



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN ASTROFÍSICA

ASTROFÍSICA TEÓRICA

MODELOS DE FOTOIONIZACIÓN CON TRES PARÁMETROS LIBRES DE
LAS REGIONES HII DEL SURVEY CALIFA

TESIS

PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS (ASTROFÍSICA)

PRESENTA:
CARLOS CRISPÍN ESPINOSA PONCE

DR. CHRISTOPHE MORISSET
INSTITUTO DE ASTRONOMÍA

CIUDAD DE MÉXICO, ENERO 2018



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Índice general

Agradecimientos	XVII
Abstract	XXI
1. El medio interestelar	1
§1.1. Introducción	1
§1.2. Regiones HII	2
§1.2.1. Equilibrio de fotoionización de una nebulosa de hidrógeno	5
§1.2.2. Balance térmico	14
§1.3. Cálculo de propiedades físicas en nebulosas	19
§1.3.1. Temperaturas y densidades nebulares	19
§1.3.2. Cálculo de abundancias	22
2. CALIFA	29
§2.1. Introducción	29
§2.2. Detección de Regiones HII	30
§2.2.1. Ajuste del espectro	32
3. Modelos de fotoionización	39
§3.1. Introducción	39
§3.2. <i>Inputs</i> y <i>Outputs</i> de los códigos de fotoionización	40
§3.3. Modelado de regiones HII de CALIFA	41
§3.4. Modelos con tres parámetros libres	46
§3.4.1. Problema de O/H	46
§3.5. Método Genético	47
§3.5.1. Primera familia de modelos	53

§3.6. Análisis de cúmulos: <i>Mean Shift</i>	62
4. Resultados	89
5. Discusión de resultados y trabajo a futuro	103
Apéndices	111
A. Modelos finales para todas las regiones	113
Bibliografía	139

Índice de figuras

1.1. Nebulosa de Orión. Imagen:NASA/JPL-Caltech/STScI	3
1.2. Nebulosa planetaria Ojo de gato. Imagen:NASA, ESA, HEIC, and The Hubble Heritage Team (STScI/AURA)	4
1.3. Diagrama de estructura de una nebulosa. Se representa la estrella con un asterisco y el inicio del gas como la primera línea gruesa cerca de la estrella. La distancia al gas es R_{in} y el radio de Strömngren como R_s . La segunda línea negra representa el final del gas ionizado. Si la nebulosa tiene un radio delimitante menor al R_s se dice que la nebulosa está limitada por masa. Por el contrario si este radio es mayor que R_s entonces se dice que la nebulosa está limitada por radiación.	9
1.4. Secciones eficaces para HI, HeI, HeII, OI y NII, basado en Verner et al. (1996). Al incluir más elementos en las ecuación del equilibrio de fotoionización las secciones eficaces influirán dependiendo de que elementos se tomen en cuenta.	11

- 1.5. Diagrama para una nebulosa. a) Para una estrella con $T < 40000\text{K}$. Observamos que debido a la energía radiada por la estrella solo en sus alrededores puede ionizar una vez al hidrógeno y al helio. El gas más alejado se conforma de hidrógeno una vez ionizado y helio neutro. Esto se debe a que los iones que pueden ionizar al helio ya han sido absorbidos en las cercanías de la estrella. Esto crea dos zonas en el gas: la primera que se encuentra cerca del gas que se conforma principalmente de HII y HeII y otra más lejana de la estrella que está compuesta en su mayoría por HII y HeI. b) Para una estrella con $40000\text{K} < T < 10^5\text{K}$. A diferencia del caso a) este tipo de estrellas son capaces de ionizar todo el hidrógeno presente en el gas de la nebulosa sin crear las dos zonas. 12
- 1.6. Diagrama de transiciones entre niveles de [OIII], todos desde la configuración $2p^3$. Las líneas de emisión que se encuentran en el óptico son $\lambda 4363$, $\lambda 5007$ y $\lambda 4949$. Las otras líneas mostradas se encuentran en el infrarojo y en el ultravioleta. 20
- 1.7. Diagrama de transiciones entre niveles de [OII]. Las líneas que se encuentran en el rango óptico son $\lambda 3726$ y $\lambda 3720$ 21
- 1.8. Variación de los cocientes de línea de [OII] y [SII] con respecto a la densidad. Aproximadamente entre una densidad electrónica de 10^2 y 10^4 tenemos la mayor variación de las intensidades de línea. En este rango es donde podemos usar estas líneas para calcular la densidad. Fuera de este rango la intensidad no varía con respecto a la densidad. (Osterbrock & Ferland, 2006) 22
- 1.9. Calibración hecha por Marino et al. (2013) donde se muestra las distintas regiones HII usadas y otras calibraciones de otros trabajos. Los colores representan con qué línea se obtuvo la temperatura. Los símbolos azules son las regiones HII donde se calculó su temperatura con la línea [OIII] $\lambda 4363$. Los símbolos naranjas indican las regiones donde se usó la línea [NII] $\lambda 5755$. En las regiones que se representan con símbolos rojos se usó la línea [SIII] $\lambda 6312$ y en verde se representan las regiones donde se usaron las líneas [OIII] $\lambda 4363$ y [NII] $\lambda 5755$ 25

1.10. Degeneración del cociente $[OIII]5007/H\beta$ con respecto a O/H (Stasińska, 2002). Este comportamiento se debe a que a bajas metalicidades estos cocientes incrementan conforme la metalicidad aumenta. Pero a altas metalicidades disminuyen debido a el incremento del enfriamiento debido a las líneas infrarojas. Esto causa una disminución en la temperatura por debajo del límite para una línea de excitación colisional. 26

1.11. O/H vs $[OIII]/H\beta$ con N/O y logU del grid de modelos BOND (Vale Asari et al., 2016). Las gráficas superiores son todos los modelos disponibles. La gráfica inferior izquierda nos muestra los modelos que tienen un valor de $\log U = -3 \pm 0.1$ y el N/O es libre. En la gráfica inferior derecha son los modelos con $N/O = -1.0 \pm 0.3$ y logU libre. Es decir, los modelos inferiores es un acercamiento a las gráficas superiores. 27

2.1. Imagen de una galaxia obtenida por CALIFA. Imagen:CALIFA Survey . . . 30

2.2. Input y salida de HIIEXPLORER. Izq.)Mapa de $H\alpha$ de la galaxia NGC5947 input de HIIEXPLORER. Der.)Regiones HII detectadas con HIIEXPLORER de la galaxia NGC5947. Cada punto de color es candidato a ser una región HII. 32

3.1. Identificación de poblaciones estelares: OB y HOLMES para las regiones de gas ionizadas de CALIFA. El valor $Q_{0/1} = Q(H^0)/Q(He^0)$, el cociente del número de fotones ionizantes de H^0 y He^0 está relacionado con la propoción de fotones ionizantes que vienen de estrellas OB del número total de fotones ionizantes $Q(H^0)$ (Morisset et al., 2016) 45

3.2. Flujo básico de un algoritmo genético. Dada una población inicial se determina que tan bien ajustan algún parámetro o relación dada. Se seleccionan los mejores modelos que cumplan este ajuste. Se reproducen de manera sexual o asexual y por último se realiza una mutación la cual puede ser nula. A partir de este punto se repite el proceso hasta la convergencia. 49

3.3. Algoritmo genético aplicado a los espacios P y espacio O. *Izquierda*: De una malla de puntos en el espacio P se corren modelos con Cloudy para generar los puntos del espacio O. *Medio*: De los datos en el espacio O se seleccionan los puntos dentro de un radio dado y a su vez se identifican los puntos correspondientes en el espacio P. *Derecha*: De los puntos seleccionados en el espacio se generan crías creando una nueva malla para repetir todo el proceso. 52

- 3.4. Comportamiento del parámetro σ . Iniciando desde un valor $\sigma_0 = 0.5$ decrece hasta la generación 7 en el cual toma un valor constante de 0.05. Por lo tanto nuestros parámetros tendrán un rango mayor en el cual poder cambiar sus parámetros en las primeras generaciones. A partir de la séptima generación este rango de variación se restringe a un valor pequeño y constante suponiendo que estamos cerca de converger a la solución. 54
- 3.5. Malla inicial de la primera familia de modelos. En la parte superior el diamante representa la observación de CALIFA que se quiere reproducir. En las dos gráficas inferiores se tienen los puntos de la malla inicial. Se presentan algunos ajustes de la literatura: líneas azules (sólida y rayada) son ajustes obtenidos por Morisset et al. (2016). Línea negra punteada es el ajuste obtenido por Pérez-Montero (2014). La línea verde rayada es el ajuste obtenido por Dopita & Evans (1986) y la línea morada de punto y guión es el ajuste echo por Dopita et al. (2006). La línea cyan rayada es el ajuste realizado por Vila Costas & Edmunds (1993). La línea verde punteada es el ajuste hecho por Dopita et al. (2013) y la línea morada rayada es el ajuste hecho por Pilyugin et al. (2012). En ambas gráficas se muestra el ajuste para la abundancia de oxígeno (línea azul sólida) hecha por Marino et al. (2013). Este último ajuste se utilizó en el trabajo anterior por lo que se generaron un mayor número de puntos alrededor de esta línea. 56
- 3.6. Generación 1 de la primera familia de modelos. En puntos grises se representa los puntos de la generación anterior, es decir, el grid inicial de la familia. En puntos negros se muestran los modelos escogidos para ser progenitores. Los puntos rojos representas las crías generadas a partir de los modelos progenitores. Las líneas son las mismas mostradas en la Fig. 3.5. 58

3.7. Todas las generaciones de la primera familia de modelos de la región 2. Observamos como las primeras generaciones, representadas por tonos rojizos, abarcan una gran zona del espacio de parámetros. Conforme se realizan generaciones observamos como los modelos en el espacio O empiezan a converger hacia la observación (modelos con tonos azules). Pero en el espacio P, sus parámetros correspondientes, igualmente empiezan a converger a una pequeña zona. Esto define un conjunto solución lo cual en este caso solo es un conjunto de los dos esperados. Además el único conjunto solución encontrado se encuentra lejos de los ajustes y resultados de la literatura. Las líneas mostradas son las mismas que en la Fig. 3.5. 59

3.8. Representación de la relación entre el espacio de parámetros y el espacio de observables. Los espacios O y P se representan en una dimensión. Si la tolerancia con la cual se escogen modelos en el método genético es muy pequeña podemos dejar fuera una solución. Esto es representado con los pozos, si la tolerancia pasa uno de los dos pozos entonces se perdería una solución. 61

3.9. Relación entre el espacio de parámetros y el espacio de observables dividido en dos subespacios. Esperamos obtener dos conjuntos de soluciones por lo explicado al final del capítulo 1. Si dividimos el espacio y aplicamos el método genético en cada subespacio entonces obtendríamos los dos conjuntos solución. Esto debido a que la evolución del método genético de cada solución sería independiente una de la otra. 62

3.10. Agrupación de grupos de datos en cúmulos. Podemos ver que los puntos originales claramente para un ser humano están agrupados en dos cúmulos como se muestra en el caso b). Pero si lo requiriéramos podríamos formar subcúmulos de estos dos cúmulos. Esto se muestra en el caso c) y d) en el cual se dividen en cuatro y seis cúmulos, correspondientemente. (Tan et al., 2005). 63

- 3.11. Diferentes casos en la función bi valuada visto en el espacio P vs espacio O. En el panel superior observamos que cuando la solución esta lejos del punto máximo de la relación entre O/H y [OIII] tenemos dos soluciones bien definidas en el espacio de parámetros. En el panel medio se muestra que si estamos cerca del punto máximo en el espacio de parámetros las dos soluciones pueden superponerse una con la otra. El panel inferior muestra el caso en que estamos en el punto máximo donde solo tendremos un conjunto de soluciones. 67
- 3.12. Comparación de diferentes métodos de *clusterings* (Pedregosa et al., 2011). A pesar de que solo se representan los tres métodos considerados para nuestro proceso existen un gran número de algoritmos de identificación de cúmulos. Notemos que en cada uno de los paneles se muestra el tiempo de ejecución. Este tiempo varía de acuerdo a la cantidad de puntos como de la distribución de datos. En general vemos que el algoritmo *Mean Shift* es el que tarda más en identificar los cúmulos. 69
- 3.13. Proceso llevado a cabo por *Mean Shift*. Para cada punto de los datos define una ventana, representada por el círculo verde. Dentro de esta ventana se calculará donde se encuentra la mayor densidad de puntos. Se moverá la ventana, y por lo tanto el punto relacionado a ella hacia donde se encontrará la mayor densidad de puntos. Este movimiento se representa por la línea roja. Este proceso se repite hasta que dentro de la ventana la densidad de puntos es constante. En este punto lo definirá como el centro del cúmulo. Todos los puntos que lleguen al mismo centro o estén muy cercanos a este centro se clasificarán como miembros del cúmulo definido. 71
- 3.14. Encontrando centro de los cúmulos con el algoritmo *Mean Shift*. Observamos la distribución de datos en el panel superior izquierdo. En cada paso del algoritmo los puntos se mueven hacia el centro de su respectivo cúmulo. En la iteración final vemos como todos los puntos caen en el centro de su respectivo cúmulo. Todos los puntos que hayan convergido al mismo centro se clasificarán en el mismo cúmulo. 75

3.15. Algoritmo *Mean Shift* aplicado a nuestros modelos de la región 2. Observamos que todos los modelos reproducen la observación con una diferencia de 0.5 dex en el espacio O. En el espacio P el *clustering* nos divide el espacio en dos cúmulos. A pesar de que en el panel inferior izquierdo se sobreponen los cúmulos recordemos que el espacio P es un espacio en tres dimensiones y una proyección en dos dimensiones hace que encimen los cúmulos identificados. Con esta división podemos aplicar el método genético en cada cúmulo de manera independiente. 76

3.16. Generación 1 y 2 de la segunda familia de modelos para la región 2. La generación 1 (puntos rojos) es el grid inicial generado de forma aleatoria. La segunda generación (puntos turquesa) son generados a partir de los modelos de la primera generación que mejor ajustan las observaciones. 78

3.17. Resultado final de la segunda familia con una tolerancia de 0.05 para la región 2. Observamos como en esta región obtenemos los dos conjuntos de soluciones esperadas. Las líneas son las mismas mostradas en la Fig. 3.5. . . 80

3.18. Resultado final de la segunda familia con una tolerancia de 0.05 para la región 25. A diferencia de la región 2, en esta región solo se obtiene un conjunto solución. Esto se puede deber a que aún perdemos una solución o estemos en el punto máximo de relación entre O/H vs [OIII] λ 5007. Todos los modelos de estos parámetros reproducen la observación con una discrepancia máxima del 0.05 dex. Las líneas son las mismas mostradas en la Fig. 3.5. . . 81

3.19. Resultado después de siete generaciones de la tercera familia con una tolerancia de 0.2 para la región 25. A diferencia de la familia anterior donde los parámetros se ajustaban en cada generación a ojo. En esta generación que se trato de automatizar el proceso el conjunto solución que se obtiene está alejado de los resultados de la literatura. Esto se debe posiblemente a la automatización de la búsqueda del parámetro *quantile* para el método *Mean Shift*. Las líneas son las mismas mostradas en la Fig. 3.5. 83

3.20. Posibles cúmulos obtenidos para diferentes valores del parámetro *quantile*.
 Dependiendo del valor que se utilice podemos obtener diferentes tipos de
 cúmulos. Para un valor igual a *quantile* = 1 es el caso a) tomará todos los
 elementos y solo identificará un cúmulo. Para valores muy cercanos a 0 para
 el parámetro *quantile* el método puede identificar cúmulos muy pequeños. El
 caso ideal para algún valor de *quantile* que queremos obtener es el del caso
 c) donde los dos cúmulos son de tamaño comparable. 84

3.21. Ejemplo de un análisis de la calidad de *clustering* usando el coeficiente de
 Silhouette. Tenemos una muestra de datos en los paneles derechos. Cada
 uno se le ha aplicado un *clustering* usando el método *Kmeans* para que
 identificara 2, 3 y 4 cúmulos. El coeficiente de Silhouette puede ser calcu-
 lado para cada cúmulo identificado, mostrado en los paneles izquierdos de
 diferente color y su respectiva etiqueta. El coeficiente promedio de Silhou-
 ette es mostrado como una línea discontinua roja, el cual es calculado a
 partir de todos los coeficientes de Silhouette. Según la muestra de datos y
 del número de cúmulos identificados el coeficiente tendrá un mayor o menor
 valor. Basado en un ejemplo de *Scikit-learn* (Pedregosa et al., 2011) 87

4.1. Esquema de una generación de nuestro método final. A partir de un grid
 inicial seleccionan los modelos que mejor reproducen las observaciones dentro
 de una tolerancia dada. A estos modelos se les aplica el algoritmo *mean shift*
 buscando el valor de *quantile* que identificará de dos a tres cúmulos. Se aplica
 el método genético seleccionando los mejores modelos para ser progenitores
 y generar *m* crías. Con estos modelos se corren modelos en Cloudy e inicia
 el proceso nuevamente. 91

4.2. Generación de puntos por ajuste lineal de las generaciones 14 y 15. La genera-
 ción 12 (puntos negros) son tomados como bases para hacer un ajuste lineal.
 Alrededor de esta línea se generan los puntos para la generación 14 y 15. Se
 generan puntos desde un valor de O/H igual a 7 hasta aproximadamente 9.
 Esto se debe al σ utilizado para generar estos puntos. Este proceso se realizó
 con el fin de poblar aún más las zonas donde no había una alta densidad de
 puntos. 94

4.3. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la última familia de modelos de la región 2. Observamos como nuestros modelos reproducen la observación con una diferencia máxima de 0.05 dex en el espacio O. En el espacio P se encontraron los dos conjuntos solución principales más un tercer conjunto muy pequeño el cual corresponde a lo explicado al final del capítulo 1 (ver Fig. 1.11) Las líneas mostradas son las mismas que las de la Fig. 3.5 95

4.4. Modelos finales que difieren de la observación un máximo de 0.05 de la última familia de modelos de la región 25. Las líneas mostradas son las mismas que las de la Fig. 3.5. Tenemos un comportamiento similar que en la región 2 en los cuatro primeros paneles. En la tercera fila se muestran el valor de $[OIII]4363/5007$ vs $[OIII]$ y $[NII]5755/6584$ vs $[OIII]$ en el cual vemos que de conocerse el valor de $[OIII]4363$ ó $[NII]5755/6584$ podríamos diferenciar qué conjunto solución es correcto. Esto también se observa en el panel que muestra $[ArIII]/[NeIII]$. Por último se muestra EWa vs $[OIII]$. EWa está relacionado con el cociente entre el número de fotones procesados por el gas y el número de estrellas ionizantes, pero no hay diferencia alguna entre los dos conjuntos de soluciones. 97

