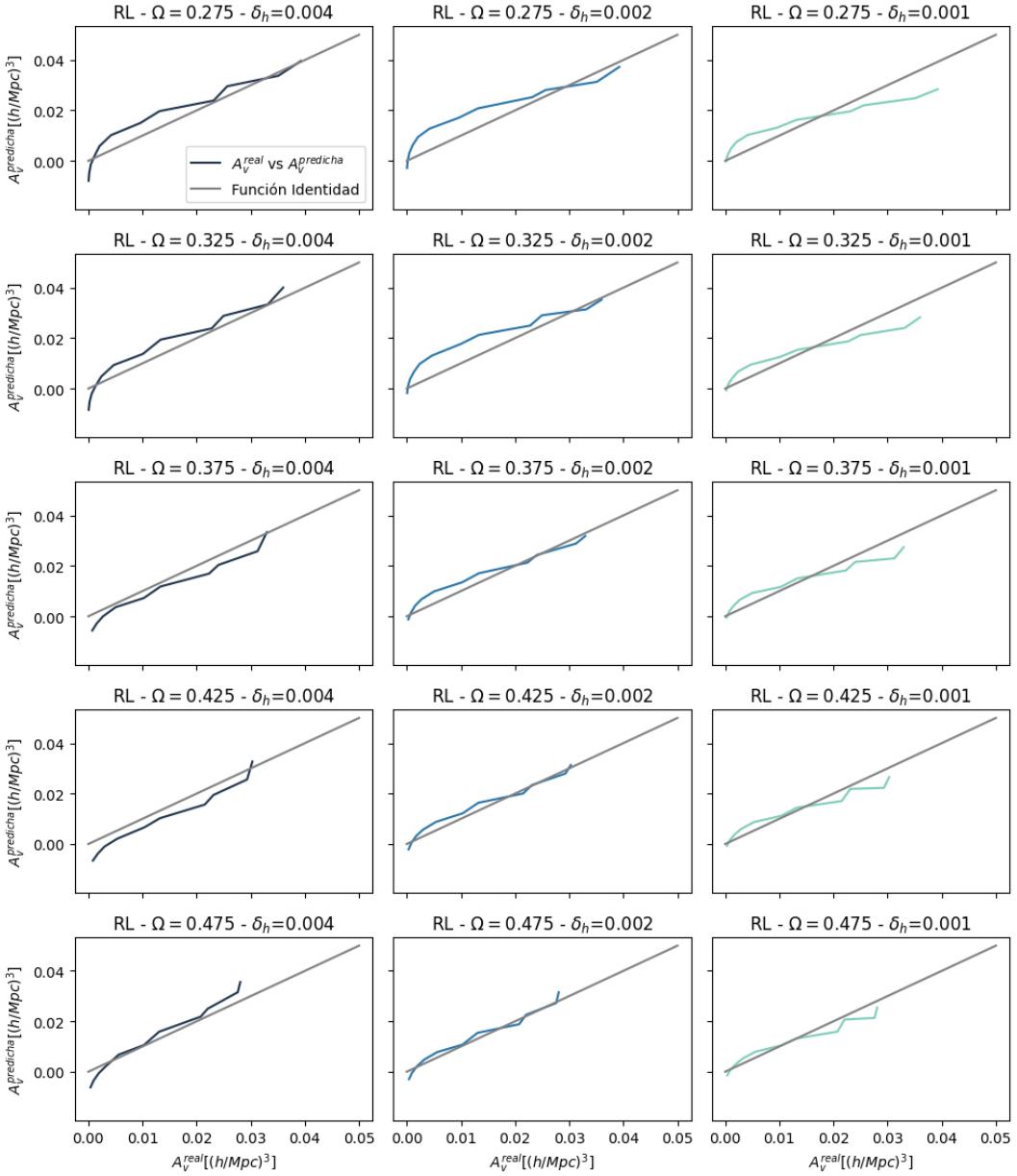
**Figura B.3:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



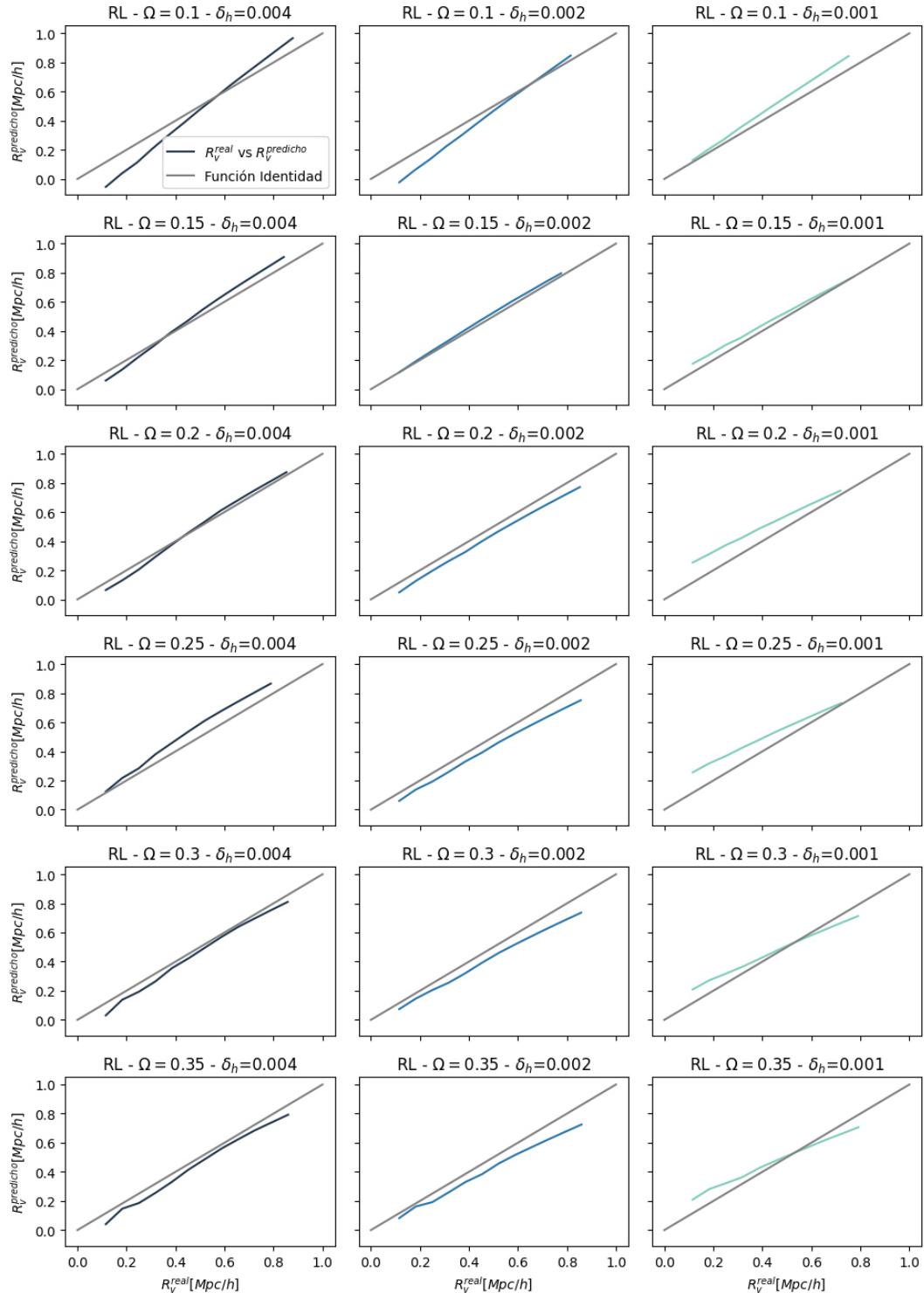
**Figura B.4:** Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



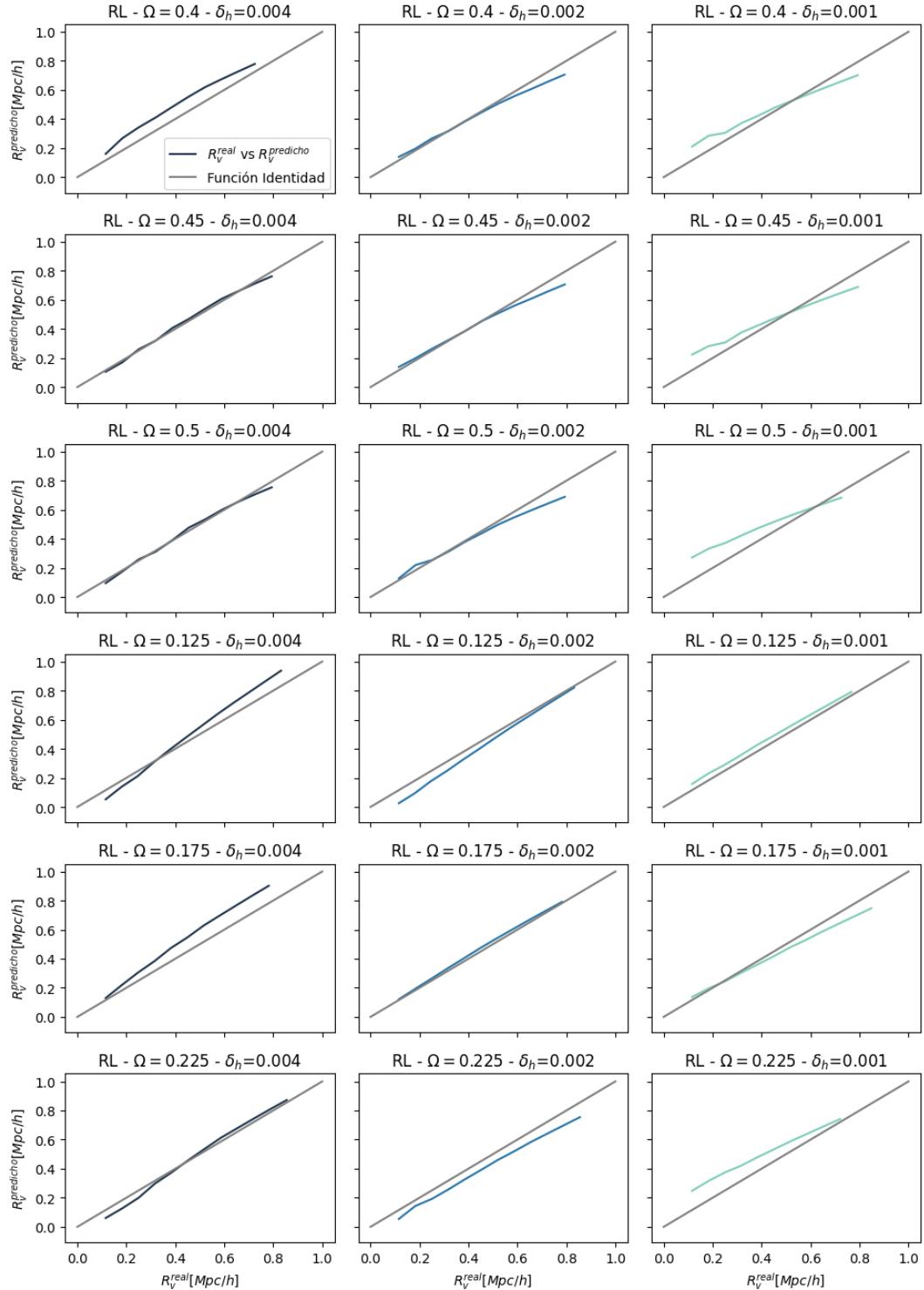
**Figura B.5:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



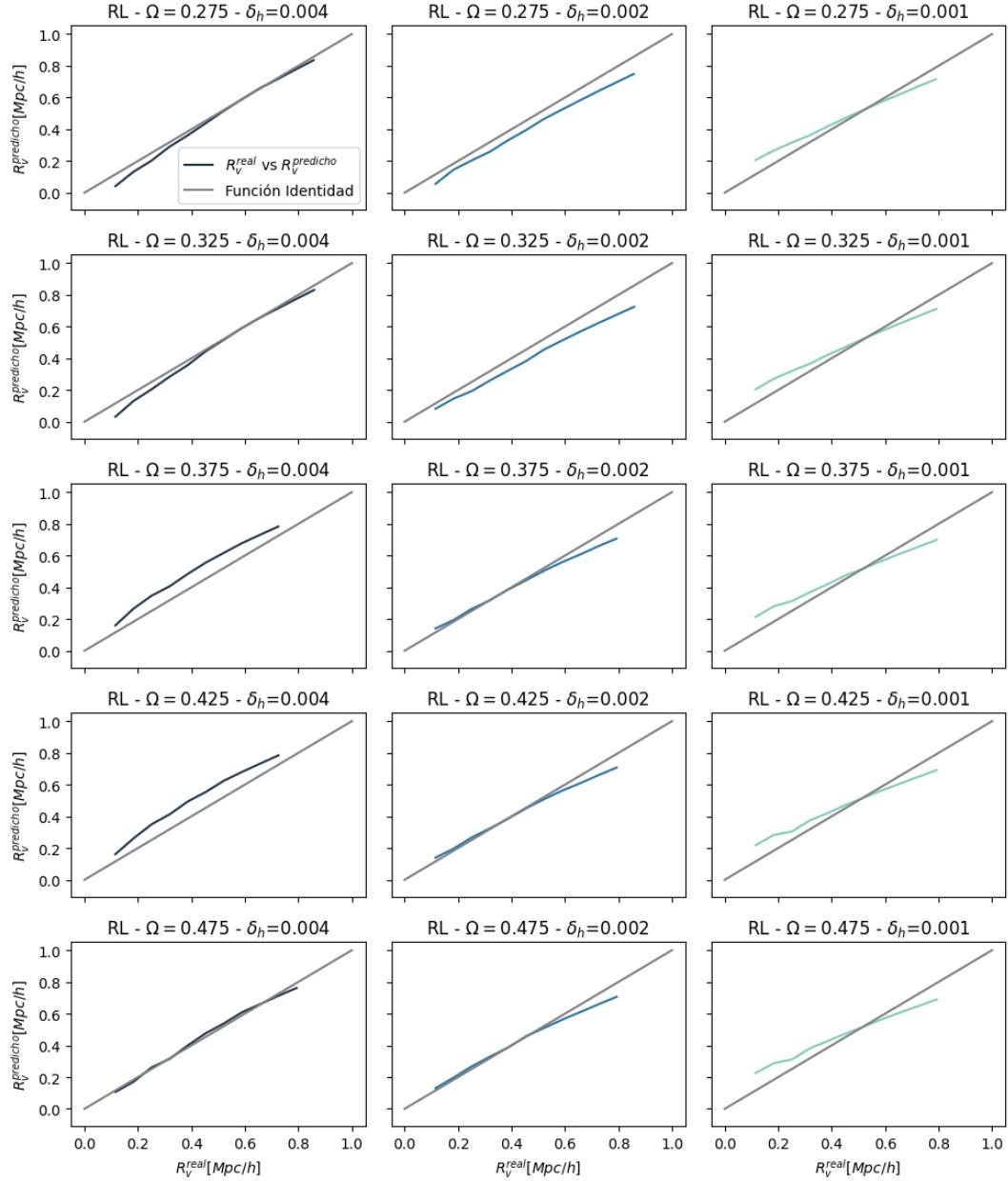
**Figura B.6:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



**Figura B.7:** Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



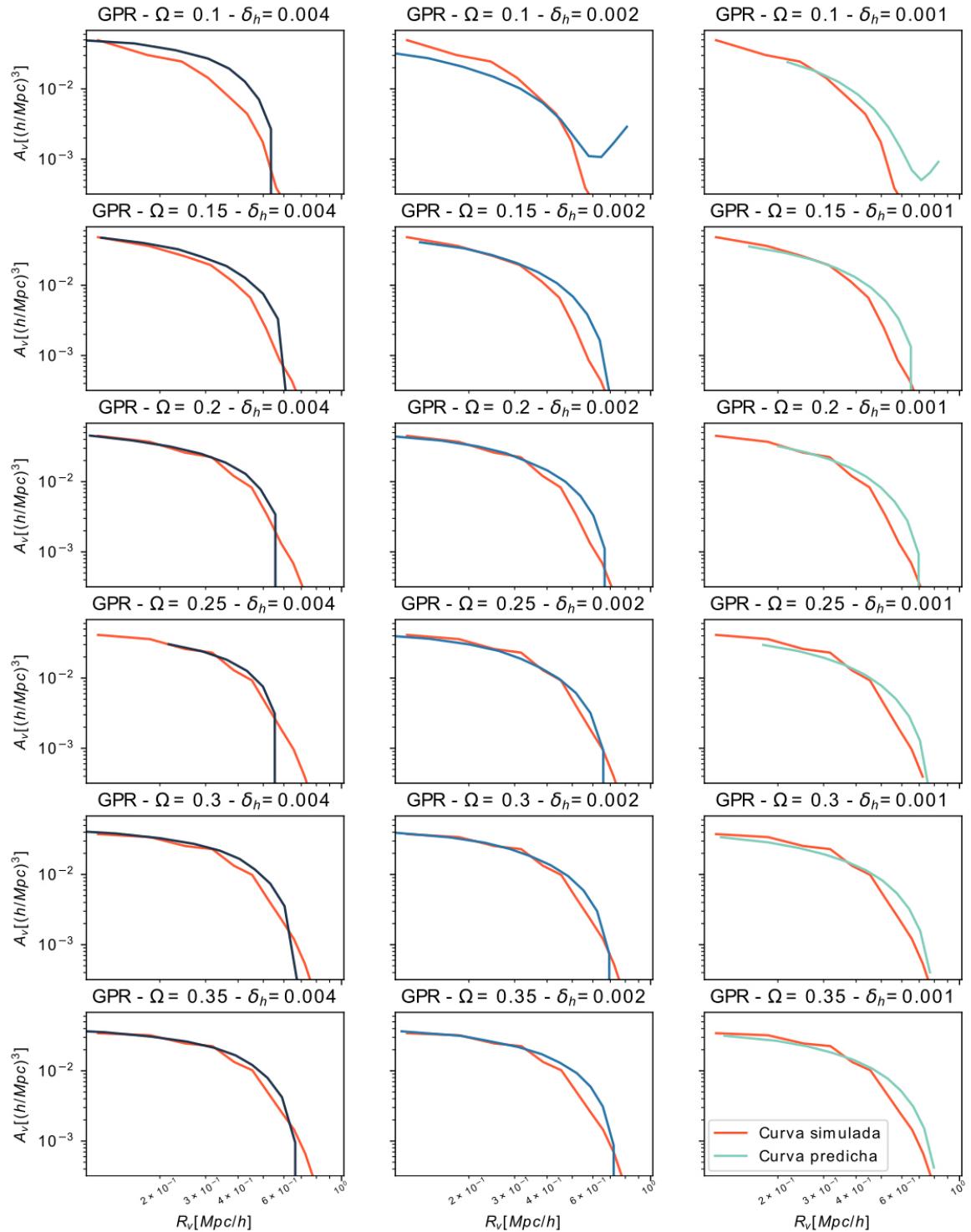
**Figura B.8:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



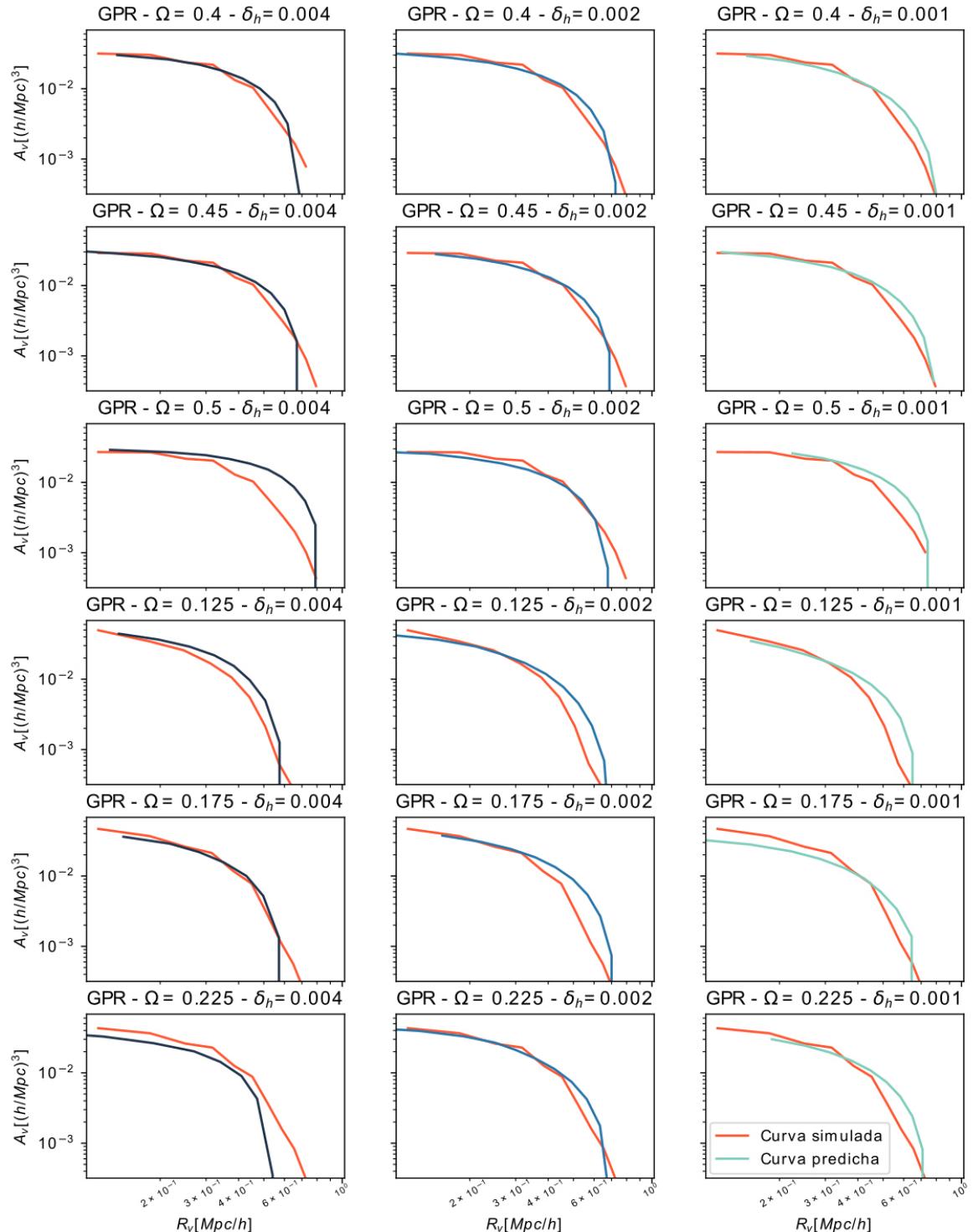
**Figura B.9:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Regresor lineal, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.