4.5. Valores promedios de modelos finales para las dos geometrías ($fr = 0.03$ y 3.0). Los puntos mostrados son el promedio de todos los modelos que reproducen las observaciones para cada región. Se muestran con diamantes los valores promedio de los modelos que son astrofísicamente posibles. Los valores representados con asteriscos son aquellos que se encuentran apartados de las zonas de los calibradores de la literatura. Estos modelos y sus respectivos valores promedios los consideramos astrofísicamente improbables. 98

4.6. Modelos con diferentes edades para la población estelar ionizante sólo realizados para la región 25. En los paneles superiores se muestran que todos nuestros modelos para todas las edades utilizadas reproducen las observaciones con una discrepancia máxima de 0.05 dex. En los paneles inferiores mostramos nuestro espacio P. Observamos que el parámetro de ionización U cambia respecto a la edad de la población estelar. Notemos cómo el valor de U da la vuelta alrededor de una edad de 2.5 Myr. Para los valores de N/O y O/H observamos que no hay una gran diferencia entre los parámetros calculados con distintas edades. 101

- 5.1. Conjuntos de soluciones de todas las regiones modeladas para $f_r = 0.03$. Se muestran todos los modelos que reproducen la observación en color gris claro. También se muestran los valores promedio por región de todos los modelos correspondientes. Cada color representa una región diferente. Los diamantes representan soluciones que tienen una gran probabilidad de representar regiones HII reales. Mientras que las estrellas representan modelos que son físicamente correctos pero hay una baja probabilidad de que sean objetos astrofísicos reales. El criterio para clasificarlos de esta forma es la cercanía que tienen con las líneas mostradas. Las líneas discontinuas es el ajuste de Morisset et al. (2016) llamado PaperI. La línea morada discontinua representa el calibrador de Pilyugin et al. (2012) y la línea verde punteada el calibrador de Dopita et al. (2013). Junto a cada punto de valor medio es mostrado un contorno. Este contorno representa dónde se concentran la mayor cantidad de modelos que reproducen la observación para la región correspondiente. Es decir, el área que delimita el contorno es un conjunto solución de la región. Toda esa área puede reproducir la observación correspondiente. Los contornos sólidos son aquellos que corresponden a una región HII real y los contornos discontinuos representarán los modelos que tienen una alta probabilidad de no representar un objeto astrofísico real. 107
- 5.2. Conjuntos de soluciones de todas las regiones modeladas para $f_r = 3.0$ con sus respectivos contornos. Está figura es análoga a la Fig. 5.1. 108
- 5.3. Abundancia de oxígeno vs el coeficiente O3N2. Cada color corresponde a la solución astrofísicamente posible de una región modelada con un $f_r = 3.0$. Se muestran los valores de la abundancia de O para cada región. Sólo se muestran los modelos que tienen una alta probabilidad de representar un objeto astrofísico real. La Línea negra representa el calibrador realizado por Marino et al. (2013). Las líneas definidas por nuestros modelos representan los conjuntos solución que son compatibles con el calibrador de Marino. . . 109
- 5.4. Abundancia de oxígeno vs el coeficiente O3N2. Cada color corresponde a la solución astrofísicamente posible de una región modelada con un $f_r = 0.03$. Esta imagen es análoga a la Fig. 5.3 110
- A.1. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 2 114

A.2. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 3	115
A.3. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 5	116
A.4. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 6	117
A.5. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 7	118
A.6. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 9	119
A.7. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 10	120
A.8. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 12	121
A.9. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 13	122
A.10. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 14	123
A.11. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 16	124
A.12. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 18	125
A.13. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 19	126
A.14. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 20	127
A.15. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 21	128
A.16. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 24	129
A.17. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 25	130
A.18. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 26	131

A.19. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 27	132
A.20. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 28	133
A.21. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 31	134
A.22. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 33	135
A.23. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 35	136
A.24. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 37	137
A.25. Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 38	138

Agradecimientos

Este trabajo no hubiera sido posible de completar sin la ayuda de mucha gente tanto del Instituto de Astronomía como de fuera de este.

A mi sínodo: Dr. Sebastián Sánchez, Dr. Manuel Peimbert, Dr. Fabían Rosales, Dra. Aida Wofford y Dra. Irene Cruz. Muchas gracias por tomarse el tiempo de revisar y darme los comentarios para que la tesis tuviera la formalidad y el nivel que requiere una tesis de maestría.

Agradezco a los proyectos PAEP, DGAPA/PAPIIT-107215 y CONACyT-CB2015-254132, por brindarme los apoyos necesarios para desarrollar y complementar mi aprendizaje a través de la asistencia a diferentes congresos nacionales e internacionales como a diferentes escuelas que me permitieron discutir sobre diversos temas así como presentar mi trabajo.

Gracias a el Instituto de Astronomía por proporcionarme el espacio y las herramientas necesarias para concluir mis estudios. A Bertha y Heike por estar siempre para ayudarnos en todos los trámites necesarios.

Gracias a los miembros de mi comité tutor, Dra. Leticia Carigi y Dr. Sebastián Sánchez por darle seguimiento a mi trabajo, por sus valiosas sugerencias y consejos tanto para complementar mi trabajo como para mi desarrollo durante la maestría.

Agradezco a la escuela Nebulatom, a la escuela Guillermo Haro y a los distintos congresos que tuve la suerte de asistir por darme la oportunidad de conocer a tantas personas que ayudaron a que comprendiera mejor lo que significa ser un investigador. Asimismo me permitieron adquirir nuevos conocimientos que ayudaron a desarrollar mi trabajo.

A mis compaeros en la maestría con cuales he compartido el mayor tiempo durante este proceso. Gracias por todas las experiencias, discusiones y comidas compartidas.

Al Dr. Christophe Morisset por ser mi guía y tutor durante esta etapa. Por darme tanto de su tiempo y su paciencia. Por sus valiosas enseñanzas de astrofísica y de programación que son la base de este trabajo. Y por siempre procurar que tuviera las mejores condiciones de trabajo.

*Para todas aquellas personas que he conocido
e influyeron directa o indirectamente
para que llegara a este punto.
Gracias a todos ustedes*

Abstract

Photoionization models to fit some of the giant HII regions from the CALIFA survey are presented. Each model adjusts three free parameters: O/H, N/O and U (ionization parameter) to reproduce the $[\text{NII}]\lambda 6584/\text{H}\alpha$, $[\text{OIII}]\lambda 5007/\text{H}\beta$ and $[\text{OII}]\lambda 3727/\text{H}\beta$ line ratios simultaneously.

Two algorithms were used together to find the parameters of the HII regions: i) genetic method algorithm and ii) Mean Shift algorithm from cluster analysis. Combining these methods we reproduce the observed lines with a maximum difference of 0.05 dex.

Since our models suppose no observational information. We have a bi-valued relationship between $[\text{OIII}]\lambda 5007/\text{H}\beta$ and O/H. Therefore we find two solution sets that generate models that reproduce the observed values. Based on the results of previous work, we identify the solution sets that have a high probability of representing real HII regions. If we did not have this information, we would not be able to classify the parameters that correspond to real HII regions.

Finally, we explore the effect of the poor determination of the age of the stellar population on the determination of the parameters. We found that the age change of the ionizing stellar population does not significantly influence the derived parameters.

Resumen

Se realizaron modelos de fotoionización que ajustaran algunas de las regiones HII gigantes del *survey* CALIFA. Cada uno de los modelos ajusta tres parámetros libres: O/H, N/O y el parámetro de ionización (U) para reproducir de manera simultánea los cocientes de línea: $[\text{NII}]\lambda 6584/\text{H}\alpha$, $[\text{OIII}]\lambda 5007/\text{H}\beta$ y $[\text{OII}]\lambda 3727/\text{H}\beta$.

Se utilizaron una combinación de dos algoritmos para encontrar los parámetros de las regiones HII: i) algoritmo de método genético y ii) algoritmo *Mean Shift* de análisis de cúmulos. Con estos dos algoritmos se logró reproducir las líneas observadas con una diferencia máxima de 0.05 dex.

Dado que nuestros modelos suponen ninguna información observacional. Tenemos una relación bi-valuada entre $[\text{OIII}]\lambda 5007/\text{H}\beta$ y O/H. Por lo cual encontramos dos conjuntos solución que generan modelos que reproducen las líneas observadas. En base a los resultados de trabajos anteriores identificamos los conjuntos solución que tienen una alta probabilidad de representar regiones HII reales. Si no tuviéramos esta información no podríamos clasificar los parámetros que corresponden a regiones HII reales.

Por último, se exploró el efecto de la pobre determinación de la edad de la población estelar en la determinación de estos parámetros. Se encontró que el cambio de la edad de la población estelar ionizante no influye de manera significativa en los parámetros derivados.

Capítulo 1

El medio interestelar

§1.1. Introducción

El medio interestelar (MS por sus siglas) es toda la materia que se encuentra entre las estrellas en las galaxias, siendo aproximadamente el 75 % de masa bariónica total (Bisbas, 2016). Está conformado por nubes de gas interestelares que absorben y emiten radiación, no necesariamente en todas las longitudes de onda. Algunas presentan formas bien definidas como esferas, mientras que otras no presentan una estructura definida.

En el medio interestelar se presentan una gran variedad de condiciones físicas (temperatura y presión) y está iluminado por diferentes tipos de radiaciones. Debido a esto la materia tiene diferentes estados: atómico neutro, ionizado, molecular y en estado sólido, como el polvo (Kwok, 2007). Aunque, tanto los fotones, vientos estelares o una supernova pueden modificar sus propiedades físicas.

La mayor información sobre el MS que conocemos es por el estudio del espectro observable de las nubes interestelares. La forma del espectro está relacionado directamente con propiedades de las nubes, así como temperatura, densidad y composición química.

El MS de nuestra galaxia está compuesto principalmente de hidrógeno, $\sim 70\%$, $\sim 28\%$ de helio y $\sim 2\%$ de otros elementos más pesados. Comparando con la composición resultado de la nucleosíntesis ocurrida durante el Big Bang ($\sim 75\%$ H y $\sim 25\%$ He), el medio interestelar es más *metálico*. La metalicidad es la proporción de masa de los metales (elementos pesados) en comparación con la masa de todos los elementos.

Las estrellas son formadas por la materia del medio interestelar. En el interior de estos objetos se llevan a cabo procesos nucleares que dan como resultado helio y otros elementos

pesados. Por diversos procesos la estrella devuelve material al MS, como vientos solares, explosiones de supernova, etc. De esta manera, a lo largo del tiempo, la abundancia de elementos pesados en el medio interestelar va aumentando poco a poco. Por lo que el gas del medio es una mezcla de elementos pesados y gas prístino.

Por otra parte, hay un intercambio de energía entre la radiación estelar, el gas interestelar y el polvo, lo cual causa una excitación de estos últimos; también pueden ser afectados por rayos cósmicos, rayos X, entre otros.

En el medio interestelar encontramos diversos objetos que se clasifican por sus propiedades. Casi todos estos objetos están asociados con estrellas brillantes, en particular con estrellas tipo OB. Los fotones emitidos por estas estrellas son lo suficientemente energéticos, por lo que pueden interactuar con el gas interestelar y las partículas de polvo.

Si las nubes de gas y polvo no están asociadas con estrellas brillantes, el polvo puede absorber la radiación interestelar de campo, a estos objetos se les llama nebulosas oscuras. Si la nube está asociada con una estrella con alrededor de $T_{\text{eff}} \simeq 25000$ K, las partículas de polvo pueden dispersar la radiación estelar, a este tipo de objetos se les llama nebulosas de reflexión.

Si la estrella tiene $T_{\text{eff}} \gtrsim 25000$ K, el gas de los alrededores es fotoionizado, dando lugar a una nebulosa difusa o una región HII. En las regiones HII también puede haber moléculas, como CO, OH y NH₃. Estas nebulosas se les llama nubes moleculares y, por ejemplo, sólo pueden ser detectadas con técnicas de radio-astronomía (Maciel, 2013).

Las nebulosas planetarias, otro objeto del MS, se forman a partir del material expelido por estrellas entre $1 M_{\odot}$ y $8 M_{\odot}$ en las etapas finales de su vida. La estrella se transforma en una enana blanca y el gas que la rodea es el gas que ha sido expulsado por la propia estrella (Nebulosa planetaria del ojo de gato fig. 1.2).

§1.2. Regiones HII

Una región HII típica es el gas ionizado que está alrededor de una o más estrellas masivas y calientes recién formadas. Por esto, las regiones HII son un indicador de zonas de formación estelar. El tipo de estrellas asociadas a estas nebulosas son del tipo O y B. Los fotones que emiten este tipo de estrellas tienen la suficiente energía ($h\nu > 13.6$ eV) para fotoionizar al hidrógeno neutro. La nebulosa de Orión es un ejemplo de región HII típica (fig. 1.1).

La densidad de las regiones HII en general puede tener un valor de entre 1 a 10^4 cm⁻³.



Figura 1.1: Nebulosa de Orión. Imagen:NASA/JPL-Caltech/STScI



Figura 1.2: Nebulosa planetaria Ojo de gato. Imagen: NASA, ESA, HEIC, and The Hubble Heritage Team (STScI/AURA)

La temperatura de una región HII va de 5000 a 15000 K. La estrella central, o el conjunto de estrellas, puede tener una temperatura efectiva de entre 35000 y 50000 K, (Kwok, 2007). La edad, tanto de la estrella como de la región HII, es del orden de algunos millones de años (la región HII se produce al mismo tiempo que la estrella se forma).

Las masas observadas de estas regiones son del orden de 10^2 a $10^4 M_{\odot}$ (Osterbrock & Ferland, 2006). Algunas regiones HII supergigantes, como NGC 5471, pueden llegar a tener una masa de gas ionizado del orden de $10^6 M_{\odot}$ (García-Benito et al., 2011). Estas regiones pueden abarcar desde 1 pc hasta cientos de parsecs, por ejemplo la nebulosa de Orión que tiene un diámetro de ≈ 8 pc. Para regiones HII gigantes se han reportado que abarcan una distancia de casi 1 kpc, como lo es NGC 5471 (García-Benito et al., 2011).

Los espectros de las regiones HII presentan fuertes líneas de recombinación de hidrógeno y helio, además se producen líneas de excitación colisional, por ejemplo de iones de N y O. También se produce un continuo generado por radiación ‘free-free’ o *bremsstrahlung* (Dyson & Williams, 1997). La radiación free-free es producida cuando electrones libres se mueven cerca de un ion, lo cual produce que el electrón se acelere, sin ser capturado, emitiendo un fotón (Kwok, 2007).

Para el estudio de la física de las regiones HII se suponen dos equilibrios: el equilibrio de ionización y el equilibrio térmico.

§1.2.1. Equilibrio de fotoionización de una nebulosa de hidrógeno

La fotoionización es el proceso de interacción entre un fotón y un átomo, ion o una molécula dando como resultado un ion o la disociación de la molécula. Dado que el elemento mas abundante es hidrógeno, como primera aproximación podemos suponer que el gas que rodea a la estrella joven es puramente de H.

Supongamos que tenemos una densidad de hidrógeno neutro (n_{H^0}) y una intensidad de radiación ionizante promedio J_{ν} en cada punto de la nebulosa. Entonces el número de fotones incidentes por unidad de área, tiempo e intervalos de frecuencias es $4\pi J_{\nu}/h\nu$. La tasa de fotoionización volumétrica es:

$$n_{H^0} \int_{\nu_0}^{\infty} \frac{4\pi J_{\nu}}{h\nu} a_{\nu} d\nu = \left(\# \text{Átomos de HI/volumen} \right) \times \left(\text{Flujo ionizante de fotones} \right) \times \left(\text{sección eficaz de fotoionización} \right). \quad (1.1)$$

En la ec. (1.1) aparece a_{ν} , que es la sección eficaz de ionización la cual depende de la

frecuencia ν . Para un átomo de hidrógeno en el estado n (donde $n = 1$ es el estado base) y se cumpla que $\nu > \nu_c$, donde ν_c es el límite de ionización (el límite de Lyman: $\lambda_c = 912\text{\AA}$, $\nu_c = 3.3 \times 10^{15}\text{Hz}$) se puede expresar con una buena aproximación por la fórmula de Kramers (Sobelman, 1972):

$$a_\nu = \frac{64}{3\sqrt{3}}\alpha \left(\frac{\nu_c}{\nu}\right)^3 \frac{\pi a_0^2}{n^5},$$

donde $\alpha = e^2/\hbar c$ es la constante de estructura fina, πa_0^2 es la sección eficaz geométrica de un átomo con a_0 el radio de Bohr (Letokhov & Johansson, 2008). Reescribiendo lo anterior tenemos que:

$$a_\nu \approx a_{\nu_0} \left(\frac{\nu}{\nu_0}\right)^{-3} \quad \text{para } \nu \geq \nu_0,$$

donde $a_{\nu_0} = 6 \times 10^{-18} \text{ cm}^2$ (Hummer & Seaton, 1963).

En las nebulosas, se emiten líneas de recombinación que es el proceso inverso a la ionización, es decir, un electrón libre que cae a un átomo ionizado liberando energía al quedar ligado al átomo ionizado.

En el caso de puro H, queda ligado a un protón: $p + e^- \rightarrow \text{H} + \gamma$, por lo tanto, la tasa de radiación por recombinación por unidad de volumen es:

$$n_e n_p \alpha(T). \tag{1.2}$$

Donde n_e y n_p es la densidad de electrones y protones, respectivamente y $\alpha(T)$ es el coeficiente de recombinación. La recombinación es el decaimiento radiativo de un electrón de un estado no ligado a un estado ligado del hidrógeno. Por lo que la tasa de esta reacción es la probabilidad de que un protón y un electrón cercano (a una distancia de algunos radios de Bohr) se encuentren por la tasa de decaimiento atómico (en unidades de s^{-1}) es el coeficiente de recombinación en unidades de recombinaciones por $\text{cm}^{-3}\text{s}^{-1}$

Si suponemos que tenemos una distribución termalizada de las velocidades de los electrones, el coeficiente de recombinación para una recombinación de un estado atómico con un número principal cuántico n y un momento angular orbital L es:

$$\alpha_{nL} = \int_0^\infty v \sigma_{nL}(v) f(v) dv, \tag{1.3}$$

con $f(v)$ la distribución de velocidad Maxwell-Boltzmann y σ_{nL} la sección eficaz de recombinación la cual es el área donde un electrón puede ser capturado por un protón. Dado que

la recombinación es el proceso inverso a la fotoionización la sección eficaz de la recombinación puede ser obtenida a partir de la sección eficaz de la fotoionización.