# **Apéndice C**

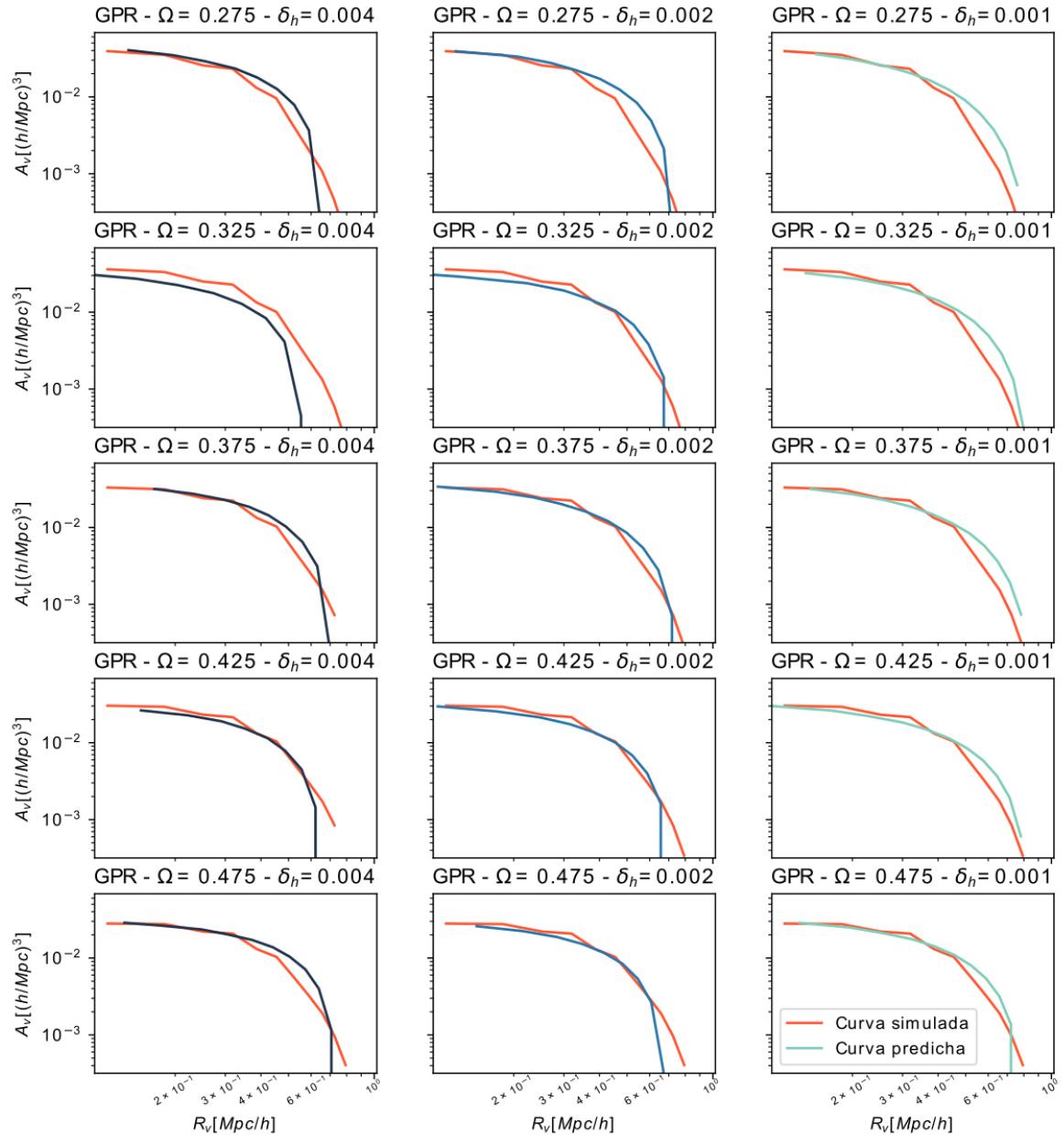
## **Gaussian Process Regressor**



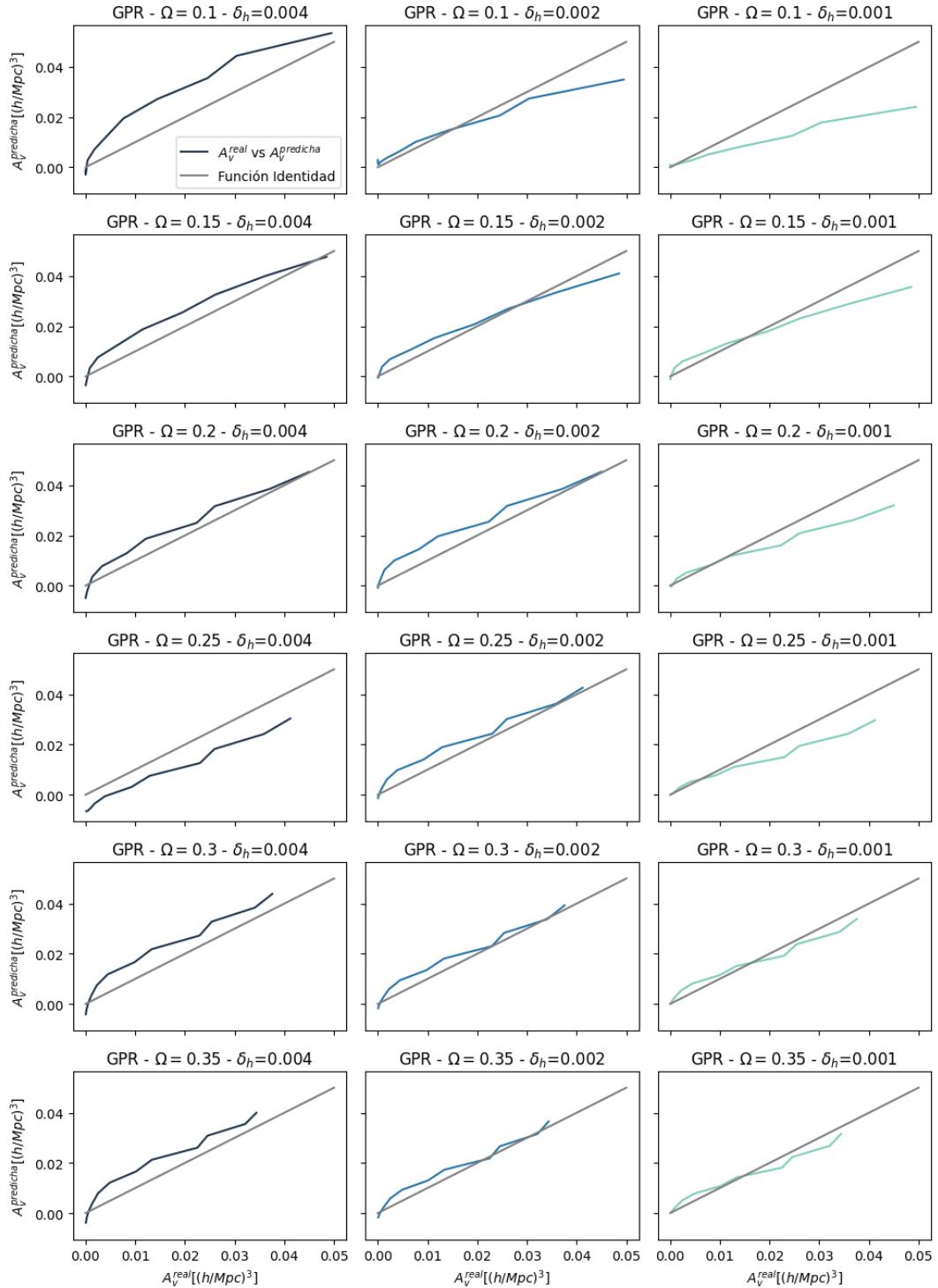
**Figura C.1:** VSF en materia simuladas vs VSF en materia predichas con un GPR, para  $\Omega_m = [0,1;0,5]$  con un paso de  $\Omega = 0,025$ .



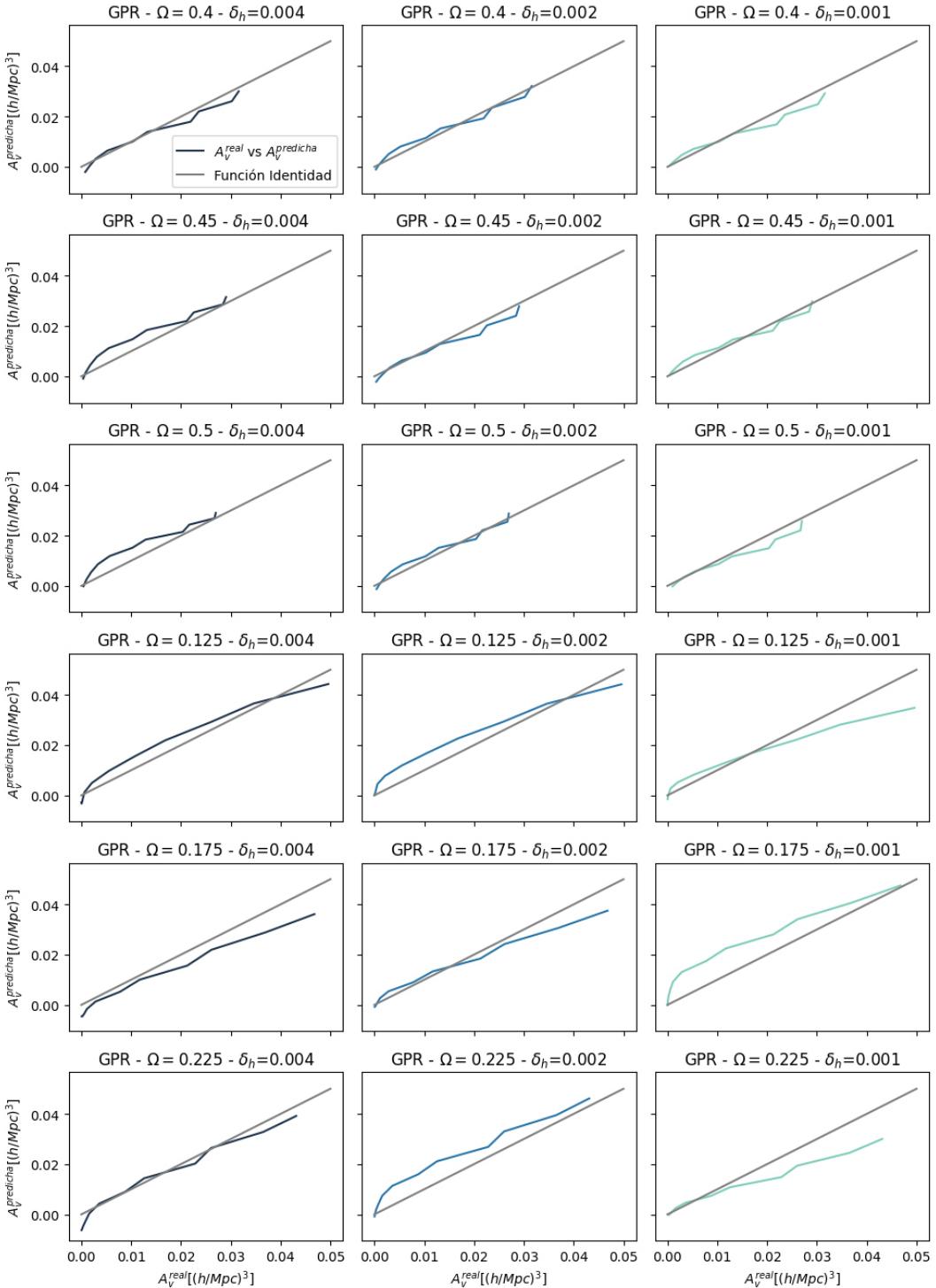
**Figura C.2:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un GPR, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