Cuando un electrón se recombina con un protón, existe la probabilidad que éste caiga en el estado base (1^2S). Aunque el electrón puede caer en cualquier nivel y seguir una rápida cascada hasta el estado base. El coeficiente de recombinación total es la suma sobre los números cuánticos principales n y los momentos angulares L de todos los coeficientes de recombinación:

$$\begin{aligned}\alpha_A &= \sum_{n,L} \alpha_{n,L}(T), \\ &= \sum_n \sum_{L=0}^{n-1} \alpha_{nL}(T), \\ &= \sum_{n=1} \alpha_n(T).\end{aligned}\tag{1.4}$$

Pero, los electrones que se recombinan directamente al estado base emiten fotones que son ionizantes, y entonces son absorbidos en otros lugares de la nebulosa. La energía de estos fotones es mayor a 13.6 eV, esto causa que sea absorbido rápidamente por el gas. Estos fotones no afectan el balance general de ionización, una aproximación mejor es no considerar estas recombinaciones con $n = 1$ (Osterbrock & Ferland, 2006). Esto define un caso B del coeficiente total de recombinación:

$$\alpha_B = \sum_{n=2} \alpha_n(T).\tag{1.5}$$

Si consideramos el equilibrio, la tasa de fotoionización volumétrica es igual a la tasa de radiación por recombinación:

$$n_{H^0} \int_{\nu_0}^{\infty} \frac{4\pi J_\nu}{h\nu} a_\nu d\nu = n_e n_p \alpha_B(T).\tag{1.6}$$

Radio de Strömgren

La condición de equilibrio, ec. (1.6), es un balance entre la fotoionización y recombinaciones de electrones e iones. Existe un punto de transición donde la nebulosa es casi neutra y la radiación ionizante ya está totalmente absorbida, a este punto se define como el radio de Strömgren.

Supongamos una nebulosa esférica que contiene solamente H. Sea el radio de Strömgen denotado por R_s . Si estamos fuera de la nebulosa, $r > R_s$, la nebulosa es neutra, si estamos dentro de la nebulosa, es decir, $r \leq R_s$, todos los fotones ionizantes son absorbidos (si estamos en $r = R_s$, el hidrógeno está casi completamente ionizado). Podemos decir que la profundidad óptica es $\tau_{\nu_0} \gg 1$.

La profundidad óptica esta definida como (Rybicki & Lightman, 2008):

$$\tau_{\nu} = \int_0^r n_{HI} a_{\nu} ds.$$

Por otra parte, en una primera aproximación la intensidad promedio J_{ν} es simplemente la radiación emitida por la estrella entre la distancia al cuadrado debido al efecto de dilución geométrica (Osterbrock & Ferland, 2006). Si estamos a cierta distancia r en la nebulosa tenemos que:

$$4\pi J_{\nu} = \pi F_{\nu}(r).$$

$$4\pi J_{\nu} = \pi F_{\nu}(R) \left(\frac{R}{r}\right)^2 e^{-\tau_{\nu}(r)}.$$

De la condición de equilibrio, ec. 1.6, imponiendo que conforme nos acercamos a un radio R_s la profundidad óptica se vuelve infinita ($\tau_{\nu} \rightarrow \infty$ como $r \rightarrow R_s$)

$$4\pi R^2 \int_{\nu}^{\infty} \frac{\pi F_{\nu}(R)}{h\nu} d\nu \int_0^{\infty} d(-e^{-\tau_{\nu}}) = \int_0^{R_s} n_e n_p \alpha(T) 4\pi r^2 dr.$$

Pero $4\pi J_{\nu} = \frac{L_{\nu}}{4\pi r^2}$

$$\int_{\nu_0}^{\infty} \frac{L_{\nu}}{h\nu} d\nu = 4\pi R^2 \int_{\nu}^{\infty} \frac{\pi F_{\nu}(R)}{h\nu} d\nu. \quad (1.7)$$

Pero esto es la cantidad de fotones emitidos por la estrella que ionizan el hidrógeno por segundo

$$Q_0 \equiv \int_{\nu_0}^{\infty} \frac{L_{\nu}}{h\nu} d\nu. \quad (1.8)$$

Por lo que, integrando la parte derecha de la ec. (1.7), suponiendo que $n_e = n_p = n_H$:

$$Q_0 = \frac{4\pi}{3} R_s^3 n_H^2 \alpha_B(T). \quad (1.9)$$

Despejando R_s

$$R_s = \left(\frac{3Q_0}{4\pi n_H^2 \alpha_B(T)} \right)^{1/3}. \quad (1.10)$$

Este es el radio de Strömgen de una esfera llena, la cual es la distancia máxima donde los fotones provenientes de la estrella pueden ionizar al HI.

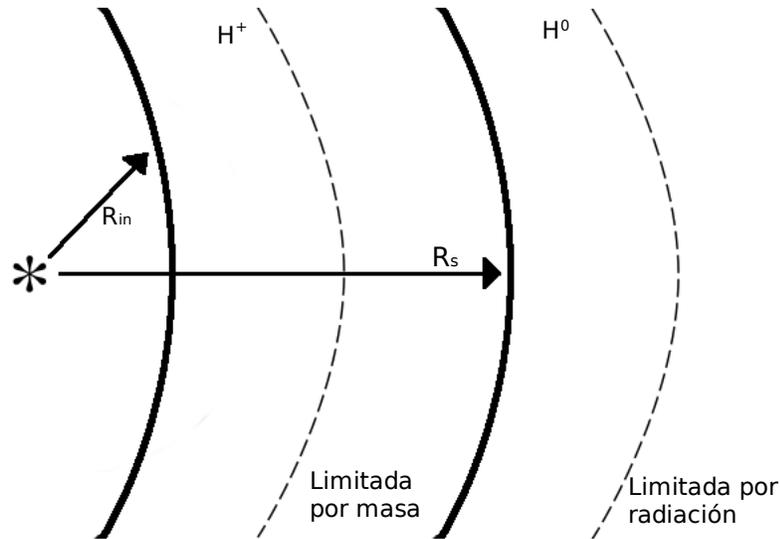


Figura 1.3: Diagrama de estructura de una nebulosa. Se representa la estrella con un asterisco y el inicio del gas como la primera línea gruesa cerca de la estrella. La distancia al gas es R_{in} y el radio de Strömgen como R_s . La segunda línea negra representa el final del gas ionizado. Si la nebulosa tiene un radio delimitante menor al R_s se dice que la nebulosa está limitada por masa. Por el contrario si este radio es mayor que R_s entonces se dice que la nebulosa está limitada por radiación.

En la fig. 1.3 se muestra un diagrama de la estructura de una nebulosa. Una nebulosa de emisión se dice que es limitada por masa (*matter bounded*) si la estrella asociada (o estrellas) ioniza todo el gas circundante, es decir, el radio exterior es menor a R_s . Por otro lado, si los fotones ionizantes no son suficientes para ionizar el gas de los alrededores, se dice que es una nebulosa limitada por radiación (*radiation bounded*), o sea, que el radio exterior es mayor a R_s (Basu et al., 2010).

Fotoionización con más elementos

En la realidad, aunque el hidrógeno sea el elemento más abundante en una nebulosa, puede contener muchos más elementos. El segundo más abundante es el helio y en mucha menor proporción puede haber O, N, S, etc. con sus respectivos iones.

Al agregar He a nuestra descripción de la nebulosa, debemos tomar en cuenta que la física atómica de este átomo es más compleja. Al tener dos electrones tiene dos posibles formas iónicas: He^+ y He^{++} . La primera tiene un potencial de ionización de 24.6 eV y la segunda 54.4 eV.

Las estrellas masivas de tipo O pueden emitir un número de fotones significativo con energías mayores a 24.6 eV, pero emiten muy pocos fotones con energías mayores a 54.4 eV. Por lo que podemos omitir He^{++} en regiones HII y solo tomar en cuenta el balance entre He y He^+ (Osterbrock & Ferland, 2006).

La sección eficaz de He también es diferente a la sección eficaz de H. En la fig. 1.4 se muestra la sección eficaz de H, He y otros elementos (Verner et al., 1996). La sección eficaz para la ionización de He es mayor que para el H para 24.6 eV. Eso nos dice que los fotones ionizantes que tengan una energía mayor a 24.6 eV ionizarán en primera instancia al He en lugar del H. Esto tendrá un impacto sobre el balance de equilibrio de ionización y sobre la estructura nebulosa que también dependerá del espectro estelar ionizante.

Si tenemos una estrella con una $T < 40000\text{K}$, esta tendrá fotones entre 13.6 eV y 24.6 eV capaces de ionizar H, pero también tiene algunos con energías mayores a 24.6 eV los cuales ionizan al He. Además, la recombinación de He puede emitir fotones con energía mayor a 13.6 eV que pueden ionizar H, lo que forma una zona central de He^+ y H^+ que rodea a la estrella. Alrededor de esta zona, ya no hay fotones que puedan ionizar He, lo que da origen a una zona de H^+ y He. La fracción de fotones con energía mayor a 24.6 eV absorbidos por H es (Osterbrock & Ferland, 2006):

$$y = \frac{n(\text{H})a_{\nu_2}(\text{H})}{n(\text{H})a_{\nu_2}(\text{H}) + n(\text{He})a_{\nu_2}(\text{He})}, \quad (1.11)$$

donde $1 - y$ es la fracción absorbida por He.

Una estrella con una temperatura entre 40000K y 100000K, la cual tenga una fuerte absorción en su atmósfera mayor de 54.5 eV, provocará una ionización de H y He en toda la nebulosa. Para estrellas calientes con $T > 10^5\text{K}$ están las estrellas centrales de nebulosas planetarias y estrellas Wolf-Rayet. Estas estrellas emiten fotones suficientemente

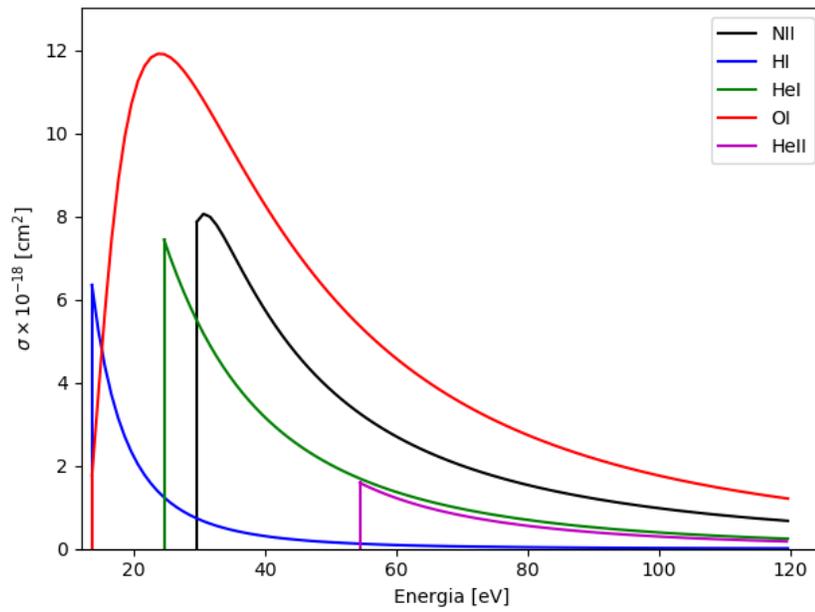


Figura 1.4: Secciones eficaces para HI, HeI, HeII, OI y NII, basado en Verner et al. (1996). Al incluir más elementos en las ecuación del equilibrio de fotoionización las secciones eficaces influirán dependiendo de que elementos se tomen en cuenta.

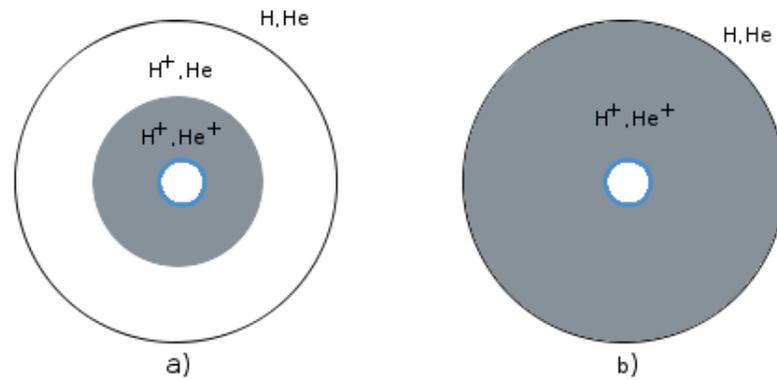


Figura 1.5: Diagrama para una nebulosa. a) Para una estrella con $T < 40000K$. Observamos que debido a la energía radiada por la estrella solo en sus alrededores puede ionizar una vez al hidrógeno y al helio. El gas más alejado se conforma de hidrógeno una vez ionizado y helio neutro. Esto se debe a que los iones que pueden ionizar al helio ya han sido absorbidos en las cercanías de la estrella. Esto crea dos zonas en el gas: la primera que se encuentra cerca del gas que se conforma principalmente de HII y $HeII$ y otra más lejana de la estrella que está compuesta en su mayoría por HII y HeI . b) Para una estrella con $40000K < T < 10^5K$. A diferencia del caso a) este tipo de estrellas son capaces de ionizar todo el hidrógeno presente en el gas de la nebulosa sin crear las dos zonas.

energéticos para formar el ion He^{++} . Alrededor de la estrella se forma una zona de He^{++} . La recombinación de He^{++} pueden producir fotones capaces de ionizar H, por lo que la zona está conformada por He^{++} y H^+ .

Al agregar helio a una nebulosa puramente de hidrógeno provoca un acoplamiento entre la ionización y la recombinación de helio e hidrógeno. Para calcular completamente la estructura de la nebulosa, se deben de resolver las ecuaciones de equilibrio de ionización acopladas. Este proceso puede ser hecho de manera numérica con códigos de equilibrio de fotoionización, como por ejemplo Cloudy (Ferland et al., 2013).

Otros elementos también pueden estar presentes en las nebulosa, por ejemplo C, N, O, S, Ne, Fe y Si. Dado que estos elementos tienen varios iones, consideremos dos etapas sucesivas de ionización, de r a $r + 1$ de un elemento X, entonces la ecuación de equilibrio de ionización (1.6) se escribe (Osterbrock & Ferland, 2006):

$$n(X^r) \int_{\nu_r}^{\infty} \frac{4\pi J_\nu}{h\nu} a_\nu(X^r) d\nu = n(X^{r+1}) n_e \alpha_G(X^r, T), \quad (1.12)$$

donde α_G es el coeficiente de recombinación del nivel base de X^{r+1} a todos los niveles de X^r , $a_\nu(X^r)$ es la sección eficaz del nivel base de X^r y $n(X^r)$ y $n(X^{r+1})$ son las densidades numéricas de dos sucesivos estados de ionización. Si suponemos que la abundancia de un elemento X es la suma de las abundancias de cada ion de X, entonces se podría determinar el equilibrio de ionización en cada punto de la nebulosa.

La sección eficaz es diferente para cada elemento, fig. 1.4. Los metales, y sus iones, tiene una estructura electrónica bastante compleja. Según el número de electrones que son expulsados del átomo, el ion tendrá diferentes condiciones de fotoionización. (Dopita & Sutherland, 2003).

Otros de los procesos presentes cuando hay componentes iónicos son las reacciones de intercambio de cargas. Estas reacciones son importantes en la determinación del balance de ionización de O. Esto se debe que O y H pueden colisionar, donde se puede dar un intercambio de carga. El potencial de ionización de O es 13.62 eV y el del H es 13.59 eV por lo que la reacción es casi resonante, lo cual incrementa su eficiencia. Esto causa que las fronteras de ionización de H^+ y O^+ estén ligados (Dopita & Sutherland, 2003).

Otros intercambios de carga pueden ser posibles si los potenciales de ionización son muy cercanos para algunos niveles. Estos efectos deben de ser tomados en cuenta en las ecuaciones acopladas que describen el equilibrio de ionización.

§1.2.2. Balance térmico

En una nebulosa estática existe un equilibrio entre el calentamiento generado por la fotoionización y el enfriamiento dado por la recombinación y otras pérdidas radiativas. Los electrones libres son rápidamente termalizados a una distribución de velocidades Maxwelliana por las colisiones entre ellos (Ferland et al., 2016).

Calentamiento por fotoionización

La fuente principal de calentamiento en una región ionizada es la fotoionización. El electrón libre producido por la fotoionización tiene una energía cinética igual a la diferencia entre energía del fotón y el potencial de ionización:

$$\frac{1}{2}mu^2 = h(\nu - \nu_0). \quad (1.13)$$

Supongamos que tenemos una nebulosa puramente de H. Asumiendo que las ionizaciones secundarias y que todo exceso de energía es convertida en calor podemos escribir la tasa de calentamiento por fotoionización como (Osterbrock & Ferland, 2006):

$$G(\text{H}) = n(\text{H}) \int_{\nu_0}^{\infty} \frac{4\pi J_{\nu}}{h\nu} h(\nu - \nu_0) a_{\nu}(\text{H}) d\nu \quad [\text{erg cm}^{-3} \text{s}^{-1}]. \quad (1.14)$$

El primer término dentro de la integral es el número de fotones ionizantes provenientes de la estrella, el segundo la energía que es proporcionada por la fotoionización y el tercer término es la sección eficaz de la fotoionización. Si la nebulosa está en equilibrio de fotoionización, entonces se cumple la ec. 1.6, por lo tanto, podemos escribir $G(\text{H})$ como:

$$G(\text{H}) = n_e n_p \alpha_A(\text{H}, T) \frac{\int_{\nu_0}^{\infty} \frac{J_{\nu}}{h\nu} h(\nu - \nu_0) a_{\nu}(\text{H}) d\nu}{\int_{\nu_0}^{\infty} \frac{J_{\nu}}{h\nu} a_{\nu}(\text{H}) d\nu}. \quad (1.15)$$

Donde la división de integrales es la energía térmica promedio de los electrones, por lo que (Osterbrock & Ferland, 2006):

$$G(\text{H}) = \frac{3}{2} n_e n_p \alpha_A(\text{H}, T) kT_e. \quad (1.16)$$

Procesos de enfriamiento. Recombinación

Uno de los tres principales procesos de enfriamiento en una nebulosa es el enfriamiento por recombinación. Cuando los electrones libres se recombinan con protones para formar H y emiten fotones, el plasma de electrones pierde energía. La pérdida de energía del gas de electrones es (Osterbrock & Ferland, 2006):

$$L_R(\text{H}) = n_e n_p kT \beta_A(\text{H}, T) \quad [\text{erg cm}^{-3} \text{s}^{-1}]. \quad (1.17)$$

Donde $\beta_A(\text{H}, T)$ es el coeficiente de recombinación promediado por energía:

$$\beta_A(\text{H}, T) = \sum_{n=1}^{\infty} \beta_n(\text{H}, T) = \sum_{n=1}^{\infty} \sum_{L=0}^{n-1} \beta_{nL}(\text{H}, T) \quad [\text{cm}^3 \text{s}^{-1}] \quad (1.18)$$

con

$$\beta_{nL}(\text{H}, T) = \frac{1}{kT} \int_0^{\infty} u \sigma_{nL}(\text{H}, T) \frac{1}{2} m u^2 f(u) du, \quad (1.19)$$

donde $\sigma_{nL}(\text{H}, T)$ es la sección eficaz de recombinación y depende de las velocidades relativas de los electrones y protones, es decir, es proporcional a u^{-2} . Esto nos indica que los electrones que son lentos son mayormente recapturados que los electrones rápidos.