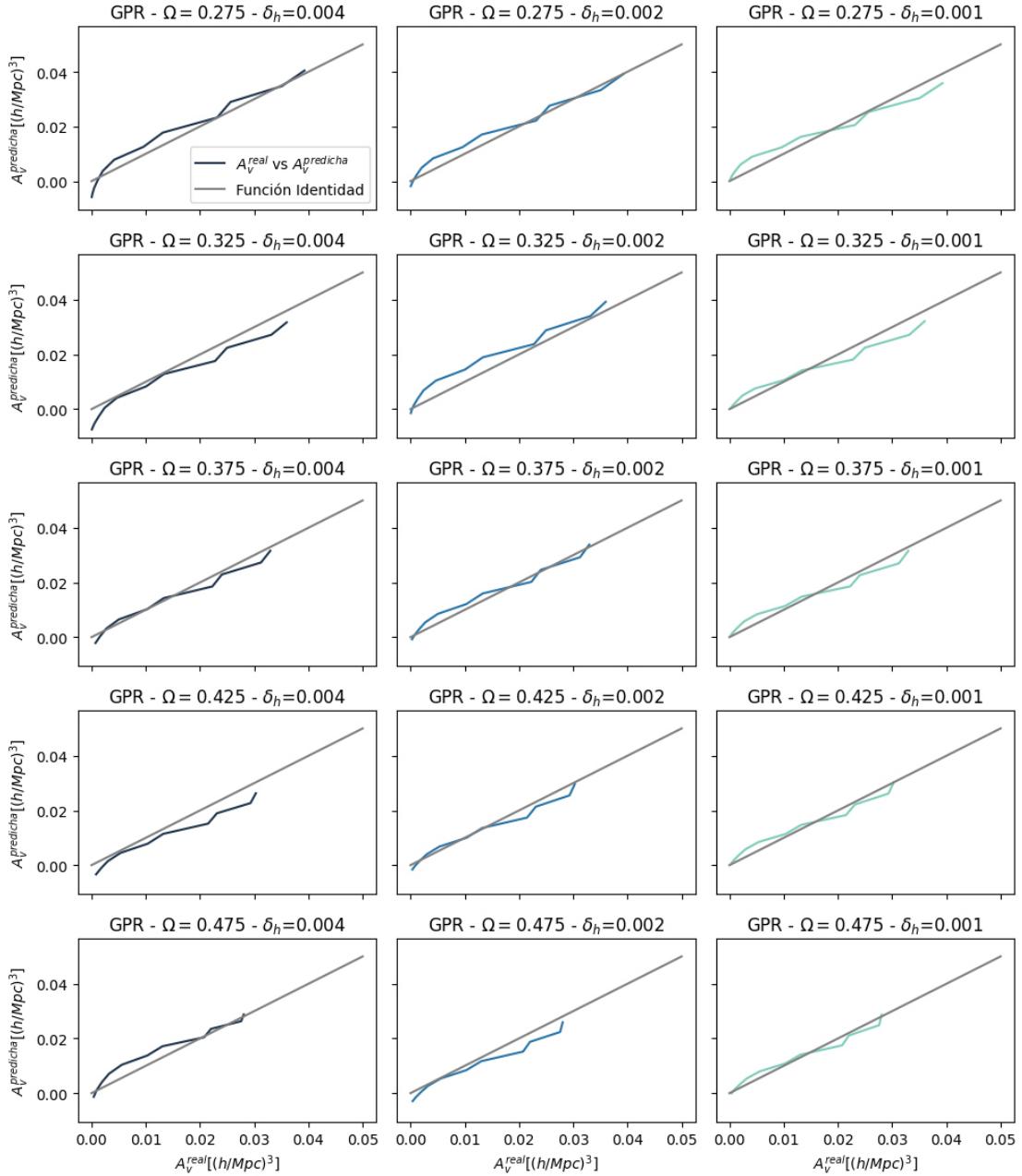
**Figura C.3:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un GPR, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



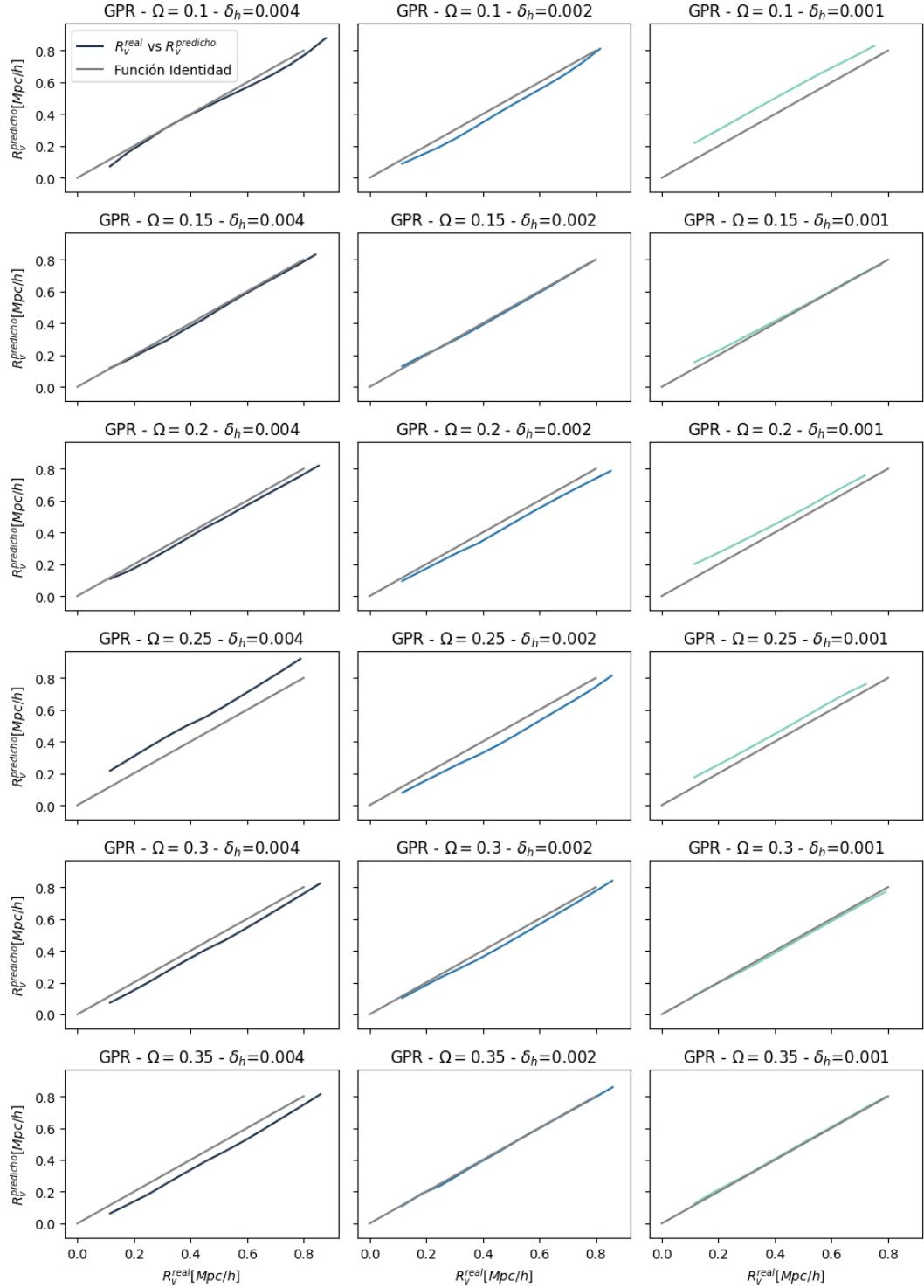
**Figura C.4:** Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



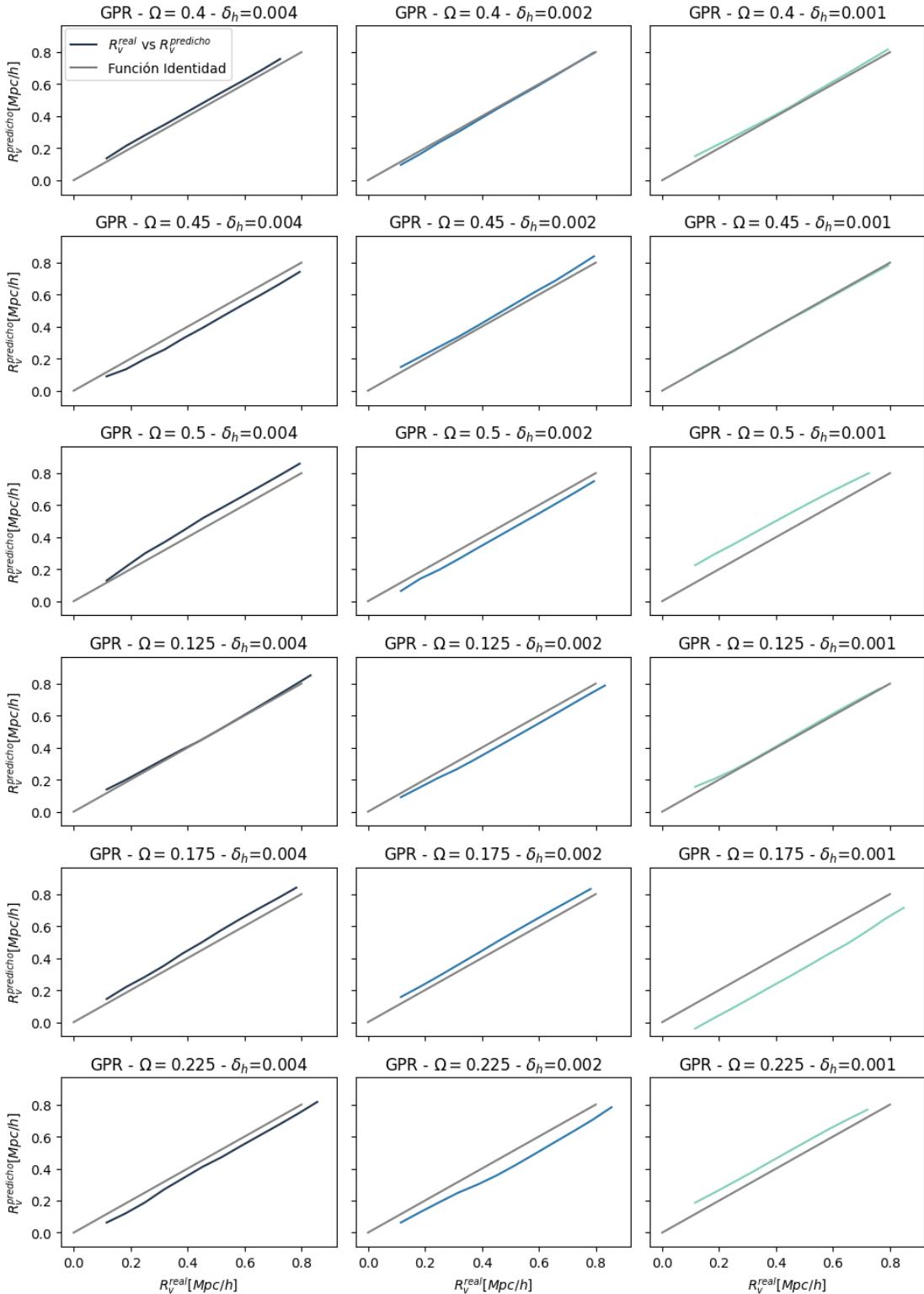
**Figura C.5:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