Si suponemos que la recombinación es el único mecanismo de enfriamiento, es decir:

$$G(\text{H}) = L_R(\text{H}), \quad (1.20)$$

esto causará que la temperatura electrónica, T_e , sea un poco mayor que la temperatura inicial, T_i (Osterbrock & Ferland, 2006). Esto se debe a que los electrones lentos se recombinan, haciendo que la distribución de velocidades de los electrones libres restantes se incline hacia energías más altas y por lo tanto temperaturas más altas.

El anterior análisis es sólo para el caso de una nebulosa de H, si agregamos otros elementos, como He, entonces tendríamos que la tasa de calentamiento por fotoionización, sólo del He, es:

$$G(\text{He}) = n_e n_{\text{He}^+} \alpha_A(\text{He}, T) \frac{\int_{\nu_2}^{\infty} \frac{J_{\nu}}{h\nu} h(\nu - \nu_2) a_{\nu}(\text{He}) d\nu}{\int_{\nu_2}^{\infty} \frac{J_{\nu}}{h\nu} a_{\nu}(\text{He}) d\nu}. \quad (1.21)$$

Y para el enfriamiento, se debe de sumar la contribución de ambos elementos, H y He, tal

que:

$$L_R = L_R(\text{H}) + L_R(\text{He}), \quad (1.22)$$

donde

$$L_R(\text{He}) = n_e n_{\text{He}^+} kT \beta_A(\text{He}, T). \quad (1.23)$$

De manera similar se agregan las contribuciones al enfriamiento de otros átomos o iones.

Procesos de enfriamiento. Radiación *Free-Free*

Cuando un electrón libre pasa cerca de un ion, éste se acelera y emite radiación. Los electrones tienen diferentes energías y pasan a diferentes distancias. Esto causa que la radiación resultante sea continua, la cual es llamada radiación *free-free*. Este nombre se debe a que el electrón no queda ligado al átomo o ión. (Kwok, 2007).

La tasa de enfriamiento por radiación *free-free* para un ion con carga Z es (Osterbrock & Ferland, 2006)

$$L_{ff}(Z) = 4\pi j_{ff} = \frac{32\pi e^6 Z^2}{3^{3/2} h m_e c^3} \left(\frac{2\pi kT}{m_e} \right)^{1/2} n_e n_+ g_{ff}, \quad (1.24)$$

donde g_{ff} es el factor promedio Gaunt para emisión *free-free*, el cual es función de la densidad y la temperatura. Y n_+ es la densidad numérica de los iones con una carga nuclear Z . Para una nebulosa con distintos iones, por ejemplo, H^+ y He^+ , n_+ es la suma de ambas densidades, n_P Y n_{He^+}

A pesar de que el enfriamiento por *free-free* es bastante ineficiente, contribuye lo suficiente para disminuir un poco la temperatura. Si suponemos que los únicas fuentes de enfriamiento son recombinación y emisión *free-free*, la temperatura resultando será un poco más pequeña que la T_i .

Procesos de enfriamiento. Emisión de líneas por excitación colisional.

La excitación colisional es el proceso de enfriamiento más eficiente. Esto se debe a que algunos iones, como O^+ , O^{++} y N^+ , poseen potenciales de excitación colisional de algunos eV. Si consideramos que una temperatura típica de una nebulosa es de 10^4K , la energía térmica de los electrones será igualmente de algunos eV. Esto hace que la excitación por colisiones de los iones metálicos sea muy eficiente a comparación del H y He. A pesar de que H y He son los elementos más abundantes, la energía de excitación por encima del

estado base es de $\sim 10\text{eV}$ haciendo que su excitación colisional sea ineficiente.

El enfriamiento por excitación colisional es el mecanismo de enfriamiento dominante para nebulosas. Una nebulosa con una baja metalicidad tendrá una temperatura mayor comparada a una nebulosa con la misma estrella central, pero con una metalicidad más alta.

Cuando un átomo o ion es excitado colisionalmente, éste puede desexcitarse de dos formas diferentes. El átomo o ion puede desexcitarse espontáneamente emitiendo un fotón que puede escapar de la nebulosa. Este fotón se lleva parte de la energía del plasma. Por otra parte, el átomo o ion puede ser desexcitado por otro choque con otro electrón antes de emitir un fotón. Este último proceso no generará un cambio neto de la energía del plasma.

La mayoría de las líneas colisionales corresponden a líneas prohibidas. Estas líneas reciben este nombre por que las transiciones que las generan violan las reglas de selección de paridad. Pero pueden violar estas reglas por transiciones de cuadru-polo eléctrico o transiciones de dipolo magnético pero su coeficiente de emisión espontanea son muy pequeños. En las condiciones interestelares, el tiempo entre colisiones es lo suficientemente alto para que el nivel pueda decaer emitiendo un fotón de una línea prohibida. En la tierra, las colisiones son mucho más frecuentes por la alta densidad, por lo que el átomo o ion es desexcitado colisionalmente antes de que se desexcite espontáneamente.

Para calcular el enfriamiento provocado por la excitación colisional se necesita calcular la población del nivel superior para todos los niveles más bajos para cada ión. El equilibrio estadístico dentro del átomo es el balance entre la tasa de excitación colisional fuera del estado base y la tasa de desexcitaciones colisionales y radiativas del estado excitado. En general para un nivel j tenemos la condición de equilibrio estadístico es:

$$\sum_{k \neq j} n_k n_e q_{kj} + \sum_{k > j} n_k A_{kj} = \sum_{k \neq j} n_j n_e q_{jk} + \sum_{k < j} n_j A_{jk}. \quad (1.25)$$

Donde A_{kj} es el coeficiente de Einstein para la transición espontánea $k \rightarrow j$ y es la probabilidad de la transición por unidad de tiempo (Rybicki & Lightman, 2008). q_{kj} son los coeficientes de excitación ($k < j$) y de desexcitación ($k > j$). El coeficiente q_{kj} para $k > j$ es de la forma (Kwok, 2007):

$$q_{kj} = \frac{8.629 \times 10^{-6} \gamma_{kj}}{T_e^{1/2} g_k} \quad [\text{cm}^3 \text{s}^{-1}]. \quad (1.26)$$

Mientras que para $j > k$ es

$$q_{jk} = \frac{g_k}{g_j} e^{-E_{kj}/kT_e} q_{kj}. \quad (1.27)$$

Con γ_{kj} las fuerzas colisionales, esta función tiene una lenta dependencia con respecto a la temperatura. Y g_k es el peso estadístico del nivel en el cual la transición inicia.

En la ec. 1.25 el lado izquierdo son todas las excitaciones colisionales y desexcitaciones colisionales y radiativas que pueblan el estado j . El lado derecho son las excitaciones colisionales y desexcitaciones colisionales y radiativas que despueblan el estado j .

Para estimar el enfriamiento por líneas colisionales de un nivel N de un átomo o ión, se suman todas las emisiones de líneas originadas por el átomo tal que:

$$L_C = \sum_{j=1}^N n_j \sum_{k<j} A_{jk} h\nu_{jk}. \quad (1.28)$$

Numéricamente primero se resuelve la ec. 1.25 para las poblaciones del nivel. Estos valores son usados en la ec. 1.28 con sus respectivas probabilidades de transición y energías para derivar el enfriamiento total. Se puede definir la densidad crítica para un estado excitado como:

$$n_{crit}(j) = \frac{\sum_{k<j} A_{jk}}{\sum_{k \neq j} q_{jk}}. \quad (1.29)$$

Si la densidad electrónica es menor que la densidad crítica la desexcitación colisional es despreciable. En el caso contrario, la desexcitación colisional se torna importante haciendo que el enfriamiento sea eficiente. Esto se debe a que las colisiones dominan los procesos de desexcitación. Si la densidad es igual a la densidad crítica, entonces la desexcitación colisional es igual a la desexcitación radiativa. (Draine, 2010).

Equilibrio térmico

Tomando en cuenta todos los procesos, de calentamiento y enfriamiento, tenemos que el equilibrio térmico en la nebulosa está dado por:

$$G = L_R + L_{ff} + L_C. \quad (1.30)$$

La evaluación de la expresión anterior implica resolver las ecuaciones de ionización en cada punto de la nebulosa para encontrar qué iones están presentes. Una vez teniendo estos datos y combinándolos con cálculos de modelos atómicos se evalúa el balance térmico de

ese punto. Todo este proceso puede ser calculado numéricamente por códigos como Cloudy.

Además de los procesos antes mencionados para el enfriamiento en nebulosas el polvo puede provocar otros efectos. Uno de ellos es la sección eficaz de absorción del polvo la cual aumenta con altas energías provocando que en promedio la energía de los fotoelectrones disminuya. Por otro lado, la depleción de las abundancias de algunos iones que son importantes enfriadores causa que la eficiencia del enfriamiento se reduzca.

§1.3. Cálculo de propiedades físicas en nebulosas

Las regiones HII tiene un espectro que está caracterizado por tener fuertes líneas de emisión resultado de: i) cascadas de recombinación del hidrógeno, helio y en menor grado de otros iones. ii) Excitaciones colisionales resultado del impacto entre electrones libres e iones.

Las intensidades de las líneas de recombinación del H y He depende tanto de la luminosidad total y el espectro del campo de radiación ionizante. Por lo tanto, pueden ser usadas para derivar a partir de ellas propiedades globales de las nebulosas.

La cantidad de un ion para cierto átomo (por ejemplo O^{++} o S^+) depende de las propiedades térmicas locales del gas ionizado, principalmente de la densidad y la temperatura, y del campo de radiación. La cantidad de cierto ion se verá reflejada en la intensidad de las líneas que sean emitidas por éste. Podemos usar estas líneas emitidas para poder derivar la densidad de la nebulosa y su temperatura donde la línea es producida.

Además las intensidades de las líneas de los otros elementos diferentes al H y He (llamados *metales*) respecto al hidrógeno son sensibles a la abundancia relativa al ion y a la abundancia total del elemento en cuestión así como a la forma del espectro ionizante. Esto nos permite una manera de medir las abundancias de los elementos dentro de la nebulosa.

§1.3.1. Temperaturas y densidades nebulares

Podemos usar algunas líneas de emisión para calcular temperaturas y densidades de la nebulosa. Dependiendo de la transición de los electrones entre los distintos niveles de los átomos la línea puede ser sensible a la temperatura o a la densidad. Usando esta dependencia podemos estimar estas propiedades físicas de la zona donde se emite la línea.

Para calcular la temperatura a partir de líneas de emisión debemos tomar aquellas que son sensibles a la temperatura. Por ejemplo, para el OIII (fig. 1.6) los estados 1S y

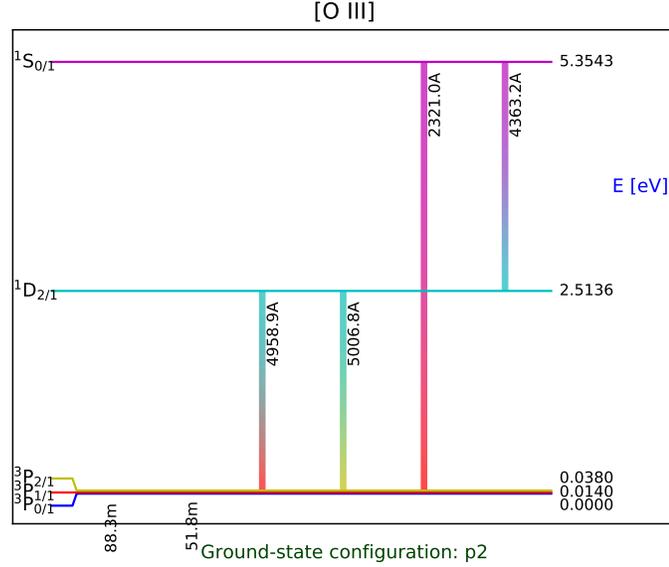


Figura 1.6: Diagrama de transiciones entre niveles de [OIII], todos desde la configuración $2p^3$. Las líneas de emisión que se encuentran en el óptico son $\lambda 4363$, $\lambda 5007$ y $\lambda 4949$. Las otras líneas mostradas se encuentran en el infrarrojo y en el ultravioleta.

$1D$ dependen de la temperatura ya que tienen diferentes energías de excitación. Como se mencionó anteriormente si la densidad del medio es baja, más baja que la densidad crítica, las desexcitaciones colisionales pueden despreciarse. Por lo que cada excitación del nivel $1D$ resulta en la emisión de un fotón con $\lambda 5007$ o $\lambda 4949$. Por otro lado cada excitación del nivel $1S$ será seguido de un fotón con $\lambda 4363$ o $\lambda 2321$. Cada emisión de un fotón con $\lambda 4363$ será seguido por la emisión de un fotón con $\lambda 5007$ o $\lambda 4949$, pero esta contribución es pequeña comparada con la excitación directa de $1D$.

De las ecuaciones de equilibrio estadístico para la población de niveles podemos derivar el cociente $(j_{\lambda 4949} + j_{\lambda 5007})/j_{\lambda 4363}$ (Osterbrock & Ferland, 2006):

$$\frac{j_{\lambda 4949} + j_{\lambda 5007}}{j_{\lambda 4363}} = \frac{7.9 \exp(3.29 \times 10^4/T)}{1 + 4.5 \times 10^{-4} n_e / T^{1/2}}. \quad (1.31)$$

Esta última expresión es una solución de primer orden y además tiene una débil dependencia sobre la densidad. Otro cociente comúnmente usado es el que involucra las líneas

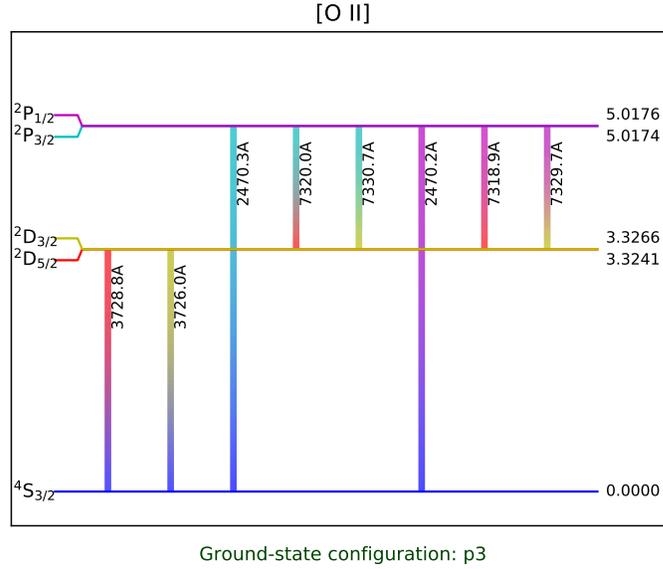


Figura 1.7: Diagrama de transiciones entre niveles de [OII]. Las líneas que se encuentran en el rango óptico son $\lambda 3726$ y $\lambda 3720$.

de nitrógeno:

$$\frac{j_{\lambda 6548} + j_{\lambda 6583}}{j_{\lambda 5755}} = \frac{6.91 \exp(2.5 \times 10^4/T)}{1 + 2.5 \times 10^{-3} n_e / T^{1/2}}. \quad (1.32)$$

Ambos cocientes de líneas están en el rango óptico. Estas soluciones no son exactas dado que hay efectos que se han despreciado. Por otra parte, las líneas [OIII] $\lambda 4363$ y [NII] $\lambda 5755$ son líneas de emisión muy débiles en una nebulosa típica. Por lo que los errores en la medición de estas líneas pueden provocar una temperatura errónea.

Para la determinación de la densidad se toman en cuenta aquellos cocientes de línea que son sensibles a la desexcitación colisional. Usualmente se comparan líneas que vienen de un nivel con un doblete muy cercano por ejemplo [OII] y su nivel $2D$ (fig. 1.7).

Estas líneas generadas por la transición de $2D_{5/2,3/2} \rightarrow 4S_{3/2}$ tienen una diferencia en energía muy pequeña en comparación a la temperatura. Por lo que si se hace el mismo tratamiento que para la temperatura se tendría que el factor exponencial que depende de la temperatura sea despreciable. El cociente de estas líneas dependerá únicamente de la densidad.

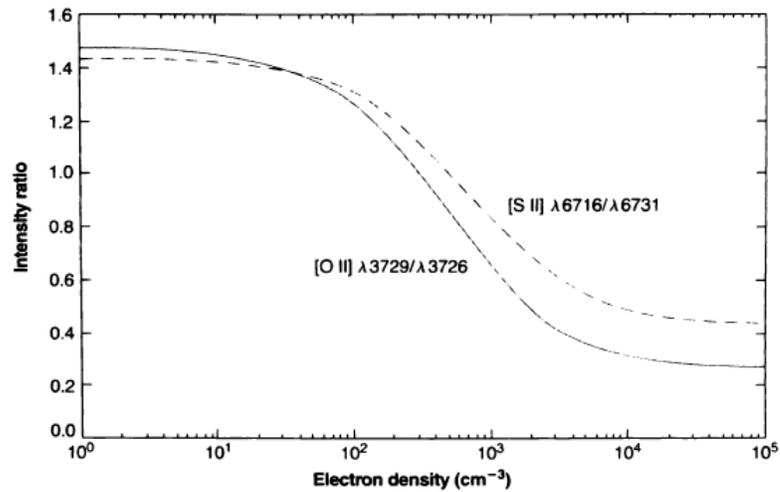


Figura 1.8: Variación de los cocientes de línea de [OII] y [SII] con respecto a la densidad. Aproximadamente entre una densidad electrónica de 10^2 y 10^4 tenemos la mayor variación de las intensidades de línea. En este rango es donde podemos usar estas líneas para calcular la densidad. Fuera de este rango la intensidad no varía con respecto a la densidad. (Osterbrock & Ferland, 2006)

En la fig.1.8 podemos ver los dos cocientes de líneas más usados para determinar la densidad. Las líneas de S proveen una buena estimación de densidad en el rango de $n_e \approx 10^2 - 10^4 \text{ cm}^{-3}$ en un rango de temperatura de $T_e = 5000 - 20000 \text{ K}$. Aunque la abundancia de S es pequeña, por lo que las líneas tienden a ser débiles. Por otra parte, S y sus especies iónicas primarias solamente se encuentran dentro de una pequeña región de la nebulosa.

El cociente de [OII] tiene ventaja sobre el cociente de [SII] ya que el ion O^+ es abundante, por lo que la línea generada es muy brillante. Pero O^+ es la especie menos abundante de las especies de O en una nebulosa por lo que no proveerá una buena estimación de la densidad.

§1.3.2. Cálculo de abundancias

El cálculo de abundancias elementales e iónicas puede calcularse de diferentes maneras dependiendo de las líneas que se observan. Para obtener las abundancias directamente a partir de las líneas emitidas por los iones es necesario conocer la temperatura del gas. Si las líneas que permiten calcular la temperatura no están disponibles entonces se deben de

usar métodos empíricos basados en calibraciones usando muestras observadas.

El método directo utiliza cocientes de líneas para derivar las abundancias iónicas. Si suponemos que la composición química es uniforme en la nebulosa tenemos

$$\frac{n(X_i)}{n(X_{i'})} = \frac{I_{ijl}/I_{i'j'l'}}{\epsilon_{ijl}(T_l, n_l)/\epsilon_{i'j'l'}(T'_l, n'_l)} icf. \quad (1.33)$$

Donde I son las intensidad de las líneas, icf es el factor de corrección de ionización dado por

$$icf = \frac{\int n(X_i^j)/n(X_i)n_e dV}{\int n(X_{i'}^{j'})/n(X_{i'})n_e dV}. \quad (1.34)$$

Para usar los métodos directos debemos de tener un cálculo muy bueno de la temperatura y densidad. Pero en algunas nebulosas las líneas que son indicadores de temperatura como [OIII]4363 o [NII]5755 no están disponibles ya sea por que la temperatura es muy baja o hay baja señal al ruido.

En estos casos podemos usar los métodos de líneas fuertes. Este tipo de métodos se basan en relaciones empíricas de objetos similares. Pero estos métodos necesitan ser calibrados ya sean con modelos de fotoionización o de métodos directos o en dado caso de objetos con la misma composición química.

Por ejemplo Pagel et al. (1979) usó $([OII]+[OIII])/H\beta$ como indicador de O/H calibrado con el grid de modelos de regiones HII de Stasińska (1978). Este método ha sido calibrado en diversas ocasiones. En años posteriores McGaugh (1991) refinó el método tomando en cuenta el parametro de ionización U y Pilyugin también lo ha refinado varias veces pero usando observaciones para calibrarlo, por ejemplo: Pilyugin (2000); Pilyugin et al. (2012).

Otro ejemplo de un método de líneas fuertes es el realizado por Marino et al. (2013). Este método fue realizado para las regiones HII de CALIFA. Usaron el cociente de dos líneas fuertes de emisión: [OIII]5007 y [NII]6584 tal que :

$$O3N2 = \log \left(\frac{[OIII]\lambda 5007}{H\beta} \times \frac{H\alpha}{[NII]\lambda 6583} \right). \quad (1.35)$$

Usando regiones HII donde se midieran las líneas [OIII]5007 y [NII]6584 además de al menos una línea auroral para determinar las temperaturas electrónicas obtuvieron la relación:

$$12 + \log(O/H) = 8.533 - 0.214 \times O3N2. \quad (1.36)$$

En la fig.1.9 podemos observar el ajuste de Marino et al. (2013) como una línea continua verde. Los diferentes símbolos representan la fuente de donde fueron obtenidas las muestras de regiones HII. Estas regiones fueron recopiladas de 17 trabajos diferentes. Los colores representan con qué línea se obtuvo la temperatura. Los símbolos azules son las regiones HII donde se calculó su temperatura con la línea [OIII] λ 4363. Los símbolos naranjas indican las regiones donde se usó la línea [NII] λ 5755. En las regiones que se representan con símbolos rojos se usó la línea [SIII] λ 6312 y en verde se representan las regiones donde se usaron las líneas [OIII] λ 4363 y [NII] λ 5755.

Se comparó esta nueva calibración con otras de otros trabajos de la literatura. La línea negra y su rango de validez con las líneas verticales grises punteadas es la calibración dada por Pettini & Pagel (2004). La calibración hecha por Pérez-Montero & Contini (2009) está señalada por la línea gris discontinua. Por último la línea gris discontinua y tres puntos representa el ajuste hecho por Nagao et al. (2006)

Como podemos observar la calibración O3N2 coincide bastante bien con las muestra de regiones HII. Esta relación será usada más adelante para comparar nuestros modelos con los modelos realizados en un anterior trabajo como se explicará al inicio del capítulo 3.

Un problema de los métodos de líneas fuertes es que si se toma una línea emitida por excitación colisional y una línea de recombinación se tendrá una relación bi-valorada. Por ejemplo, el cociente [OIII]5007/H β presenta esta degeneración con respecto a O/H como se muestra en la fig. 1.10. Esto se debe a que a bajas metalicidades estos cocientes incrementan conforme la metalicidad aumenta. Pero a altas metalicidades disminuyen debido a el incremento del enfriamiento debido a las líneas infrarojas. Esto causa una disminución en la temperatura por debajo del límite para una línea de excitación colisional.

En estos casos se debe de buscar un argumento para indicar si tenemos una abundancia alta o baja. Por ejemplo, la relación antes explicada (Marino et al., 2013) puede ser usada para calibrar el tipo de abundancia que se tiene.

El caso mostrado en la fig.1.10 se toma un valor fijo de N/O. Pero si se deja como parámetro libre la curva mostrada en la fig.1.10 puede variar de posición y de forma. En la fig.1.11 se muestra el efecto de dejar libre N/O y logU. El parámetro de ionización U se define como:

$$U = \frac{1}{4\pi r^2 c n_e} \int_{\nu_0}^{\infty} \frac{L_\nu}{h\nu} d\nu, \quad (1.37)$$

donde L_ν es la luminosidad de la fuente por unidad de frecuencia y r es la distancia a la fuente. Físicamente, U representa el cociente entre la densidad de fotones ionizantes y

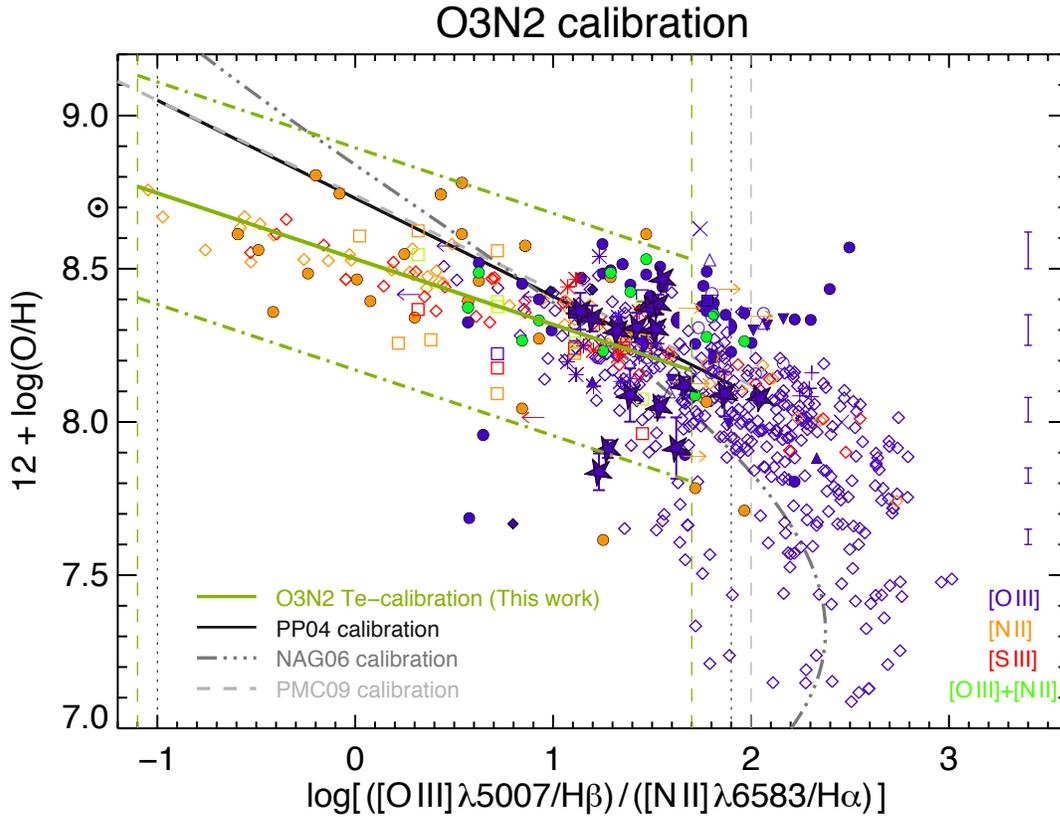


Figura 1.9: Calibración hecha por Marino et al. (2013) donde se muestra las distintas regiones HII usadas y otras calibraciones de otros trabajos. Los colores representan con qué línea se obtuvo la temperatura. Los símbolos azules son las regiones HII donde se calculó su temperatura con la línea [OIII] λ 4363. Los símbolos naranjas indican las regiones donde se usó la línea [NII] λ 5755. En las regiones que se representan con símbolos rojos se usó la línea [SIII] λ 6312 y en verde se representan las regiones donde se usaron las líneas [OIII] λ 4363 y [NII] λ 5755.

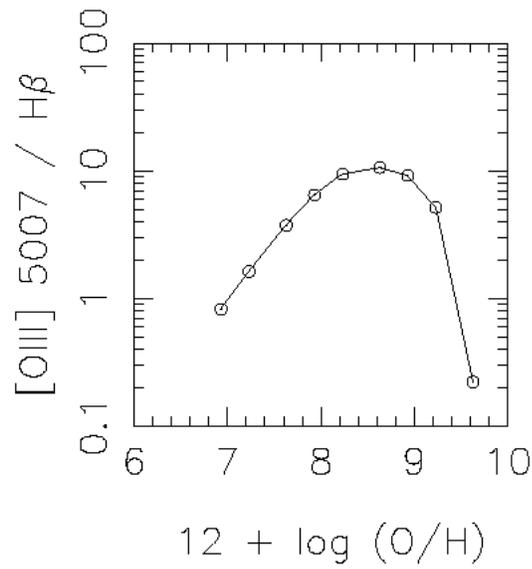


Figura 1.10: Degeneración del cociente $[OIII]5007/H\beta$ con respecto a O/H (Stasińska, 2002). Este comportamiento se debe a que a bajas metalicidades estos cocientes incrementan conforme la metalicidad aumenta. Pero a altas metalicidades disminuyen debido a el incremento del enfriamiento debido a las líneas infrarojas. Esto causa una disminución en la temperatura por debajo del límite para una línea de excitación colisional.

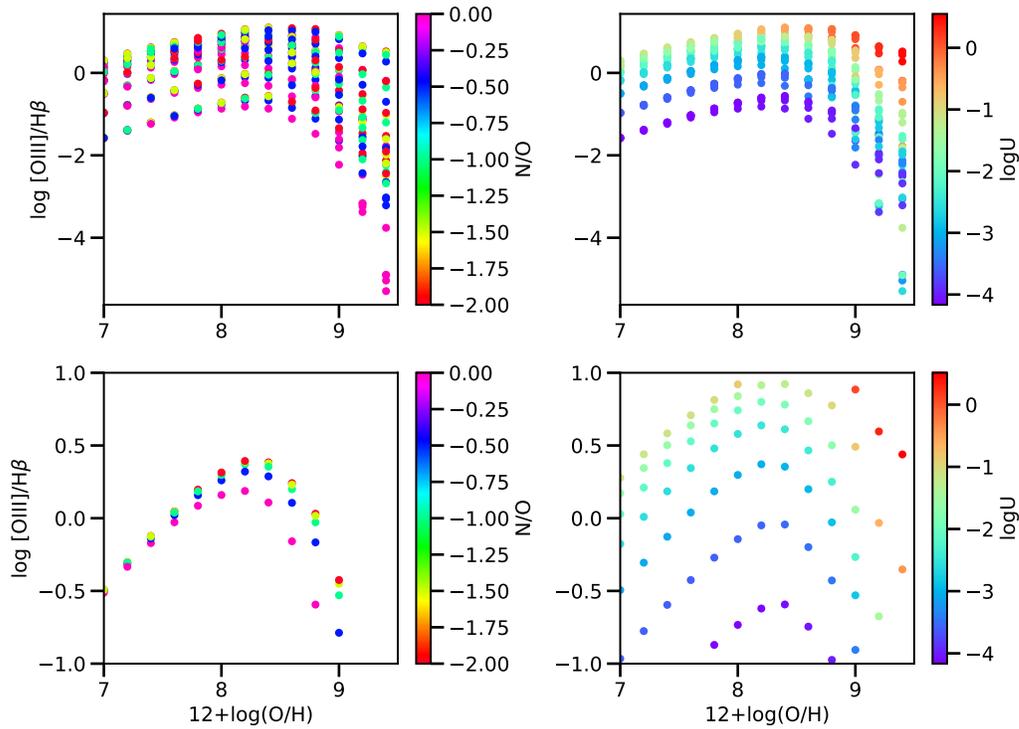


Figura 1.11: O/H vs $[\text{OIII}]/\text{H}\beta$ con N/O y $\log U$ del grid de modelos BOND (Vale Asari et al., 2016). Las gráficas superiores son todos los modelos disponibles. Las gráfica inferior izquierda nos muestra los modelos que tienen un valor de $\log U = -3 \pm 0.1$ y el N/O es libre. En la gráfica inferior derecha son los modelos con $\text{N/O} = -1.0 \pm 0.3$ y $\log U$ libre. Es decir, los modelos inferiores es un acercamiento a las gráficas superiores.

la densidad electrónica. En la fig. 1.11 se tienen los valores de O/H vs $[\text{OIII}]/\text{H}\beta$ de la base de modelos BOND (Vale Asari et al., 2016). Se tomaron los datos de los modelos que tienen una población estelar ionizante con una edad de 1Myr, un $f_r = 0.03$, o sea una cáscara delgada, y una fracción de $\text{H}\beta > 0.9$. En la gráfica superior izquierda se toman todos los valores disponibles. El código de colores representa el valor de N/O. Es decir, en esta gráfica vemos un conjunto de curvas análogas a la mostrada en la fig.1.10. El efecto de tener como parámetro libre N/O es tener un conjunto de curvas, por lo tanto, un gran número de posibles soluciones.

En la gráfica superior derecha se tienen los mismos modelos de la gráfica anteriormente mencionada pero ahora el parámetro es $\log U$. Al igual que en el caso de N/O al tener $\log U$

libre se tiene una serie de curvas en este gráfico. Por lo tanto, al variar $\log U$ igualmente se tienen un conjunto de soluciones posibles. Podemos ver como en el caso de N/O libre las curvas se entrecruzan formando una malla. Mientras para el caso de $\log U$ las líneas siguen una tendencia.

Tomemos valores alrededor de $N/O = -1.0 \pm 0.3$ y $\log U = -3 \pm 0.1$. En la gráfica inferior izquierda se muestran los valores seleccionados para N/O . Aquí observamos que en la zona de baja metalicidad las curvas son cercanas mientras que en el régimen de alta metalicidad las curvas se separan considerablemente con respecto a la otra zona. Esto provocará que se tengan dos conjuntos de soluciones. Un conjunto será menos disperso que el otro pero si tenemos la suficiente resolución podríamos identificar varias soluciones que cumplan con las observaciones.

Por otro lado en la gráfica inferior derecha es el caso donde $\log U$ es libre. Se observa cómo las curvas son bien definidas y no se sobreponen al menos en el rango mostrado. Sin embargo sí causa que se tengan un conjunto de soluciones. En el caso en que N/O y $\log U$ sean libres simultáneamente difícilmente podremos decir el número exacto de soluciones posibles. Podemos identificar a gran escala al menos dos conjuntos de soluciones pero si la resolución aumenta encontraríamos un gran número de soluciones posibles que reproduzcan $[OIII]/H\beta$.

En la realidad astrofísica, existe una relación entre N/O y O/H de un lado y entre $\log U$ y O/H de otro lado, así se puede eliminar este problema. Se observa que N/O aumenta conforme O/H aumenta, al menos en altas metalicidades. Pero estas dos relaciones no se han determinado de manera clara. En este trabajo queremos explorar la problemática sin suponer estas relaciones.

Capítulo 2

CALIFA

§2.1. Introducción

En el capítulo 3 se hablará sobre modelos de fotoionización. Estos modelos deben de ser comparados con observaciones para garantizar la reproducción de objetos astrofísicos reales. Las observaciones utilizadas para este trabajo fueron tomadas del survey *CALIFA* (Calar Alto Legacy Integral Field Area Survey).

El *survey* CALIFA es una muestra observable con espectroscopía de campo integral de ~ 600 galaxias del universo local. Las observaciones de CALIFA fueron hechas en el telescopio de 3.5m del observatorio de Calar Alto con el espectrofotómetro de campo integral PMAS (Roth et al., 2005) en la configuración PPAK (Kelz et al., 2006). El sistema de fibras PPAK consiste de un conjunto de 382 fibras, de las cuales 331 cubren un campo de visión hexagonal de $74'' \times 64''$ como se ve en la fig. 2.1. Las fibras 36 restantes son utilizadas para muestrear el cielo de fondo. Estas fibras están localizadas a una distancia de $\sim 75''$ del centro del campo de visión.

En la fig. 2.1 se muestra que entre las fibras hay un pequeño espacio, esto se soluciona con un ligero movimiento de las fibras entorno a las coordenadas nominales de cada fibra. En CALIFA se utilizaron los siguientes *offsets* en ascensión recta y declinación siguientes: $0'', 0''$; $-5.22'', -4.84''$ y $0 - 5.22'', +4.84''$. Al final, en la reducción de datos, estos diferentes *offsets* serán combinados espacialmente en un solo conjunto de datos estándar.

Como se describe en Sánchez et al. (2012a), CALIFA tiene dos configuraciones de diferente resolución: una de alta y otra de baja resolución. La configuración de baja resolución muestrea de 3745\AA a 7500\AA , con una resolución espectroscópica de $\lambda/\delta\lambda \sim 850$. Esta con-

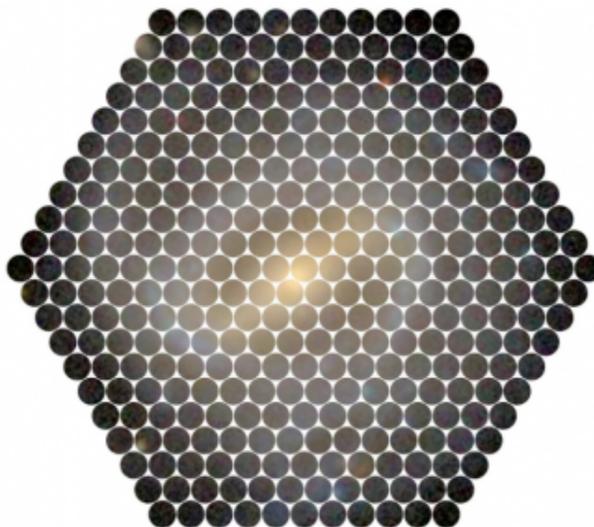


Figura 2.1: Imagen de una galaxia obtenida por CALIFA. Imagen:CALIFA Survey

figuración cubre el intervalo en longitudes de onda donde están las líneas de emisión (de $[\text{OII}]\lambda 3727$ a $[\text{SII}]\lambda 6731$) más significativas de un gas ionizado típico.