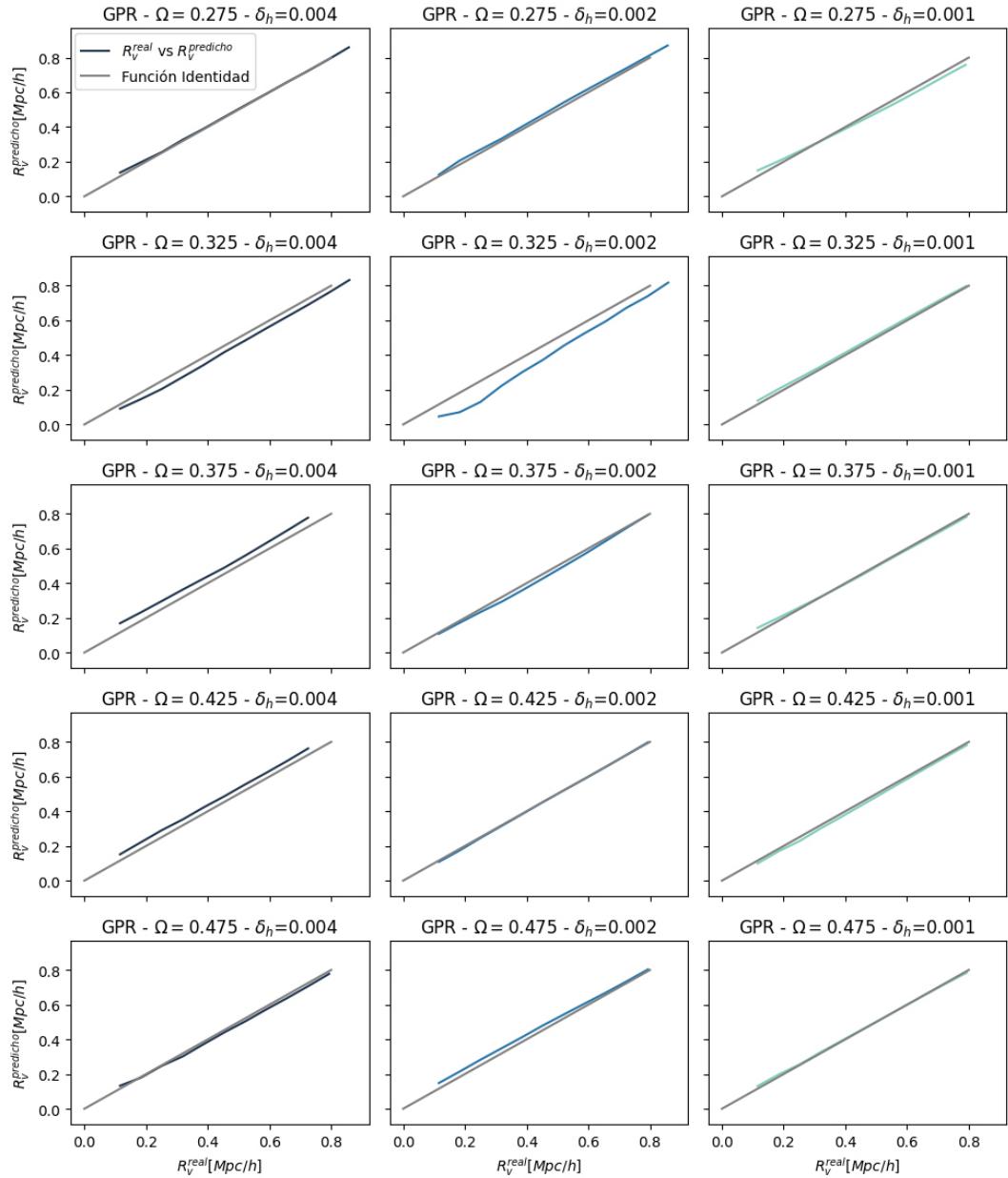
**Figura C.6:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



**Figura C.7:** Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



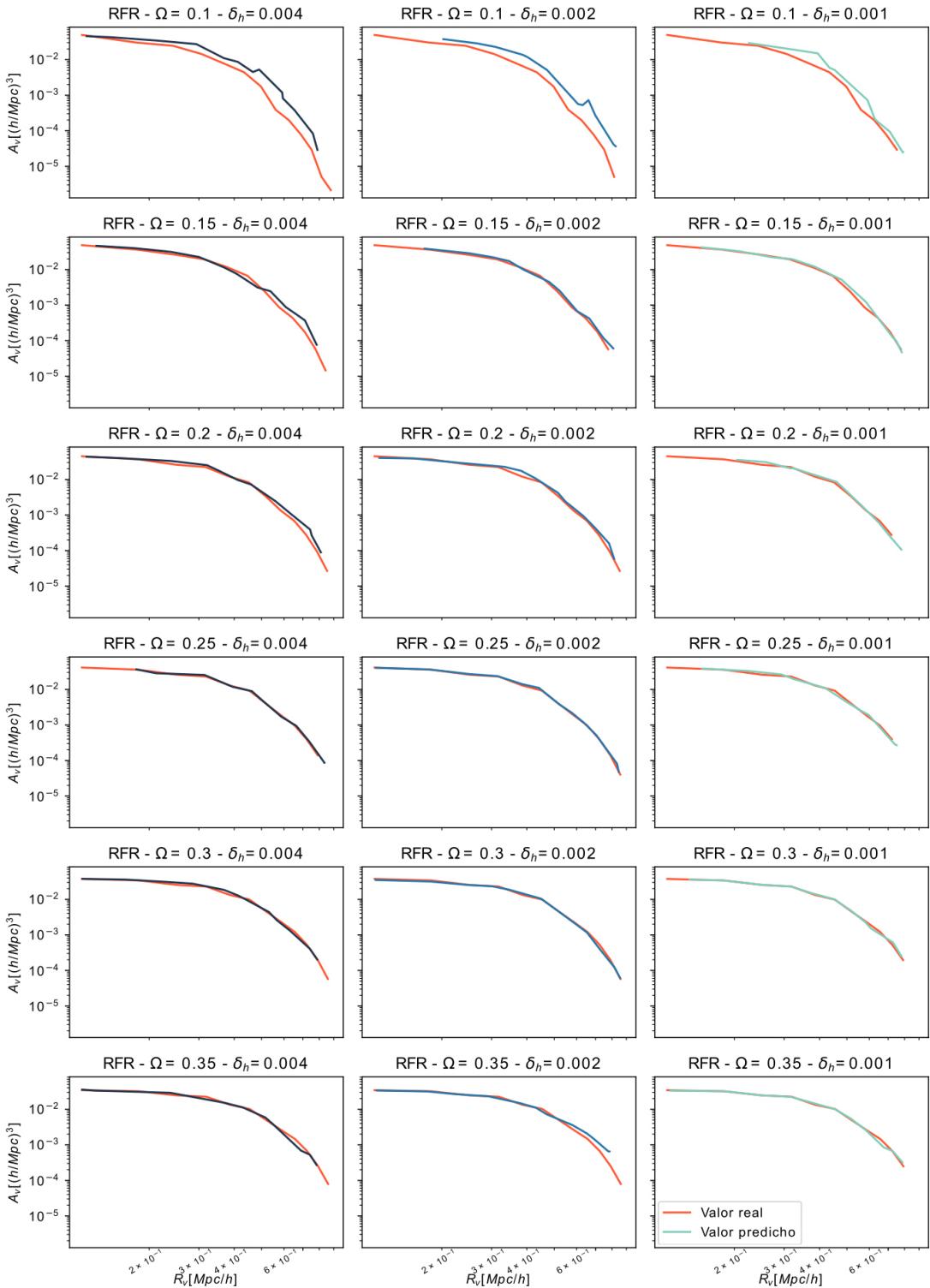
**Figura C.8:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



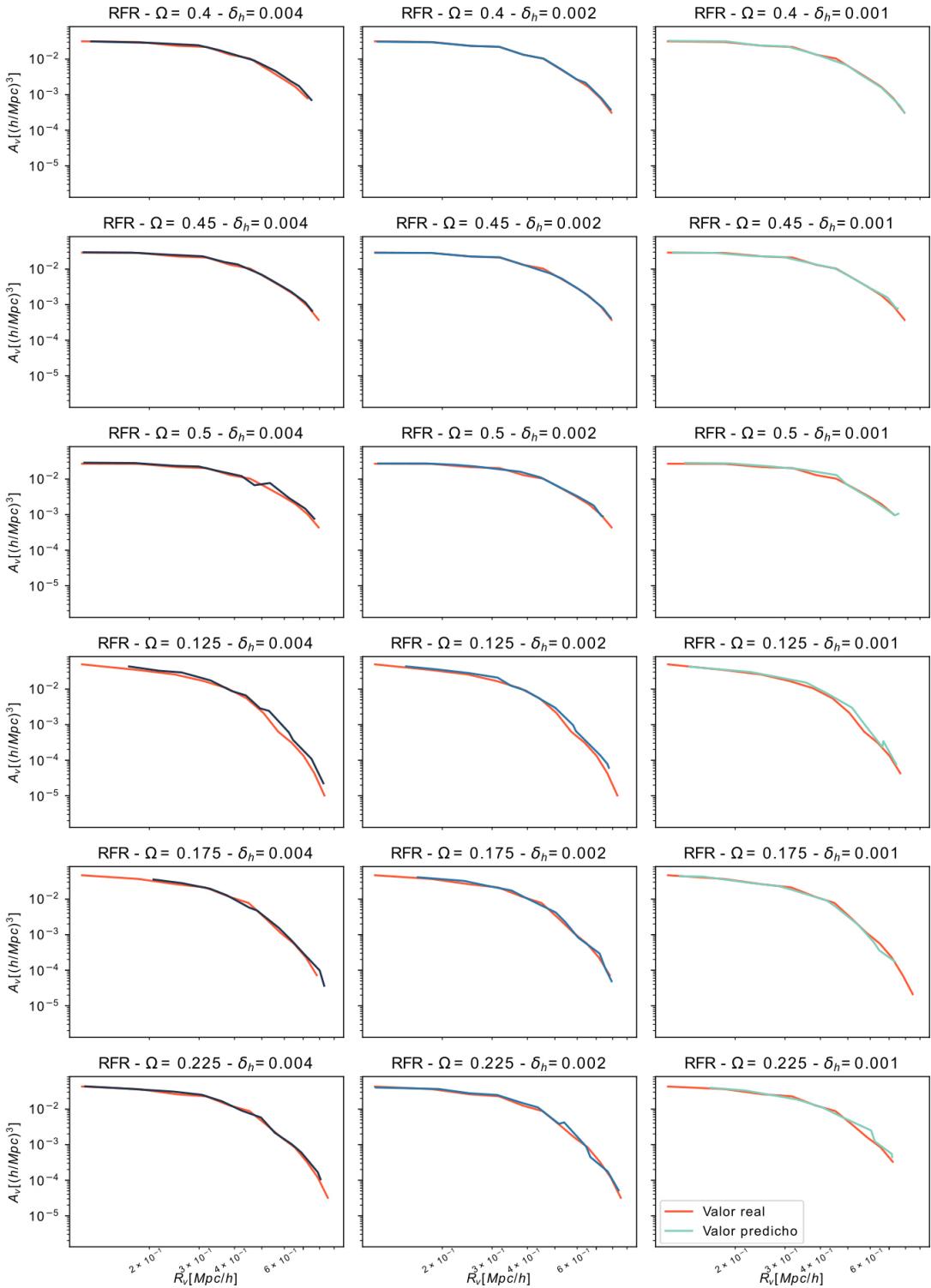
**Figura C.9:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Gaussian Process Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.