§2.2. Detección de Regiones HII

Una vez que los datos observados son procesados por el data ducto de CALIFA se genera un cubo de datos. Las coordenadas x y y son la ascensión recta y la declinación respectivamente. En la tercera coordenada z , se tiene el espectro. A partir de estos cubos de datos se tienen que detectar las regiones HII y extraerlas para trabajar con ellas.

Para esta tarea se utilizó el paquete HIIEXPLORER (Sánchez et al., 2012b). HIIEXPLORER es un procedimiento semi automático que esta basado en dos suposiciones:

- a) Las regiones HII son objetos con una fuerte emisión del gas ionizado que es claramente distinguible de la emisión del continuo. Esto puede ser utilizado para la línea de H_α que es muy intensa en estas estructuras.
- b) Las regiones HII tienen una extensión de algunos cientos de parsec, lo que equivale a un tamaño proyectado de algunos arcsec.

Dado un cubo de datos de CALIFA, podemos obtener el mapa de emisión de H_α para todo la

galaxia. Bajo la suposición a), podemos identificar el pixel más brillante que correspondería a una región HII. Se tiene que dar un límite inferior en el flujo absoluto para que no tomemos en cuenta emisión que puedan interferir en la búsqueda de este pixel de brillo máximo. Una vez localizado el pixel más brillante, se procede a guardarlo como la coordenada del centro de la región HII asignándole un ID.

Una vez identificado el centro de una región HII, se procede a identificar que tan extensa es. Esto se logra agregando pixels de alrededor del centro identificado siguiendo tres criterios. El primero es una distancia máxima de un pixel asociado a una región HII. Esta distancia va del centro identificado al pixel dentro de la cual se agregará a la región HII. Al sobrepasar esta distancia, dada por el usuario, ya no se agregarán más píxeles a la actual región HII.

La segunda condición es la intensidad mínima relativa al pico de emisión de una región HII. Por debajo de este límite ya no se agregarán más píxeles a la región HII. Por último se tiene un límite absoluto de la emisión de la región HII. Estos dos últimos límites también son dados por el usuario.

Si el pixel prospecto cumple con las tres condiciones anteriores, es agregado a la región HII. De forma iterativa se siguen agregando pixels que forman la región HII. Cuando alguna de las tres condiciones no es cumplida por los pixels circundantes, ya no se agregan más pixels terminando la identificación de la actual región HII. Dado que pueden existir más regiones HII, se enmascaran las regiones HII previamente identificadas y se vuelve al inicio del proceso, es decir, a identificar un pico de emisión intenso. Una vez que ya no es posible encontrar un nuevo pico de emisión se finaliza el proceso y se crea una máscara de segmentación de regiones HII.

Para las regiones HII extragalácticas utilizadas en este trabajo, se utilizó una distancia máxima a la que pueden estar los pixels de 3.5 spaxels (*spacial pixels*), el equivalente a 500 pc. Esta distancia representa el tamaño típico de una región HII de algunos cientos de parsecs (López et al., 2011; González Delgado & Pérez, 1997). Como límite relativo al pico de emisión se tomo del 10 %, es decir cada pixel que se agrega debe de tener un flujo mayor al 10 % del flujo del pico de emisión. Y la densidad de flujo mínima para el pico de intensidad de $2 \times 10^{-17} \text{erg s}^{-1} \text{cm}^{-2} \text{arcsec}^{-1}$.

En la fig. 2.2 se observa el mapa de H_{α} que es el input para HIIEXPLORER, a este mapa se le aplica el proceso antes descrito. La salida que nos entrega HIIEXPLORER es un mapa de segmentación que contiene todas las regiones HII identificadas. Además nos entrega las posiciones espaciales (ascensión recta y declinación) de cada una de las regiones HII.

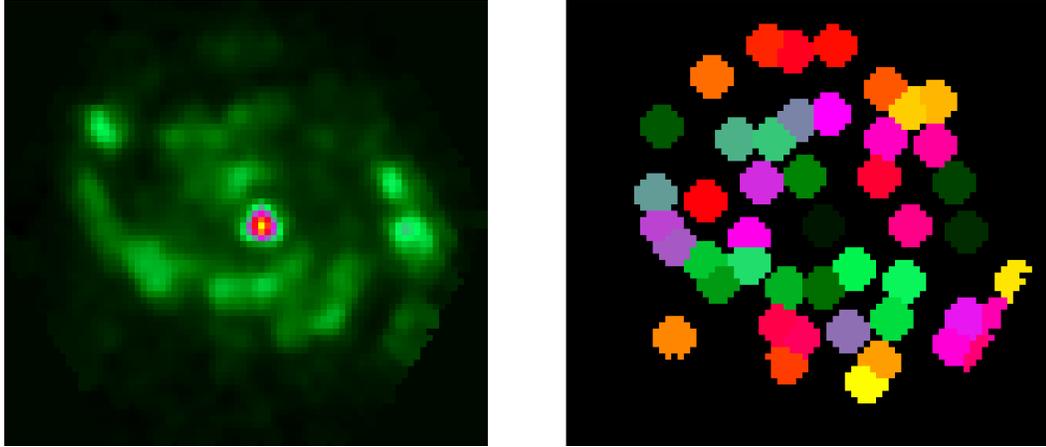


Figura 2.2: Input y salida de HIIEXPLORER. Izq.) Mapa de H_{α} de la galaxia NGC5947 input de HIIEXPLORER. Der.) Regiones HII detectadas con HIIEXPLORER de la galaxia NGC5947. Cada punto de color es candidato a ser una región HII.

§2.2.1. Ajuste del espectro

Una vez identificadas las regiones HII se extrae el espectro de dicha región. Para esto se co-suman todos los espectros en los cubos de datos originales que estén en la misma región HII. A este espectro resultante se modela el continuo estelar usando FIT3D (Sánchez et al., 2016).

La luz recibida del cual se extrae el espectro contiene tanto luz que viene de la población estelar como la luz emitida por el gas ionizado. La población estelar domina el continuo del espectro y genera líneas de absorción. Por otro lado, el gas ionizado genera líneas de emisión a ciertas longitudes de onda. Como resultado, los espectros que se obtienen de las regiones HII tiene líneas de emisión y de absorción además de un continuo dominado por la población estelar.

FIT3D ajusta estas dos componentes principales: la población estelar y las líneas de emisión de manera independiente. Además de ajustar estas dos componentes, también toma en cuenta la atenuación por polvo que está distribuido a lo largo de la galaxia, por ejemplo Tuffs et al. (2004).

Para cada uno de los dos ajustes, poblaciones estelares y líneas de emisión, se realizan ajuste de parámetros no lineales y lineales. Para el caso de las poblaciones estelares, los parámetros no lineales son aquellos que caracterizan las propiedades cinemáticas y la aten-

uación por polvo del espectro. Los parámetros lineales están relacionados al ajuste de la descomposición estelar subyacente. Los parámetros no lineales son calculados primero de manera independiente a los parámetros lineales.

Ajuste de poblaciones estelares. Parámetros no lineales

Las propiedades cinemáticas se ajustan bajo la suposición de que todas las poblaciones estelares que ionizan al gas se mueven a la misma velocidad. Además se toma que la dispersión de velocidades sigue una función Gaussiana (Rix & White, 1992).

El primer parámetro que se ajusta es la velocidad estelar. Dado un rango de valores dentro del cual está el valor de la velocidad estelar se hace una exploración tomando valores al azar. Estos valores aleatorios son generados a partir de un paso fijo. Este proceso se realiza varias veces para un mejor ajuste.

Para el ajuste de la dispersión de velocidades de las estrellas que ionizan al gas se inicia de un valor dado y la velocidad se cambia con un paso aleatorio. Para cada velocidad se calcula una población estelar que ajuste el continuo. Utilizando la mejor velocidad calculada en el proceso anterior se ajusta la dispersión de velocidades. Este proceso se repite dos veces para un mejor resultado.

De la misma forma, la atenuación por polvo, A_v , se calcula dentro de un rango dado con un paso aleatorio. Para cada valor se deriva la mejor combinación de poblaciones estelares análogamente al proceso descrito anteriormente.

Para calcular las poblaciones estelares que originan el espectro observado se utiliza una librería la cual contiene una muestra de espectros a diferentes edades y metalicidades. De esta dependerá el análisis del espectro. Para el ajuste de velocidad y dispersión de velocidades se utiliza una librería que contiene menos poblaciones estelares para acelerar el cálculo. Otra manera de acelerar el proceso es tomando solamente regiones del espectro que tengan una absorción estelar importante.

Ajuste de poblaciones estelares. Parámetros lineales

Con la velocidad, dispersión de velocidades y atenuación por el polvo calculadas, se procede al ajuste de poblaciones estelares. Este ajuste está basado en una combinación de *templates* de poblaciones estelares (*SSP* por *single stellar population*).

El template es un muestreo de SSP a diferentes edades y metalicidades. Por lo tanto, para que se tenga un buen ajuste el template usado debe de ser lo más completo posible. Es

Edades			Metalicidades Z/Z_{\odot}
1.00 Myr	3.00 Myr	4.00 Myr	
5.6 Myr	8.9 Myr	10 Myr	
12.6 Myr	14.1 Myr	17.8 Myr	
20.0 Myr	25.1 Myr	31.6 Myr	0.4
40 Myr	56.2 Myr	63 Myr	
63.1 Myr	70.8 Myr	100 Myr	
112 Myr	126 Myr	158.5 Myr	1
200 Myr	281 Myr	355 Myr	
500 Myr	708 Myr	891 Myr	1.5
1.12 Gyr	1.25 Gyr	1.4 Gyr	
2 Gyr	2.5 Gyr	3.5 Gyr	
4.5 Gyr	6.3 Gyr	8 Gyr	
10 Gyr	12.6 Gyr	14 Gyr	

Tabla 2.1: Valores de edad y metalicidad de SSP del template *gsd156*. (Cid Fernandes et al., 2013)

por esto que es complicado dado que los datos pueden ser el resultado de una combinación de un SSP. En Sánchez et al. (2016) se probaron diversos templates. Para nuestro trabajo se utilizó el template *gsd156* (Cid Fernandes et al., 2013).

El template *gsd156* consiste en una muestra de cuatro metalicidades y 39 edades entre 1 Myr y 14 Gyr que se muestran en la tabla 2.1. Este template es explicado y probado ampliamente en Cid Fernandes et al. (2013).

Con este template se implementó la técnica de ajuste multi población estelar explicado a detalle en Sánchez et al. (2011, 2006).

El ajuste se basa en el método Monte-Carlo. Se realizan K procesos Monte Carlo donde el espectro modelado final y sus incertidumbres esta dado por el promedio y desviación estándar respectivamente de cada iteración.

En cada iteración se crea una versión Monte Carlo de los datos. El flujo observado a cierta longitud de onda, O_i , más el nivel de ruido σ_i correspondiente multiplicado por un valor aleatorio R_i entre -1 y 1 crea el flujo observado Monte Carlo G_i^k .

$$G_i^k = O_i + R_i \sigma_i. \quad (2.1)$$

Por otra parte, el template de poblaciones estelares es ajustado a la velocidad sistémica del espectro observado, se convolucionan con la dispersión de velocidades y se le aplica la

atenuación del polvo anteriormente derivada. A un flujo de la i -ésima longitud de onda de un j -ésimo template después de aplicar la atenuación por el polvo se define como F_{ji}^{Av}

Se realiza un ajuste lineal de mínimos cuadrados de tal manera que χ^2 sea minimizado, donde se usó:

$$\chi^2 = \frac{1}{N - M} \sum_{i=1}^N D_i^2, \quad (2.2)$$

donde

$$D_i = w_i \left(G_i^k - \sum_{j=1}^M a_j F_{ji}^{Av} \right), \quad (2.3)$$

donde w_i es el peso del i -ésimo pixel tal que

$$w_i = \frac{1}{\sigma_i^2}; \quad (2.4)$$

a_j es el coeficiente del j -ésimo template del espectro modelado, M es el número de template usados en el proceso de ajuste y N es el número de elementos.

Pueden existir $a_j < 0$ los cuales son excluidos en la siguiente iteración del proceso de ajuste por lo que el número de templates usados se va reduciendo. Las iteraciones paran cuando todos los coeficientes a_j son positivos.

El espectro final modelado para el proceso k Monte Carlo entonces es

$$S_i^k = \sum_{j=1}^M a_j^k F_{ji}^{Av}. \quad (2.5)$$

Una vez teniendo este espectro modelado se realiza otro proceso Monte Carlo el cual inicia creando otra versión Monte Carlo de los datos.

El espectro final se obtiene como:

$$S_i = \text{promedio}(S_i^{1..K}). \quad (2.6)$$

Y la incertidumbre se calcula como

$$\sigma_{S_i} = \text{DesvEst}(S_i^{1..K}). \quad (2.7)$$

Se encuentra que si se calcula χ^2 con este último modelo no es muy diferente a los valores

de χ^2 calculados para los modelos individuales encontrados en el proceso.

De la misma manera se encuentran los coeficientes finales a_j de los componentes del template de la población estelar como

$$a_i = \text{promedio}(a_i^{1..K}), \quad (2.8)$$

$$\sigma_{a_i} = \text{DesvEst}(a_i^{1..K}). \quad (2.9)$$

Para que el proceso se realice de manera correcta algunas zonas del espectro observado se deben enmascarar para no ser tomadas en cuenta. Se deben enmascarar las zonas donde haya fuertes líneas del cielo, zonas afectadas por los defectos de los instrumentos (CCD), así como las regiones donde hay absorciones telúricas y donde hay fuertes líneas de emisión.

Una vez calculado el ajuste del espectro se pueden calcular diversas propiedades en promedio de las poblaciones estelares así como su edad o su metalicidad, entre otras. Pero para nuestro estudio necesitamos las intensidades de las líneas de emisión. Este proceso también es realizado por FIT3D.

Tomemos en cuenta que se tiene un espectro ajustado de la población estelar S_i . Para caracterizar las líneas de emisión se debe de sustraer del espectro observado O_i el espectro S_i . De esta manera, obtenemos un espectro de la emisión del gas C_i limpio de las componentes estelares. Es por esto que se necesita que el ajuste de la población estelar subyacente sea lo mejor posible.

Cada línea es ajustada a una función gaussiana y un polinomio de bajo orden. Al igual que en el ajuste de la población estelar adyacente el proceso de ajuste de las líneas de emisión es dividido en dos: componente lineales y no lineales.

Las componentes no lineales, como en el caso anterior, son la velocidad y la dispersión de velocidades. Estas componentes se buscan de manera similar a la descrita anteriormente. Se busca primero la velocidad usando una exploración pseudo aleatoria fijando la dispersión de velocidades. Una vez derivada la velocidad se fija e inicia el proceso de ajuste de la dispersión de velocidades.

En cada paso de la exploración pseudo aleatoria se realiza el ajuste de mínimos cuadrados para derivar los parámetros lineales que en este caso es la intensidad de cada línea y los coeficientes de la función polinomial. Al igual que en el ajuste de la población estelar se busca minimizar χ^2 . Los mejores valores de los parámetros lineales y no lineales darán como resultado un valor mínimo de χ^2 .

Para la estimación de errores se realizan simulaciones Monte Carlo. Análogamente al caso anterior se tiene para la iteración k y una longitud de onda λ_i

$$C_i^k = C_i + R_i \sigma_{mod}, \quad (2.10)$$

donde

$$\sigma_{mod} = \sqrt{\sigma_i^2 + \sigma_{S_i}^2}; \quad (2.11)$$

σ_i^2 es el error original que fue usado en el ajuste de las poblaciones estelares y $\sigma_{S_i}^2$ es la desviación estándar del mejor modelo del ajuste de poblaciones estelares.

El ajuste de las líneas de emisión se hace por intervalos de longitud de onda en lugar de ajustar todas a la vez. Esto simplifica el ajuste del continuo por un polinomio de grado bajo y el ajuste de las líneas seleccionadas.

De igual manera se crea un espectro de líneas de emisión utilizando una combinación de funciones gaussianas dado por

$$E_i = \sum_{k=1}^L B_k * Gauss_{ik}(\lambda_k, \sigma_k), \quad (2.12)$$

donde L es el número de líneas en el modelo, B_k es el flujo integrado de la k -ésima línea de emisión y $Gauss_{ik}$ es la función gaussiana normalizada en el i -ésimo pixel.

En este punto se puede sustraer este espectro E_i del espectro original observado G_i y con el espectro resultante calcular de nuevo la población estelar subyacente y repetir el proceso cuantas veces se crea necesario. Aunque en una tercera iteración no se producen cambios importantes.

Con este ajuste obtenemos dos cosas que necesitamos para nuestros modelos. La primera es la descripción de la población estelar subyacente a nuestras regiones HII. La segunda es la intensidad de las líneas de emisión observadas y por lo tanto razones de líneas para poder comparar nuestros modelos.

Capítulo 3

Modelos de fotoionización

§3.1. Introducción

El estudio de regiones ionizadas puede ser realizado a través del análisis del espectro que emiten. De estos podemos extraer información de las líneas de emisión que se observan. A partir de estas se puede determinar algunas propiedades físicas de la nube, por ejemplo, densidad o temperatura (Stasińska, 2009).

Otra manera de estudiar la física en las nebulosas es por medio de los modelos de fotoionización. Los modelos de fotoionización toman en cuenta las ecuaciones de equilibrio de ionización y de equilibrio térmico que se revisaron en el capítulo 1. Estas ecuaciones son complicadas de resolver pero gracias al avance constante de las computadoras y de su capacidad de cálculo se han desarrollado códigos de fotoionización que resuelven estas ecuaciones de manera rápida y eficiente.

Estos códigos toman una fuente ionizante, una distribución de densidad del gas y una composición química como parámetros de entrada. El programa devolverá como resultado un espectro sintético, fracciones iónicas, temperatura electrónica, etc. Con estos códigos nos es posible ver como el espectro emitido por la nebulosa y algunas propiedades físicas de ésta son afectadas por diversos factores, como lo es el espectro ionizante, la estructura de la nebulosa y su composición química.

Los códigos de fotoionización son usados para el estudio de diversos objetos. Son usados principalmente para modelar nebulosas planetarias (Bohigas, 2008) o regiones HII (Morisset et al., 2016). Pueden ser utilizados para el estudio de AGN (Thomas et al., 2016) y sus distintas zonas, por ejemplo las *broad regions* (Leighly & Casebeer, 2007). Los modelos

de fotoionización son usados para entender la física en distintos objetos y con diferentes propiedades (Kallman, 2010; Patel et al., 2016) .

Existen diversos códigos de fotoionización como Cloudy (Ferland et al., 2013), MAPPINGS IV (Dopita et al., 2013) y MOCASSIN (Ercolano et al., 2003). Estos códigos son los más usados y cada uno de ellos tienen características propias. En este trabajo se utilizó el código Cloudy junto a la herramienta desarrollada en python: pyCloudy (Morisset, 2013, 2014).