## **Apéndice D**

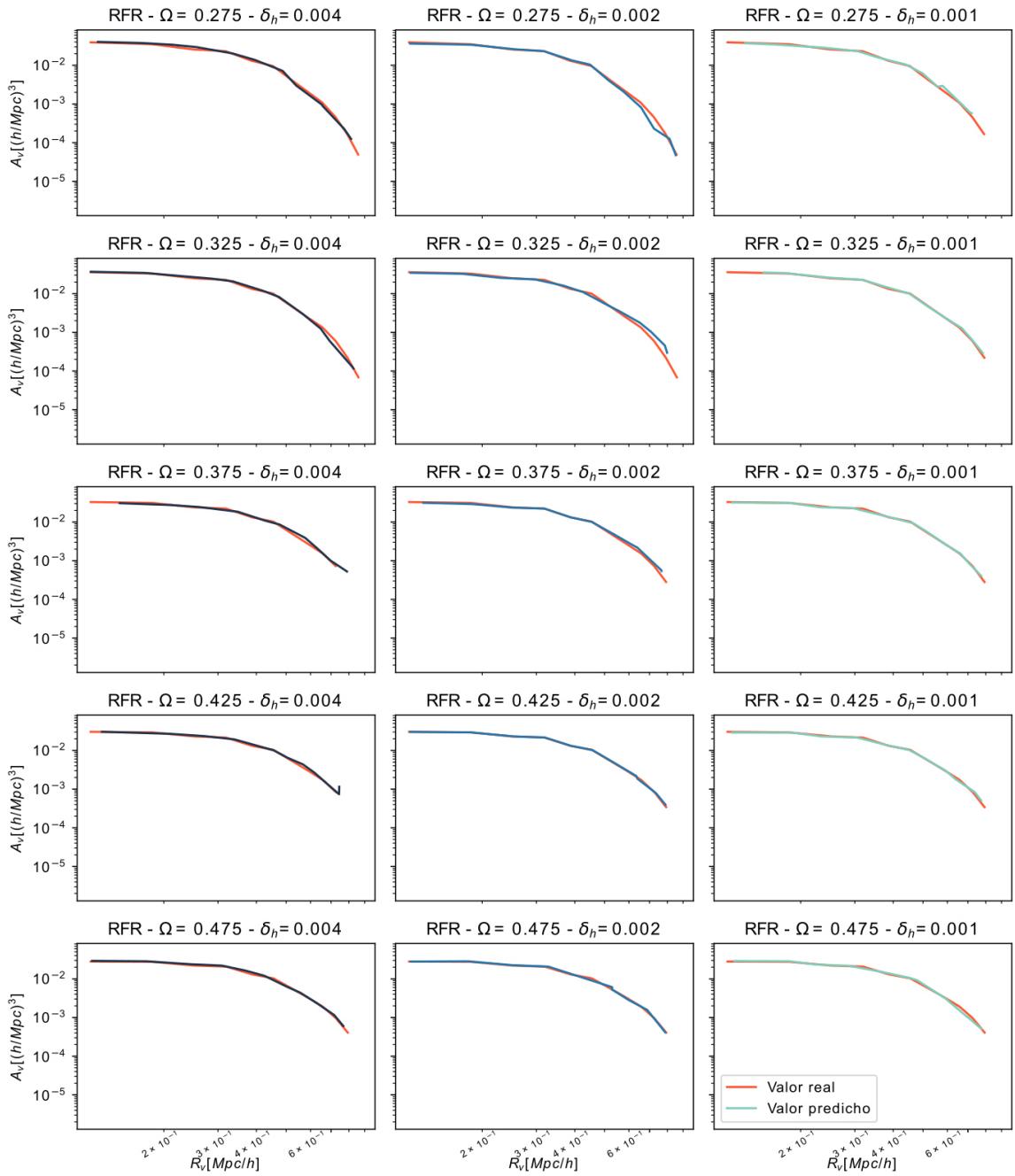
### **Random Forest Regressor**



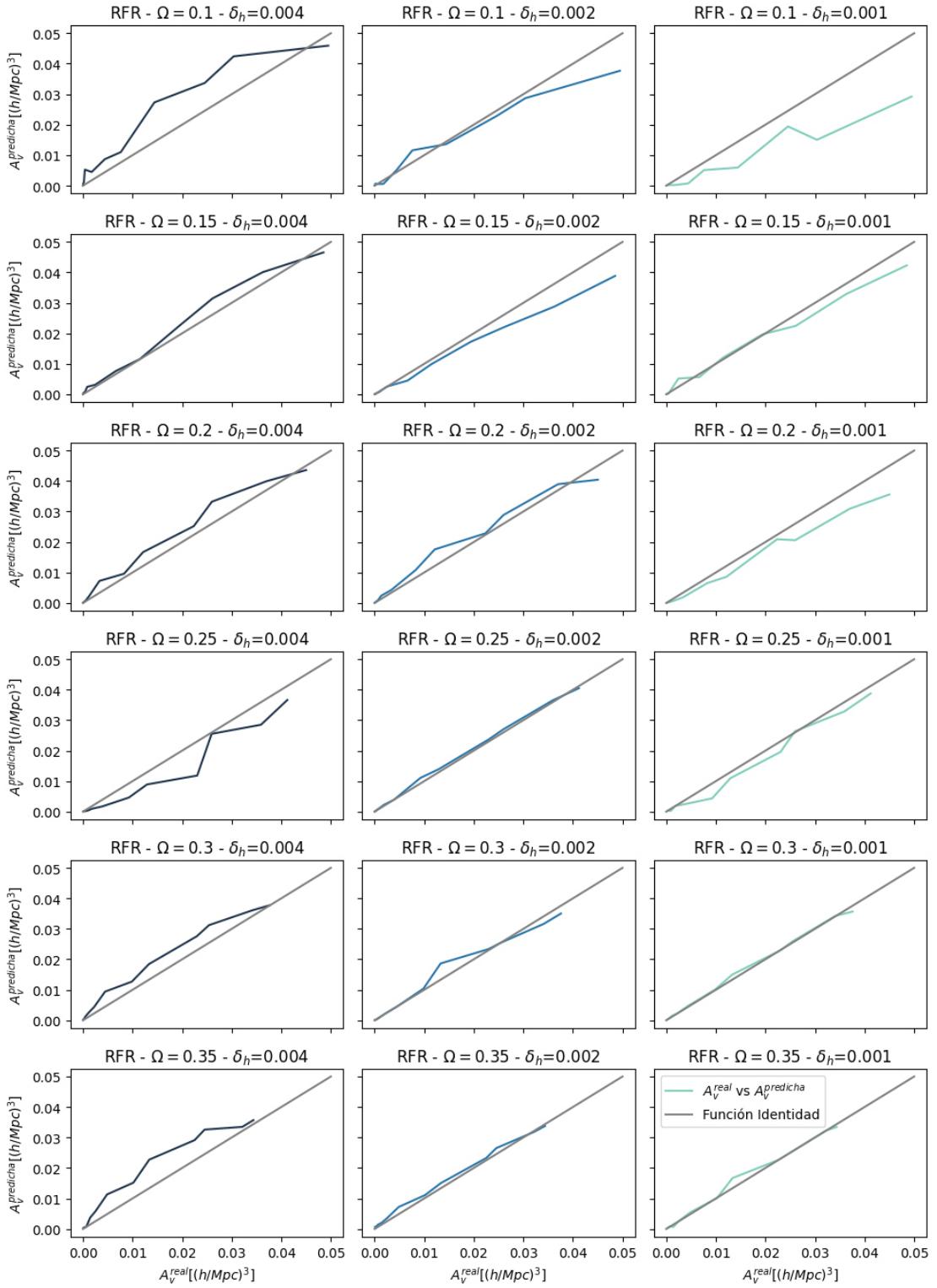
**Figura D.1:** VSF en materia simuladas vs VSF en materia predichas con un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



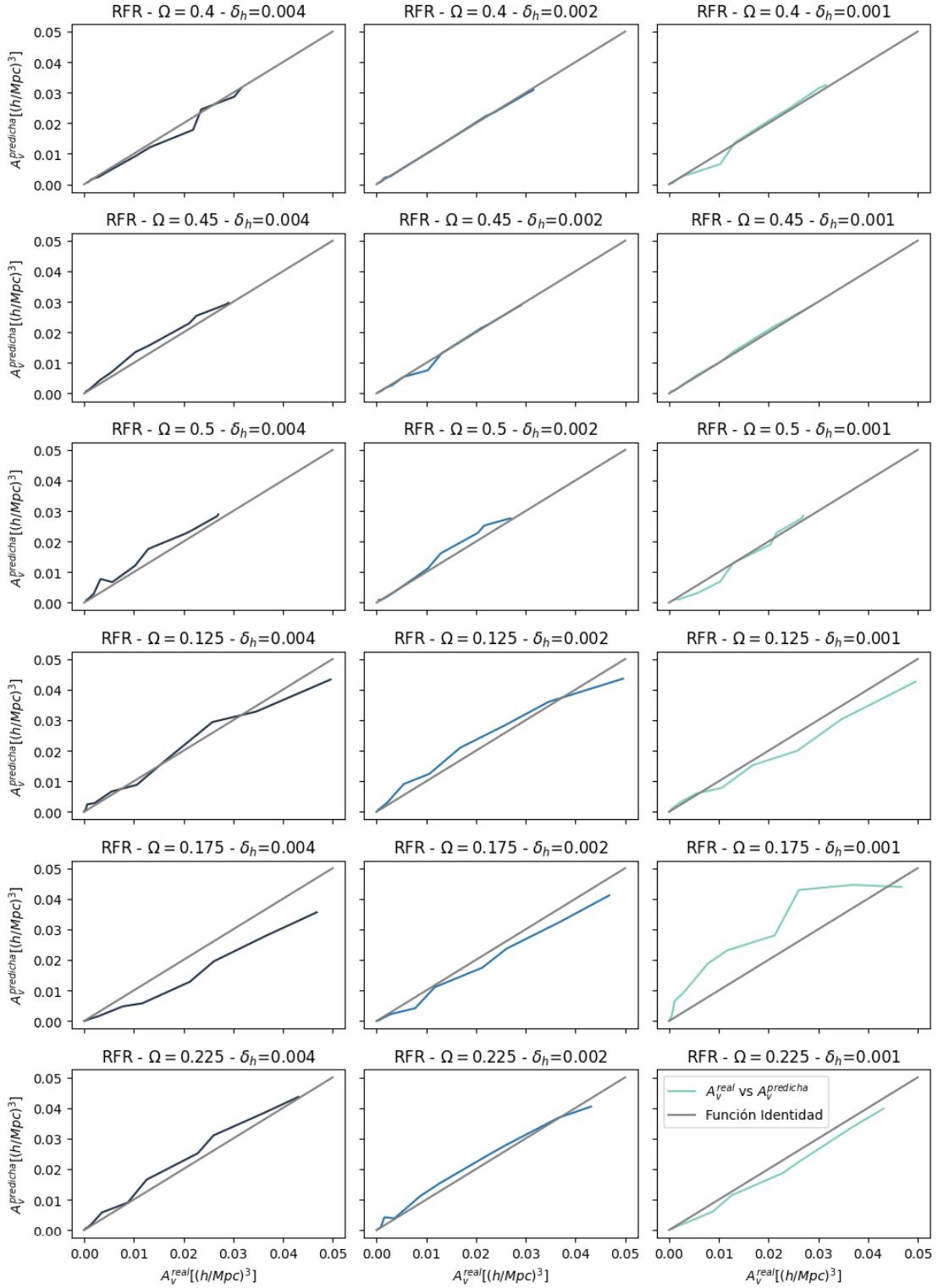
**Figura D.2:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