El paquete PYCLOUDY permite un manejo sencillo de los archivos de entrada y salida de Cloudy. Es capaz de leer los archivos de salida de Cloudy e imprimir y realizar gráficas de manera muy sencilla. Por otra parte, al poder generar archivos de entrada para Cloudy permite crear fácilmente mallas de un gran número de modelos. Esto se puede automatizar de tal manera que con unas cuantas líneas en python se realizan cientos y hasta miles de modelos en un proceso automatizado.

Cloudy es un código en una dimensión a diferencia de MOCASSIN que realiza cálculos en tres dimensiones. Con PYCLOUDY podemos realizar modelos pseudo 3D. Esto es posible dado que PYCLOUDY genera una malla de modelos cambiando el ángulo de la línea de visión del modelo. Es decir crea diferentes archivos de entrada para Cloudy a diferentes ángulos. Los archivos de salida correspondientes son unidos e interpolados logrando crear un modelo en tres dimensiones (Gesicki et al., 2016).

Además de manejar archivos de entrada y salida de Cloudy el paquete PYCLOUDY incorpora el modulo para controlar la base de datos 3MdB (Morisset et al., 2015). 3MdB es una base con cerca de 700000 de modelos de fotoionización públicos los cuales contienen muchos parámetros de salida de útiles. Esta base de datos está hecha bajo un protocolo MySQL lo cual permite un fácil acceso y esta en constante crecimiento.

§3.2. *Inputs y Outputs* de los códigos de fotoionización

Cloudy, y en general un código de fotoionización, necesita algunos parámetros de entrada para poder realizar los cálculos correspondientes. Estos parámetros son una descripción del campo de radiación ionizante y algunas de las propiedades del gas ionizado. Para poder describir la fuente ionizante o su distribución espectral de energía tenemos que proporcionar la intensidad de su radiación y las propiedades de la estrella o estrellas correspondientes que emiten el espectro ionizante.

El gas puede ser ionizado por una o varias estrellas. En este caso los parámetros que

las describen son: T_* , $\log g$ y Z . También puede ser un cúmulo de estrellas el que esté ionizando al gas. Los parámetros para describir un cúmulo ionizante es la edad del brote y su metalicidad. No obstante, el gas puede ser ionizado por una combinación de estrellas y cúmulos.

Por otra parte, los parámetros que describen el gas son su densidad y la composición química. Además, se debe de dar a que distancia se encuentra el gas de la fuente ionizante y su morfología. La morfología describe si tenemos una esfera llena de gas o un cascarón de gas. El primer caso corresponde a una capa de gas considerablemente gruesa y el segundo es una capa delgada de gas donde se puede usar una aproximación plano paralelo.

La salida de Cloudy incluye muchos datos útiles del gas ionizado, así como su espectro generado por el gas, además de proporcionar las emisividades de muchas líneas. En la tabla 3.1 se muestran los *inputs* y *outputs* en concreto de nuestros modelos.

Los modelos de fotoionización son usados frecuentemente para reproducir las observaciones de algunos objetos del universo. Al modelar de esta manera las observaciones nos permiten poder describir que condiciones están presentes en estas regiones así como las características de la fuente ionizante.

Dado que el código de fotoionización nos da un espectro sintético podemos obtener las intensidades de las líneas para comparar con los datos observados. Estos datos obtenidos del modelo dependerá de la calidad de los datos atómicos que se usen en el cálculo.

§3.3. Modelado de regiones HII de CALIFA

Las observaciones utilizadas para este trabajo fueron tomadas del *survey* CALIFA. CALIFA tiene observaciones de muchas regiones HII gigantes extragalácticas las cuales podemos modelar. En un trabajo previo (Morisset et al., 2016) se realizó el modelado de las regiones HII del *survey* CALIFA.

CALIFA proporciona líneas de emisión fuerte e intensidades de línea de [OII] λ 3727 a [SII] λ 6731 para un poco más de 18000 regiones de 612 galaxias. Además de las líneas, CALIFA también provee de una descripción de la población estelar que ioniza a las regiones que puede ser utilizado como *input* para el modelo de fotoionización. A pesar del gran catálogo de líneas disponibles, solo se usaron las líneas de emisión que estuvieran presente en casi todas las fuentes: H_α , H_β , [NII] λ 6584Å, [OII] λ 3726 + 29Å y [OIII] λ 5007Å. Además, se contaba también con las líneas de [OI] λ 6300Å, [SII] λ 6716 + 31Å y [SIII] λ 6312Å.

Para poder reproducir las observaciones de estas regiones necesitamos dar los paráme-

Input	Output
Espacio de Parámetros	Espacio de Observables
<ul style="list-style-type: none"> ▪ SED <ul style="list-style-type: none"> • Forma (Edad, T_*) • Intensidad ▪ Gas <ul style="list-style-type: none"> • Densidad • Abundancias <ul style="list-style-type: none"> ◦ O/H ◦ N/O ◦ Etc. ▪ Distancia a la estrella ▪ Morfología <ul style="list-style-type: none"> • Esfera llena • Cascarón 	<ul style="list-style-type: none"> ▪ H_α, H_β ▪ [NII]6584 ▪ [OII]3727 ▪ [OIII]5007

Tabla 3.1: Inputs y outputs para nuestros modelos de fotoionización en este trabajo

tros adecuados a nuestro código de fotoionización, en este caso Cloudy. En la tabla 3.1 vemos todo lo que requerimos. La descripción de la SED viene dada por FIT3D (Sánchez et al., 2016) el cual ajusta la población estelar subyacente a las regiones HII. Con estos datos se calcula la SED con el código POPSTAR (Mollá et al., 2009). La SED total es la suma de cada una de las poblaciones estelares ajustadas con su respectivo peso calculado por FIT3D. Por lo tanto, de la descripción de la SED sólo queda como parámetro libre su intensidad.

Para la descripción del gas necesitamos su densidad y su composición química. La composición química de una nebulosa incluye toda la abundancia de los elementos químicos en ella. Como indicador de metalicidad se usó O/H y N/O. Estas dos abundancias fueron las únicas tomadas en cuenta en este trabajo como parámetros libres ya que todas las otras abundancias se determinaron a partir de O/H.

Además de la densidad y composición química del gas, necesitamos saber a qué distancia está de la estrella y su morfología. Para tomar dos casos representativos de la morfología se utilizó el factor de morfológico $f_r = R_{in}/R_{str}$. Donde R_{in} es el radio interno de la nebulosa y R_{str} el radio de Strömgren dado por la expresión 1.10. Se usaron dos valores de factor morfológico, $f_r = 3.0$ que representa una esfera delgada y $f_r = 0.3$ que representa una esfera llena.

Tomando todo lo anterior en cuenta se reducen el número de parámetros libres a: la intensidad de la SED, la abundancia O/H y N/O, la densidad del gas y la distancia entre la población estelar ionizante y el gas. Un total de 5 parámetros libres que hay que ajustarlos al mismo tiempo. Para reducir este espacio de parámetros incluimos el parámetro de ionización U . El parámetro de ionización se define como:

$$U(r) = \frac{Q(H^0)}{4\pi r^2 N_H c}. \quad (3.1)$$

Donde $Q(H^0)$ es el número de fotones ionizantes emitidos por la población estelar por unidad de tiempo, r es la distancia entre el gas y la fuente ionizante, N_H la densidad de hidrógeno y c la velocidad de la luz. Con el parámetro de ionización tomamos tres parámetros libres y lo transformamos en un solo parámetro libre. Aunque se ocupó el parámetro de ionización promedio sobre todo el volumen de la nebulosa, \tilde{U} .

Con U y f_r podemos calcular la cantidad de fotones ionizantes Q_0 y el radio interno

R_{in} . Para el caso de Q_0 tenemos que:

$$Q_0 = \frac{4\pi c^3 \tilde{U}}{3N_H f f^2 \alpha_B w^3}, \quad (3.2)$$

donde c es la velocidad de la luz, N_H es la densidad de hidrógeno, ff es el *filling factor*, α_B es el coeficiente de recombinación caso B y $w = (1 + fr^3)^{1/3} - fr$. Con Q_0 se calcula el radio de Strömgen que en este caso es:

$$R_{Str} = \left[\frac{3Q_H}{4\pi N_H^2 \alpha_B ff} \right]^{1/3}. \quad (3.3)$$

Con el radio de Strömgen y f_r calculamos el radio interno R_{in} tal que

$$R_{in} = f_r R_{Str}. \quad (3.4)$$

Tenemos tres parámetros libres: U , O/H y N/O. En el trabajo previo de Morisset et al. (2016), se restringió un parámetro más: O/H. El O/H es un indicador de metalicidad que se fijó su valor usando un método de líneas fuertes realizado por Marino et al. (2013). Este calibrador utiliza el cociente de líneas ($[OIII]/H\beta$)/($[NII]/H\alpha$) para obtener el valor de O/H.

Finalmente sólo se tuvieron dos parámetros libres: U y N/O. Se realizó una malla de modelos variando estos dos parámetros para encontrar el valor observado de los cocientes de líneas: $[OIII]/H\beta$ y $[NII]/H\alpha$.

De todos los modelos que se realizaron se identificaron por las poblaciones estelares ionizantes. Del trabajo de Morisset et al. (2015) se pueden definir límites en la abundancia O/H y la edad de las poblaciones estelares. Se clasifican en dos poblaciones estelares: en poblaciones jóvenes (estrellas OB) y en poblaciones de gran edad, los cuales corresponden a HOLMES (HOT Low Mass Evolved Stars, Flores-Fajardo et al. (2011)). Los HOLMES son capaces de producir radiación que ionice el gas circundante. El análisis hecho por FIT3D da el mejor ajuste de acuerdo a la librería gsd159. Esta librería contiene desde estrellas jóvenes a estrellas viejas, por lo que el ajuste no solamente incluye estrellas jóvenes OB.

Si tomamos $Q(H^0)$ el número de fotones ionizantes de H^0 y $Q(He^0)$ el número de fotones ionizantes de He^0 . El cociente $Q_{0/1} = Q(H^0)/Q(He^0)$ está correlacionado con el tipo de población estelar dominante que ioniza el gas, por lo explicado en el cap II. En la Fig. 3.1 podemos ver el número de modelos que corresponden a una población tipo

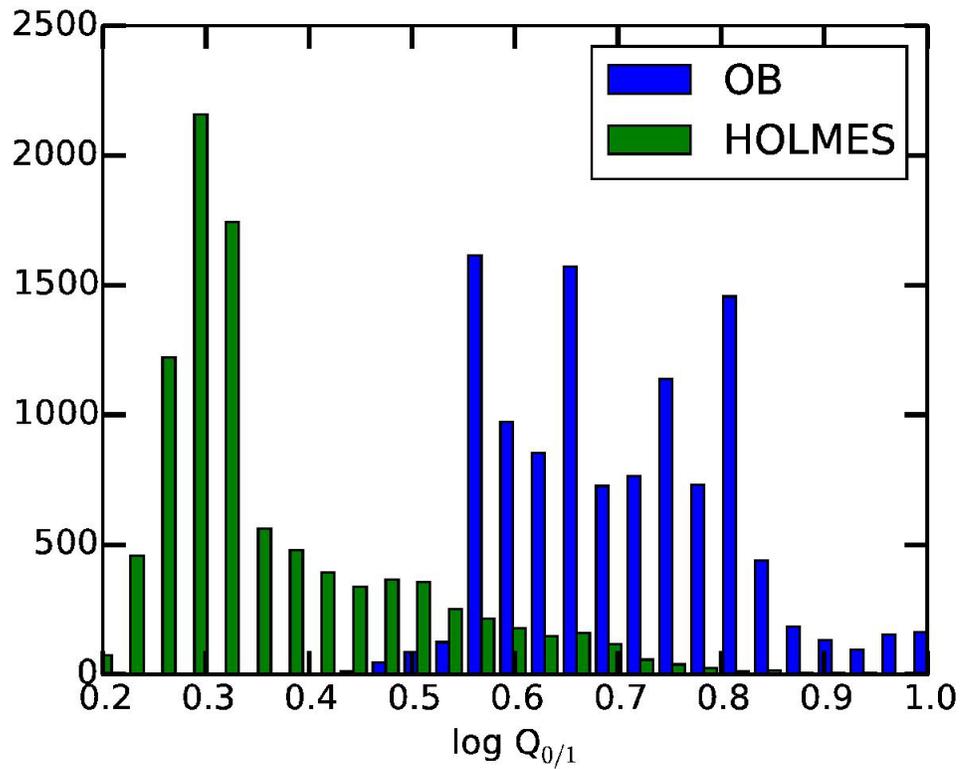


Figura 3.1: Identificación de poblaciones estelares: OB y HOLMES para las regiones de gas ionizadas de CALIFA. El valor $Q_{0/1} = Q(\text{H}^0)/Q(\text{He}^0)$, el cociente del número de fotones ionizantes de H^0 y He^0 está relacionado con la proporción de fotones ionizantes que vienen de estrellas OB del número total de fotones ionizantes $Q(\text{H}^0)$ (Morisset et al., 2016)

OB o HOLMES, donde vemos que hay una clara separación del tipo de población estelar dominante alrededor de $Q_{0/1} \approx 0.55$

§3.4. Modelos con tres parámetros libres

Tomando como base el trabajo descrito anteriormente (Morisset et al., 2016) se hicieron nuevos modelos con tres parámetros libres. Dos parámetros libres son la abundancia de nitrógeno (N/O) y el parámetro de ionización, $\log U$. El tercer parámetro en este trabajo es la abundancia de oxígeno (O/H) que en el trabajo anterior se fijó usando la relación de Marino et al. (2013).

Además de los tres parámetros libres, O/H, N/O y $\log U$, los modelos de fotoionización en este trabajo fueron realizados con base en los modelos del trabajo anterior. Se utilizaron dos valores para $f_r = 0.3$ y $f_r = 3.0$, que corresponden a una esfera llena de gas y a un cascarón de gas. La población estelar ionizante de las regiones HII son las ajustadas por medio de FIT3D (Sánchez et al., 2016). Por otra parte, se tomó que la composición química de las regiones HII relativamente a O/H está dada por Asplund et al. (2009). Para la abundancia de He se calculó en base a la abundancia de O y si en dado caso este cálculo daba un valor no válido se tomó que $\text{He} = 0.5$.

De todas las galaxias en el survey CALIFA, en este trabajo se utilizó la galaxia NGC 5947 (Fig. 2.2). De todas las regiones identificadas por HIIEXPLORER no se usaron aquellas que tienen una población estelar vieja, es decir, que son ionizadas por HOLMES. Sólo se utilizaron las regiones ionizadas por estrellas jóvenes, es decir, las regiones HII que en total son 24 regiones.

§3.4.1. Problema de O/H

El tercer parámetro libre de los modelos de fotoionización en este trabajo es O/H. Como se habló en el capítulo 1 hay una dificultad con la determinación de la abundancia de O. En la Fig. 1.11 observamos que la función de la emisividad de la línea $[\text{OIII}]\lambda 5007/\text{H}\beta$ es degenerada con respecto a la abundancia de O. Esto nos dice que para varios valores en la abundancia de oxígeno se obtiene la misma emisividad para esta línea relativamente a $\text{H}\beta$. Como se mencionó en el capítulo 1, este comportamiento se debe a que cuando O/H aumenta, es decir, aumenta la cantidad de oxígeno y la emisión de $[\text{OIII}]\lambda 5007/\text{H}\beta$ aumenta. El oxígeno es un enfriador por lo cual la temperatura, T_e disminuye. La emisividad

de las líneas ópticas decrece con T_e por la disminución de la población del nivel.

Este problema puede resolverse dando alguna calibración observacional. La calibración que se usa es utilizando la línea [OIII] λ 4363, por ejemplo (Stasińska, 2004). Pero en este trabajo se quiere ajustar de manera independiente O/H asumiendo que no hay alguna información observacional, por ejemplo la línea [OIII] λ 4363. Esto causará que tengamos una degeneración en los valores de O/H para la línea [OIII] λ 5007/H β . Y por lo tanto, también pueda haber dos valores de log U y de N/O que ajusten los datos observacionales.

Para encontrar los valores de: log U, N/O y O/H se realizó una malla de modelos. Tenemos que ajustar tres parámetros al mismo tiempo tomando en cuenta que se tienen varias soluciones por cada variable. El ajuste es realizado comparando los cocientes [OII]/H β , [OIII]/H β y [NII]/H α observados con los mismos cocientes modelados. Para automatizar el proceso utilizamos un método genético el cual después de cierto número de iteraciones encuentra las soluciones.

§3.5. Método Genético

Introducción a algoritmo genético

Encontrar los valores de O/H, N/O y log U que reproduzcan las observaciones de las regiones HII de CALIFA es un problema de optimización matemática. Como lo menciona Cantó et al. (2009) las técnicas de optimización se pueden clasificar en dos grupos: deterministas (Horst & Tuy, 1996) y estocásticas (Boender & Romeijn, 1995).

El grupo de técnicas deterministas se basan en encontrar puntos máximos o mínimos de funciones continuas y doblemente diferenciables. Las limitaciones de este tipo de técnicas está al no poder lidiar con funciones no continuas o no diferenciables. Las técnicas estocásticas tienen un aproximamiento probabilístico a este problema. Entre algunas de estas técnicas están los métodos de cúmulos y el cálculo evolutivo que incluye los algoritmos genético, la programación evolutiva, etc.

Los algoritmos genéticos están basados en cierto sentido en la evolución biológica de las especies del planeta. Como lo discute Charbonneau (1995), los algoritmos genéticos se basan en tres hechos de la evolución: selección natural, herencia y variación. La selección natural es el proceso en que los individuos se adaptan al medio ambiente, es decir, se *ajustan*. La herencia es cómo pasan sus genes a su descendencia, es decir, pasan sus cualidades a la generación siguiente para que puedan *ajustarse* al medio ambiente. Con este proceso si el

ambiente permaneciera sin cambios los individuos se adaptarían perfectamente después de cierto tiempo. Pero este no es el caso, el ambiente siempre está cambiando con el tiempo, por lo que debe haber variaciones en el proceso de evolución para que los individuos se puedan adaptar a todo tipo de ambientes.

Claro que el proceso de evolución es mucho más complicado que los lineamientos generales descritos anteriormente, ya que la principal forma en que las especies se reproducen es de forma *sexual*. Esta forma de reproducción no es una forma perfecta de pasar los genes de los progenitores a las crías. Los genes se recombinan para producir una mejor adaptación de las crías. En este punto, las variaciones que se mencionaban son *mutaciones* que modifican los genes de forma aleatoria.

Descripción de un algoritmo genético

Dada la descripción de manera muy general del proceso de evolución de las especies podemos hablar de los lineamientos generales de un algoritmo genético.

En la figura 3.2 se muestra un diagrama de flujo simplificado de un algoritmo genético. El problema es encontrar los máximos, o mínimos, de una función real de N variables. Es decir, queremos encontrar los valores de estas variables dentro de una región V donde la función está definida y la solución buscada existe. En la Fig. 3.2 el primer paso es construir una población aleatoria inicial. La población inicial es un conjunto de puntos ya sea en una malla regular o totalmente aleatorio dentro de la región V . En el contexto del algoritmo genético estos puntos son llamados individuos. Para asegurar que la evolución sea hecha de manera óptima debe haber una cantidad grande de individuos iniciales.