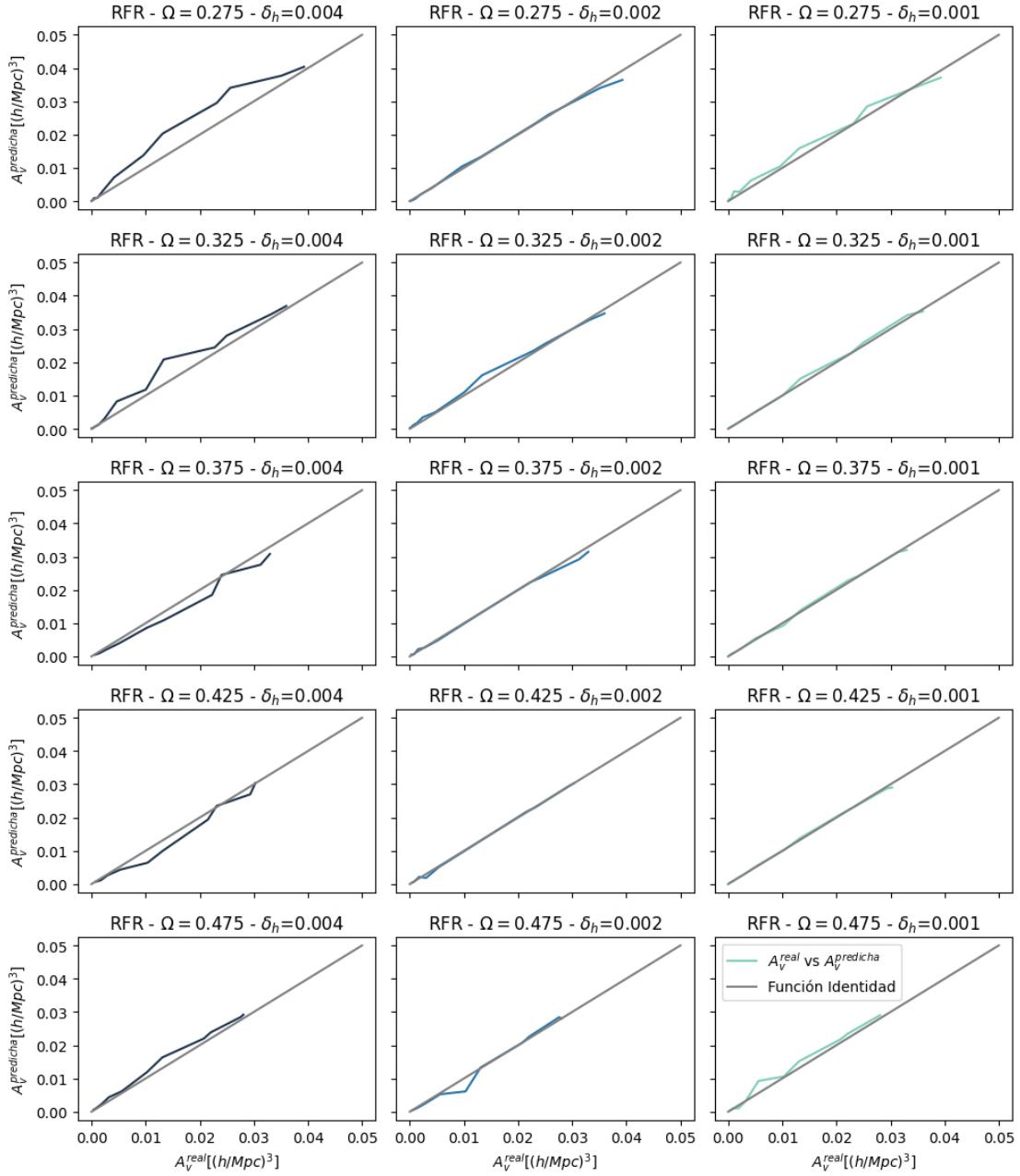
**Figura D.3:** (cont.) VSF en materia simuladas vs VSF en materia predichas con un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ .



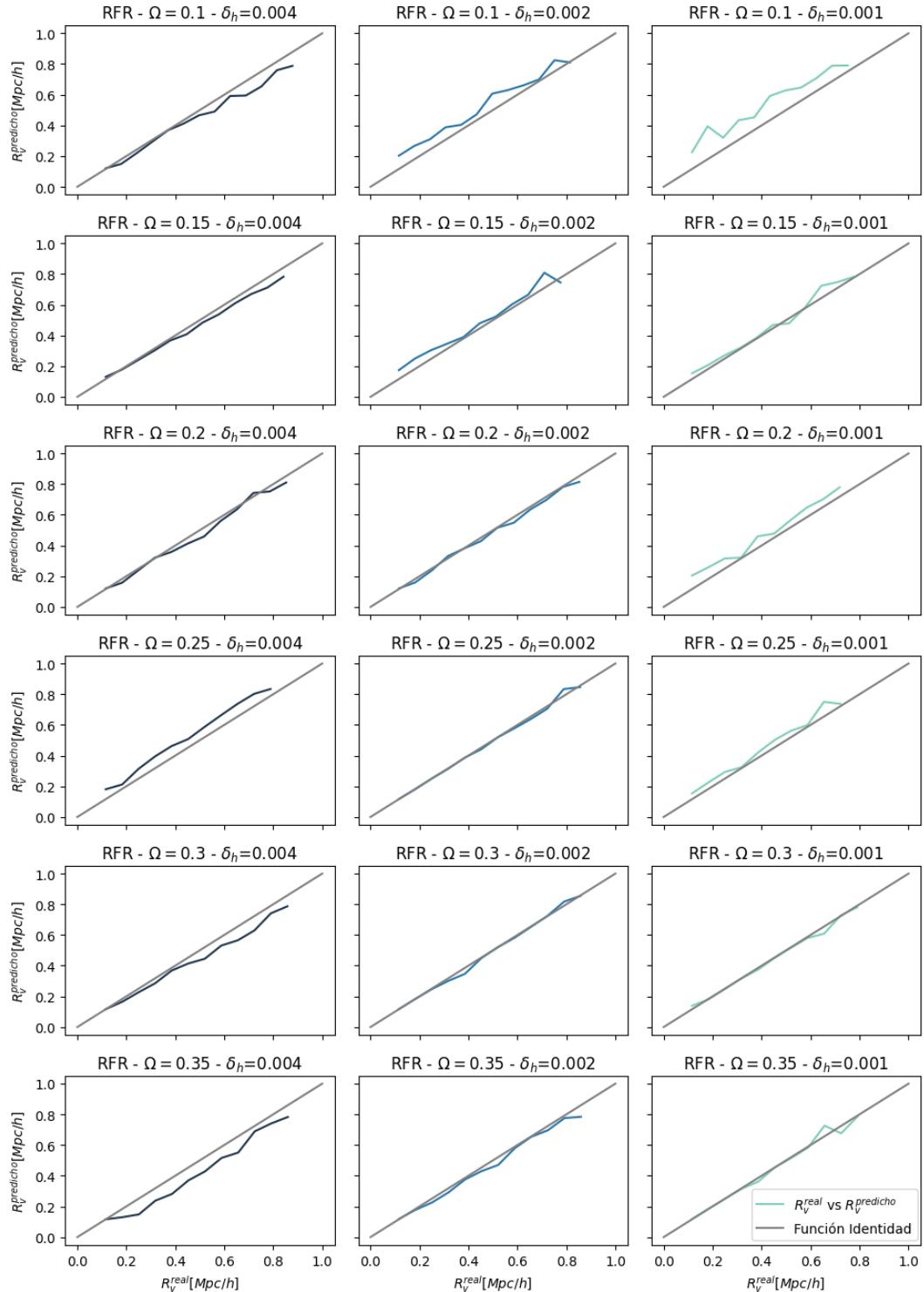
**Figura D.4:** Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad



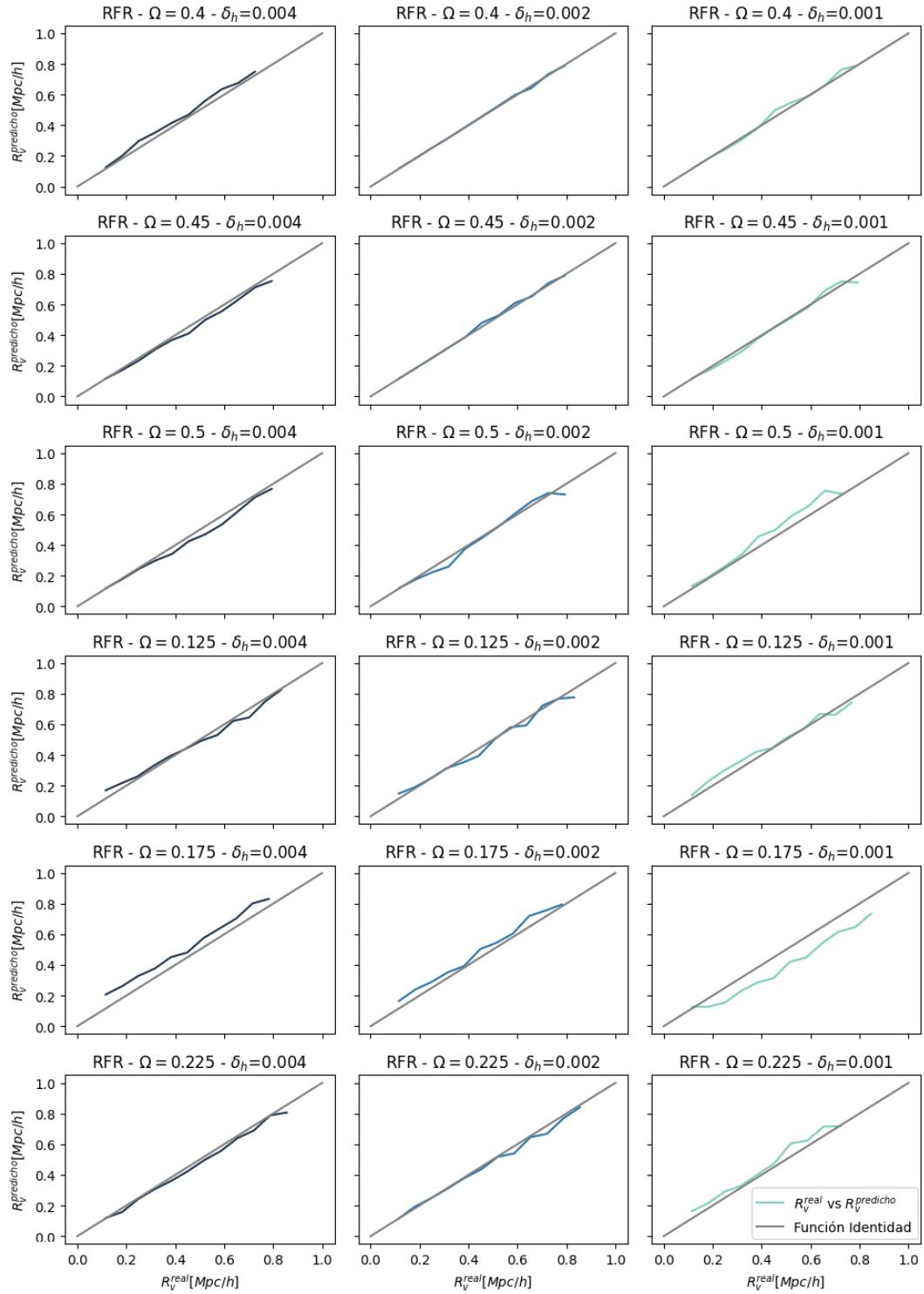
**Figura D.5:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



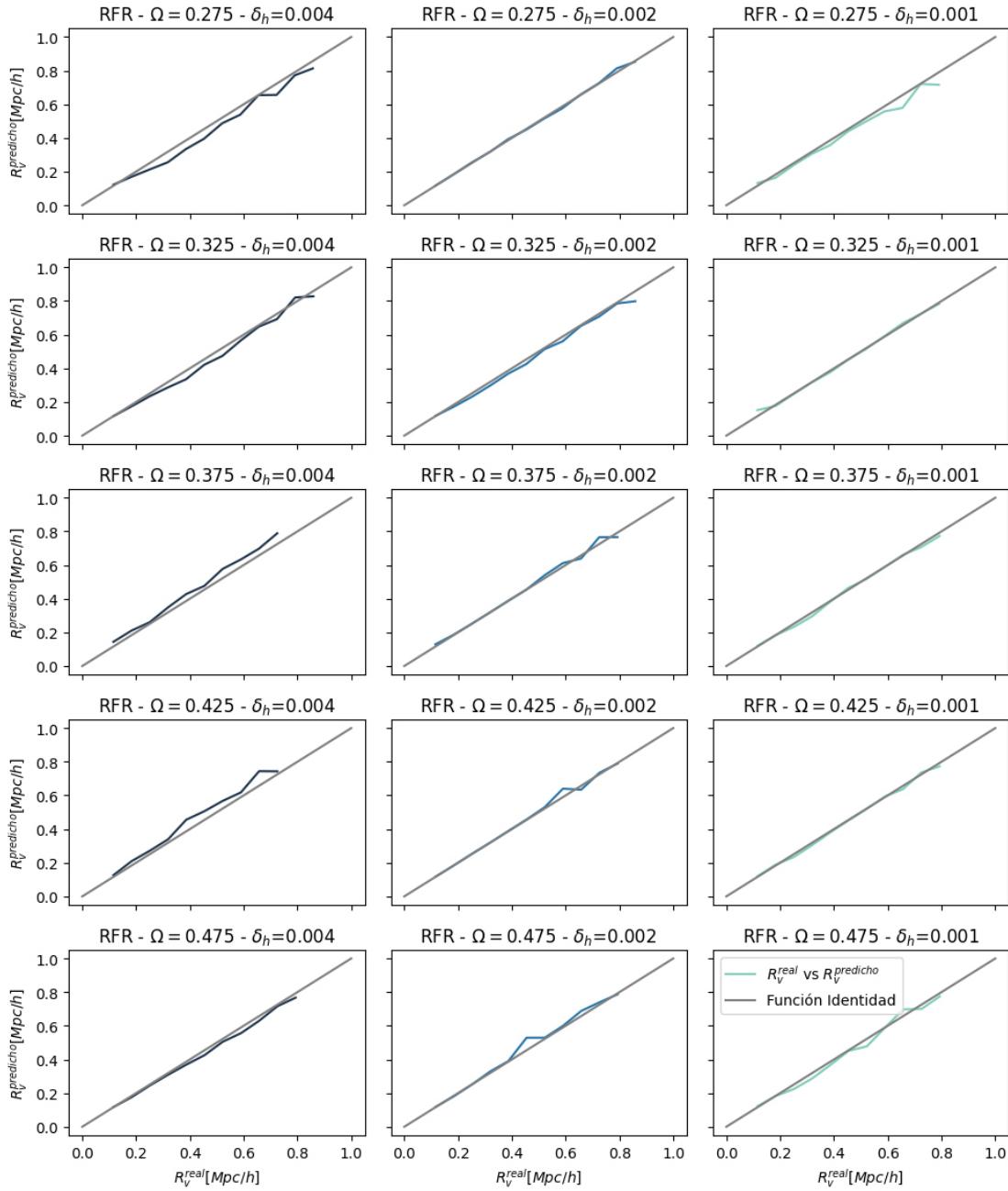
**Figura D.6:** (cont.) Gráficas comparativas del conteo de vacíos simulados en materia vs el conteo de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



**Figura D.7:** Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



**Figura D.8:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.



**Figura D.9:** (cont.) Gráficas comparativas del radio de vacíos simulados en materia vs el radio de vacíos en materia predicho por un Random Forest Regressor, para  $\Omega_m = [0,1; 0,5]$  con un paso de  $\Omega = 0,025$ . En gris una función identidad para visualizar mejor la proporcionalidad en el gráfico.

# Bibliografía

- [Adler et al., 1995] Adler, R. J. et al. (1995). Vacuum catastrophe: An elementary exposition of the cosmological constant problem. *American Journal of Physics*, 63.
- [Ashby, 1949] Ashby, W. R. (1949). The electronic brain. *Radio Electronics*.
- [Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, Volume 13*.
- [Berrar, 2019] Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*.
- [Bertschinger, 1998] Bertschinger, E. (1998). Simulations of structure formation in the universe. *Annual Review of Astronomy and Astrophysics, Volume 36*.
- [Blumenthal et al., 1984] Blumenthal, G. R., Faber, S. M., Primack, J. R., and Rees, M. J. (1984). Formation of galaxies and large-scale structure with cold dark matter. , 311:517–525.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning, Volume 45*.
- [Cabral, 2019] Cabral, J. (2019). *Análisis y diseño de procesos de minería de datos astrofísicos sobre catálogos fotométricos múltiple época*. PhD thesis, Universidad Nacional de Rosario.

[Colberg et al., 2005] Colberg, J. M. et al. (2005). Voids in a  $\lambda$ cdm universe. *Monthly Notices of the Royal Astronomical Society, Volume 360, Issue 1.*

[Contarini et al., 2019] Contarini, S. et al. (2019). Cosmological exploitation of the size function of cosmic voids identified in the distribution of biased tracers. *Oxford University Press Monthly Notices of the Royal Astronomical Society, Volume 488, Issue 3.*

[Correa, 2021] Correa, C. M. (2021). *Vacíos cósmicos como laboratorios cosmológicos.* PhD thesis, Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba.

[Criminisi et al., 2011] Criminisi, A. et al. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Computer Graphics and Vision, Vol. 7, Nos. 2–3.*

[Devore, 2011] Devore, J. L. (2011). *Probability & Statistics.* Brooks/Cole, Cengage Learning.

[Einstein, 1917] Einstein, A. (1917). Kosmologische betrachtungen zur allgemeinen relativitätstheorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften.*

[Eisenstein, 2005] Eisenstein, D. (2005). Dark energy and cosmic sound. *New Astronomy Reviews, 49.*

[Friedman et al., 2009] Friedman, J. H. et al. (2009). *The Elements of Statistical Learning.* Springer.

[Friedmann, 1922] Friedmann, A. (1922). Über die krümmung des raumes. *Zeitschrift für Physik, Volume 10.*

[Frieman et al., 2008] Frieman, J. A. et al. (2008). Dark energy and the accelerating universe. *Annual Review of Astronomy and Astrophysics, Volume 46.*

- [Furlanetto and Piran, 2006] Furlanetto, S. R. and Piran, T. (2006). The evidence of absence: galaxy voids in the excursion set formalism. *Monthly Notices of the Royal Astronomical Society, Volume 366, Issue 2, pp. 467-479.*
- [Glover et al., 2014] Glover, S. C. et al. (2014). Atomic, molecular and optical physics in the early universe: From recombination to reionization. *Advances in Atomic, Molecular, and Optical Physics, 63.*
- [Goliath et al., 2001] Goliath, M. et al. (2001). Supernovae and the nature of the dark energy. *Astronomy & Astrophysics, Volume 380, Number 1.*
- [Hey et al., 2009] Hey, T. et al. (2009). *The Fourth Paradigm*. Microsoft Research.
- [Hubble, 1929] Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences.*
- [Ichiki and Sumiya, 2022] Ichiki, K. and Sumiya, K. (2022). Measuring the cosmological density field twice: A novel test of dark energy using cmb quadrupole. *Physical Review D, Volume 105, Issue 6.*
- [Jennings et al., 2013] Jennings, E. et al. (2013). The abundance of voids and the excursion set formalism. *Monthly Notices of the Royal Astronomical Society, Volume 434, Issue 3.*
- [Karpatne et al., 2016] Karpatne, A. et al. (2016). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering, 29.*
- [Lemaître, 1931] Lemaître, A. G. (1931). The expanding universe. *Monthly Notices of the Royal Astronomical Society, Volume 91, Issue 5.*
- [Mickaelian, 2017] Mickaelian, A. M. (2017). Astronomical surveys and big data. *Open Astronomy.*

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

[Murphy, 2012] Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. The MIT Press.

[NASA, 2024] NASA (2024). Large scale structures. <https://science.nasa.gov/universe/galaxies/large-scale-structures>.

[Padilla et al., 2005] Padilla, N. D. et al. (2005). Spatial and dynamical properties of voids in a  $\Lambda$ CDM universe. *Monthly Notices of the Royal Astronomical Society, Volume 363, Issue 3*.

[Paz et al., 2023] Paz, D. J. et al. (2023). Guess the cheese flavour by the size of its holes: A cosmological test using the abundance of popcorn voids. *Monthly Notices of the Royal Astronomical Society, 522*.

[Peebles, 1982] Peebles, P. J. E. (1982). Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *Astrophysical Journal, Part 2 - Letters to the Editor, vol. 263*.

[Peebles, 1982] Peebles, P. J. E. (1982). Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. , 263:L1–L5.

[Peebles, 1993] Peebles, P. J. E. (1993). *Principles of physical cosmology*. Princeton University Press.

[Penzias and Wilson, 1965] Penzias, A. A. and Wilson, R. W. (1965). A measurement of excess antenna temperature at 4080 mc/s. *Astrophysical Journal, vol. 142*.

[Perlmutter et al., 1998] Perlmutter, S. et al. (1998). Discovery of a supernova explosion at half the age of the universe and its cosmological implications. *Nature, Volume 391*.

[Planck Collaboration, 2020] Planck Collaboration (2020). Planck 2018 results. VI. Cosmological parameters. , 641:A6.

- [Pollina et al., 2017] Pollina, G. et al. (2017). On the linearity of tracer bias around voids. *Monthly Notices of the Royal Astronomical Society, Volume 469, Issue 1*.
- [Raschka and Mirjalili, 2017] Raschka, S. and Mirjalili, V. (2017). *Python Machine Learning, 2° ed.* Packt Publishing.
- [Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* The MIT Press.
- [Riess et al., 1998] Riess, A. G. et al. (1998). Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal, Volume 116, Issue 3*.
- [Ruiz et al., 2015] Ruiz, A. N. et al. (2015). Clues on void evolution iii: Structure and dynamics in void shells. *Monthly Notices of the Royal Astronomical Society, Volume 448, Issue 2*.
- [Ryden and Melott, 1996] Ryden, B. S. and Melott, A. L. (1996). Voids in real space and in redshift space. *Astrophysical Journal, Volume 470*.
- [Schmidt et al., 2001] Schmidt, J. D. et al. (2001). The size and shape of voids in three-dimensional galaxy surveys. *The Astrophysical Journal, Volume 546, Number 2*.
- [Schneider, 2006] Schneider, P. (2006). *Extragalactic cosmology.* Springer.
- [Scikit-Learn, 2024a] Scikit-Learn (2024a). Scikit-learn: Regression metrics.  
[https://scikit-learn.org/stable/modules/model\\_evaluation.html#regression-metrics](https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics).
- [Scikit-Learn, 2024b] Scikit-Learn (2024b). sklearn.gaussian\_process.gaussianprocessregressor.  
[https://scikit-learn.org/stable/modules/generated/sklearn.gaussian\\_process.GaussianProcessRegressor.html](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html).

- [Sheth and Lemson, 1999] Sheth, R. and Lemson, G. (1999). Biasing and the distribution of dark matter haloes. *Monthly Notices of the Royal Astronomical Society, Volume 304, Issue 4.*
- [Sheth and van de Weygaert, 2004] Sheth, R. and van de Weygaert, R. (2004). A hierarchy of voids: Much ado about nothing. *MNRAS, Vol. 350, No. 2, p. 517–538.*
- [Spergel et al., 2007] Spergel, D. et al. (2007). Three-year wilkinson microwave anisotropy probe (wmap) observations: implications for cosmology. *The Astrophysical Journal Supplement Series, 170.*
- [Sun et al., 2018] Sun, Z. et al. (2018). Big data with ten big characteristics. *ICBDR '18: Proceedings of the 2nd International Conference on Big Data Research.*
- [Sutter et al., 2014] Sutter, P. M. et al. (2014). Sparse sampling, galaxy bias, and voids. *Monthly Notices of the Royal Astronomical Society, Volume 442.*
- [Thompson, 2020] Thompson, L. A. (2020). *The discovery of cosmic voids.* Cambridge University Press.
- [Tinker and Conroy, 2009] Tinker, J. L. and Conroy, C. (2009). The void phenomenon explained. *The Astrophysical Journal, Volume 691, Issue 1.*
- [Tsujikawa, 2018] Tsujikawa, S. (2018). *The Encyclopedia of Cosmology, Volumen 3.* World Scientific.
- [Wang, 2023] Wang, J. (2023). n intuitive tutorial to gaussian process regression. *Computing in Science & Engineering.*
- [Wechsler and Tinker, 2018] Wechsler, R. H. and Tinker, J. L. (2018). The connection between galaxies and their dark matter halos. *Annual review of astronomy and astrophysics, Vol. 56:435-487.*
- [Zwicky, 1933] Zwicky, F. (1933). Die rotverschiebung von extragalaktischen nebeln. *Helvetica Physica Acta, Vol. 6.*