Una vez dada la población inicial se realiza una evaluación de la función objetivo. Esta evaluación se hace si se conoce la función que se quiere ajustar o compararla con algún dato conocido de lo que se quiere ajustar.

El siguiente paso es la selección de los mejores individuos. Al hacer la determinación del ajuste se seleccionan los individuos que mejor ajusten nuestra función objetivo. A estos individuos son llamados *progenitores* bajo el contexto del algoritmo genético. Una vez seleccionados los progenitores se da la reproducción de estos. Anteriormente cuando se habló sobre el origen de los algoritmos genéticos se mencionó que se da la reproducción sexual. Esto conlleva a una recombinación de genes o valores en este caso. También puede usarse una reproducción asexual (Cantó et al., 2009) dependiendo del tipo de problema o estrategia para ajustar la función objetivo.

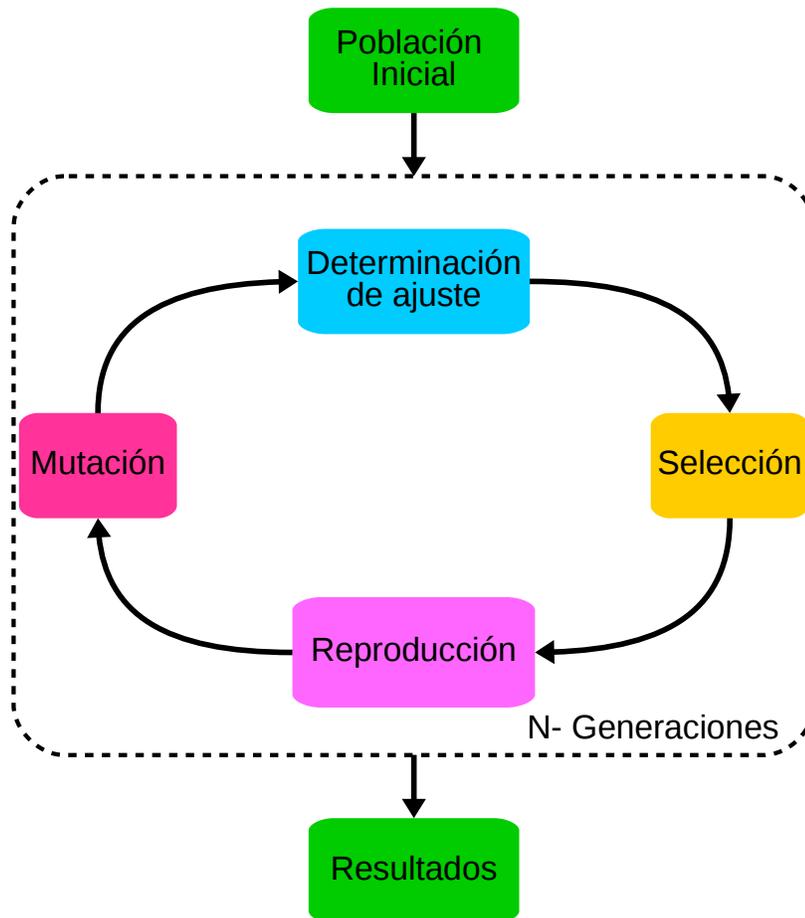


Figura 3.2: Flujo básico de un algoritmo genético. Dada una población inicial se determina que tan bien ajustan algún parámetro o relación dada. Se seleccionan los mejores modelos que cumplan este ajuste. Se reproducen de manera sexual o asexual y por último se realiza una mutación la cual puede ser nula. A partir de este punto se repite el proceso hasta la convergencia.

Si observamos la Fig. 3.2 después de la reproducción viene la mutación. La mutación en un algoritmo genético estándar, se refiere a cambiar de manera aleatoria e independiente de los progenitores el valor de los parámetros de manera análoga al proceso biológico antes mencionado. En este punto se obtienen nuevos individuos llamados *crías* los cuales conforman la nueva malla para iniciar el proceso de nuevo. A cada ciclo se le da el nombre de generación siendo la malla inicial la generación 1 (aunque en algunos algoritmos y/o trabajos prefieren llamarla generación 0 (Cantó et al., 2009)). Se realizarán tantas generaciones como sean necesarias hasta ajustar los parámetros que se requieran.

Aplicando el algoritmo genético a nuestros modelos

La descripción anterior es muy general de un algoritmo genético. Para nuestro problema haremos la distinción entre dos espacios: 1. El espacio de parámetros (Espacio P). 2. El espacio de observables (Espacio O). haciendo referencia a la tabla 3.1. El espacio de parámetros es el espacio tridimensional formado por los parámetros O/H, N/O y log U. Por lo tanto, generar un modelo es generar un punto en el espacio O a partir de un punto en el espacio P.

En nuestro problema no existe una función objetivo conocida pero podemos comparar los modelos con las observaciones de las regiones HII de CALIFA. Al generar el malla inicial de individuos (valores aleatorios de O/H, N/O y log U), estos son utilizados como input para modelos de fotoionización. Estos modelos son hechos con CLOUDY 13.05 agregando los parámetros anteriormente mencionados como f_r y la SED. Al finalizar se calcula $[\text{OIII}]\lambda 5007/\text{H}\beta$, $[\text{NII}]\lambda 6584/\text{H}\alpha$ y $[\text{OII}]\lambda 3727/\text{H}\beta$ de los modelos. Cada modelo o individuo tiene sus parámetros en el espacio P y sus observables en su espacio O. De manera representativa se muestra esto en la Fig. 3.3a.

Al comparar los modelos con las observaciones se seleccionan los mejores modelos. Se seleccionan los modelos que se encuentran a una distancia o radio mínimo dado por el usuario (Fig. 3.3b) donde la observación es representada por un rombo en el espacio O. Dado que estamos comparando con observaciones llamamos a esta distancia *tolerancia*. Todos los modelos que estén dentro de la tolerancia dada son seleccionados para ser los progenitores de la siguiente generación. Idealmente todos los individuos progenitores pueden reproducirse, pero esto significaría un gran número de modelos de fotoionización para la siguiente generación. Para disminuir el tiempo de cómputo de estos modelos se selecciona un número determinado de progenitores, es decir, se seleccionan los mejores n modelos

con respecto a las observaciones.

Al tomar la tolerancia puede hacerse de dos maneras: 1. Definiendo un cubo con una longitud igual a dos veces la tolerancia dada. 2. Definiendo una esfera con un radio igual a la tolerancia proporcionada. Se eligió construir una esfera alrededor del valor observacional con un radio igual a la tolerancia dada como se muestra en la Fig. 3.3b.

La tolerancia en este punto se convierte en un parámetro libre del algoritmo genético. Esto se debe a que dependiendo de la tolerancia es el número de modelos de los cuales se podrán escoger los n mejores. Si la tolerancia es muy alta provocará que se tomen en cuenta muchos modelos para la selección. El problema más grande es que pueden ser modelos que estén relativamente lejos de la solución buscada.

Por otro lado, si la tolerancia es muy pequeña puede que no haya ningún modelo que cumpla con esta condición. Esto provocaría que si existe una solución que cumpla con la observación en el espacio O , está se perdería en el proceso. Esto tiene similitud con una cadena de Markov dado que la probabilidad de que ocurra un evento, en este caso es que esté dentro de nuestra tolerancia, depende solamente del evento inmediatamente anterior.

Algo que se debe tomar en cuenta es que la tolerancia no puede ser fija a través de las generaciones. El procedimiento genético ajusta los valores óptimos de una función entonces en cada generación los valores encontrados ajustarán de mejor manera con respecto a la generación anterior en el espacio O . Esto causa que la tolerancia para las primeras generaciones sea muy grande para las generaciones siguientes. La tolerancia debe de ser ajustada para cada generación. Cuando se hable de la reproducción de modelos se hablará de cómo se cambia la tolerancia con el paso de las generaciones.

La selección viene después de la determinación del ajuste. Se escogieron los n modelos más cercanos a la observación para ser los progenitores de los nuevos modelos. Este n se implementó para limitar el tamaño de la población siguiente y en este caso usamos $n = 6$. Este número de modelos seleccionados en un inicio se basó en seleccionar dos modelos por cada uno de los tres parámetros. No obstante, este número fue cambiado a lo largo de las generaciones para poblar con más individuos el espacio. Esto porque entre mayor número de progenitores, mayor será el número de crías producidos.

Esta selección de modelos se hace en el espacio O y posteriormente se identifican los parámetros correspondientes en el espacio P . Se seleccionan los seis conjuntos de parámetros y estos son los que se reproducirán. En este trabajo se utilizó un tipo de reproducción *asexual*. La reproducción asexual en un algoritmo genético se refiere a que los progenitores generarán a sus propias crías de manera independiente, o sea, que no habrá cruzamiento,

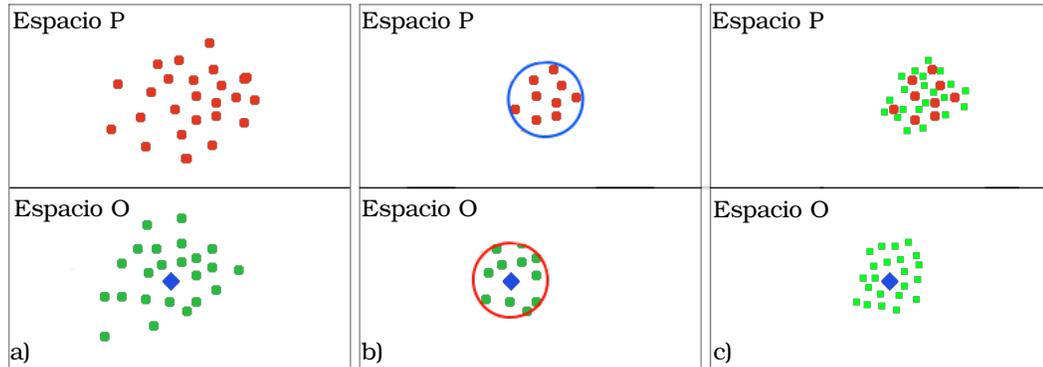


Figura 3.3: Algoritmo genético aplicado a los espacios P y espacio O. *Izquierda:* De una malla de puntos en el espacio P se corren modelos con Cloudy para generar los puntos del espacio O. *Medio:* De los datos en el espacio O se seleccionan los puntos dentro de un radio dado y a su vez se identifican los puntos correspondientes en el espacio P. *Derecha:* De los puntos seleccionados en el espacio se generan crías creando una nueva malla para repetir todo el proceso.

solo mutación. En este caso, los parámetros O/H, N/O y $\log U$ serán los mismos para progenitores y crías.

Como siguiente paso está la mutación de las crías. En la mutación se hace una variación en los parámetros de los modelos hijo (que son los mismos que los modelos padre). La variación se hace en los parámetros O/H, N/O y $\log U$. Este cambio es hecho a través de sumar o restar un número aleatorio a este valor. Con estos nuevos parámetros se generan nuevos modelos de fotoionización que bajo este contexto serían modelos *hijo*. De esta manera podemos generar el número de modelos crías que deseemos. Cada hijo tendrá valores diferentes en los tres parámetros con los cuales se generan nuevos modelos de fotoionización *crías* (ver Fig. 3.3c).

En este punto, la manera de cómo variar los parámetros también es otra propiedad ajustable del código genético. El rango de la variación de los parámetros debe ir disminuyendo con respecto a las generaciones. Cada vez que se hace un ciclo *generacional* los valores de los parámetros deben de ajustarse un poco mejor. En cada ciclo estaremos más cerca de la solución por lo que la variación debe ser menor en cada ciclo. Si tenemos un parámetro x , queremos variarlo sumándole un valor aleatorio (n_{rnd}) siguiendo una ley

normal. El cambio del parámetro sería

$$x' = x_0 + \sigma n_{\text{rnd}}. \quad (3.5)$$

Para aumentar o disminuir la variación de nuestro parámetro multiplicaremos n_{rnd} por un coeficiente σ . Este parámetro σ controlará el tamaño de la variación de cada uno de nuestros tres parámetros. Este coeficiente es una multiplicación de dos coeficientes tal que

$$\sigma = \sigma_N * \sigma_x. \quad (3.6)$$

Donde σ_N depende directamente del índice de generación en la que se esté realizando el algoritmo genético y sigue la expresión :

$$\sigma_N = \frac{\sigma_0}{2^{N/2}}. \quad (3.7)$$

Con N el número de generación actual. Se tiene que dar un valor inicial de σ_0 , el cual fue elegido en estos primeros modelos tal que $\sigma_0 = 0.5$. Además, se impone un valor de σ mínimo tal que $\sigma_{\text{min}} = 0.05$, esto significaría un cambio de 0.05 en el valor del parámetro. En la Fig. 3.4 se ve el comportamiento del parámetro σ_N . Para las primeras generaciones el valor de σ_N va disminuyendo. Esto provoca que el valor con el que se cambia el parámetro correspondiente sea mayor. Conforme el número de generaciones se incrementa el valor de sigma disminuye cada vez más, lo que provoca que el cambio sea cada vez menor. Para cuando el número de generaciones es muy grande ($N > 7$) el valor es el de σ_{min} .

El otro coeficiente σ_x depende directamente del parámetro (O/H, N/O o logU) que se esté variando. Dado que O/H, N/O y logU están en rangos diferentes de valores, se define σ_x para que cada parámetro varié de forma adecuada. En nuestros primeros modelos se tomó que $\sigma_{\text{O/H,N/O}} = 0.5$ y $\sigma_{\text{log U}} = -1.0$

Teniendo todo esto en cuenta generamos nuestra primera *familia* de modelos. Usamos la palabra familia bajo el contexto de método genético ya que engloban una serie de modelos desde su primera generación hasta la última generación.

§3.5.1. Primera familia de modelos

Nuestra primera familia de modelos son los primeros modelos aplicando el método genético. Como se mencionó arriba se necesita un malla inicial de puntos en el espacio de

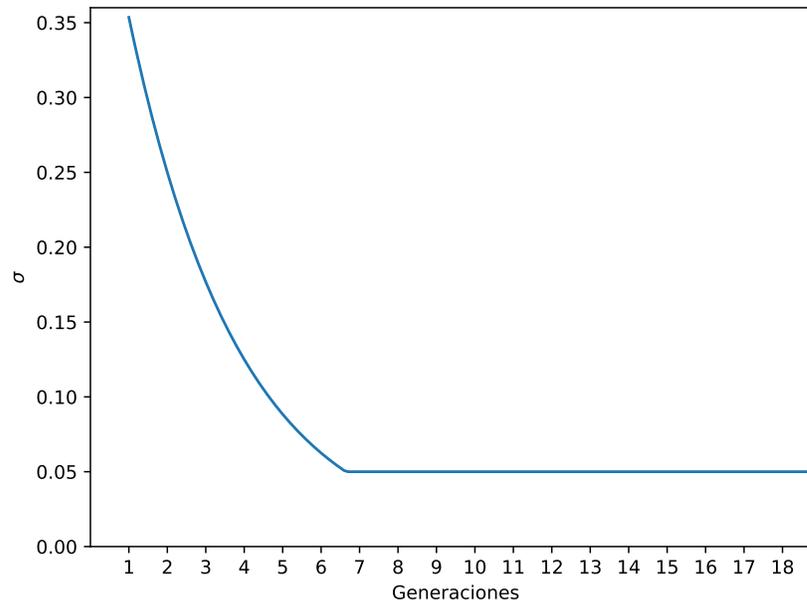


Figura 3.4: Comportamiento del parámetro σ . Iniciando desde un valor $\sigma_0 = 0.5$ decrece hasta la generación 7 en el cual toma un valor constante de 0.05. Por lo tanto nuestros parámetros tendrán un rango mayor en el cual poder cambiar sus parámetros en las primeras generaciones. A partir de la séptima generación este rango de variación se restringe a un valor pequeño y constante suponiendo que estamos cerca de converger a la solución.

parámetros a los cuales aplicarle el algoritmo genético.

El malla inicial puede ser regular o totalmente de forma aleatoria. En estos primeros modelos se optó por una malla de forma regular. En la Fig. 3.5 se muestra la malla inicial en el espacio P y su correspondiente en el espacio O.

Para esta y siguientes figuras en la gráfica $\log U$ vs O/H las líneas azules (sólida y rayada) son los ajustes obtenidos por Morisset et al. (2016). La línea negra punteada es el ajuste obtenido por Pérez-Montero (2014). La línea verde discontinua es el ajuste obtenido por Dopita & Evans (1986) y la línea morada de punto y guión es el ajuste hecho por Dopita et al. (2006). Por último, para la gráfica N/O vs O/H la línea cyan rayada es el ajuste realizado por Vila Costas & Edmunds (1993). La línea verde punteada es el ajuste hecho por Dopita et al. (2013) y la línea morada rayada es el ajuste hecho por Pilyugin et al. (2012). En ambas gráficas se muestra el ajuste para la abundancia de oxígeno hecha por Marino et al. (2013).

La malla es realizada en los tres parámetros O/H , N/O y $\log U$. Para N/O se tomaron seis valores entre -0.1 y -1.4 de $\log N/O$ con una separación aproximada de 0.25 dex. Para O/H se tomaron siete valores entre 7.6 y 9.3 de $12 + \log O/H$ con una separación de aproximadamente de 0.2 dex. Como se observa en la Fig. 3.5, alrededor de la línea de (Marino et al., 2013) se agregaron un par de valores para tener más puntos en esa área en donde podría estar una solución. Esto puede provocar que haya un sesgo en las soluciones que pueden obtenerse. Dado que probaremos el método genético por primera vez al menos esperamos encontrar la solución cerca de este calibrador.

En la Fig. 3.5 se observa que la malla es casi regular en el espacio O/H vs N/O pero en el espacio O/H vs $\log U$ no lo es. Esto se debe a que nuestros modelos contienen polvo y éste tiene efecto sobre el parámetro de ionización (Draine, 2011). Si fueran nebulosas sin polvo la malla sobre $\log U$ también sería regular. La cantidad del polvo relativamente al gas depende de la abundancia de oxígeno por lo que en una nebulosa con polvo necesitamos un parámetro de ionización mayor al de una nebulosa sin polvo. Todas las imágenes siguientes son de la misma región para ejemplificar el proceso con el método genético.

La malla inicial a la cual en esta familia de modelos le llamamos generación cero consta de 540 modelos que pueden ser seleccionados como posibles progenitores. Mientras que en el espacio P, la malla es regular, en el espacio O estos modelos cubren un área de forma no regular. Recordemos que los parámetros en el espacio P son inputs para modelos de fotoionización hechos con Cloudy. En el diagrama BPT (en la Fig.3.5 arriba a la izquierda) la malla inicial cubre toda la zona donde se encuentran las regiones HII.

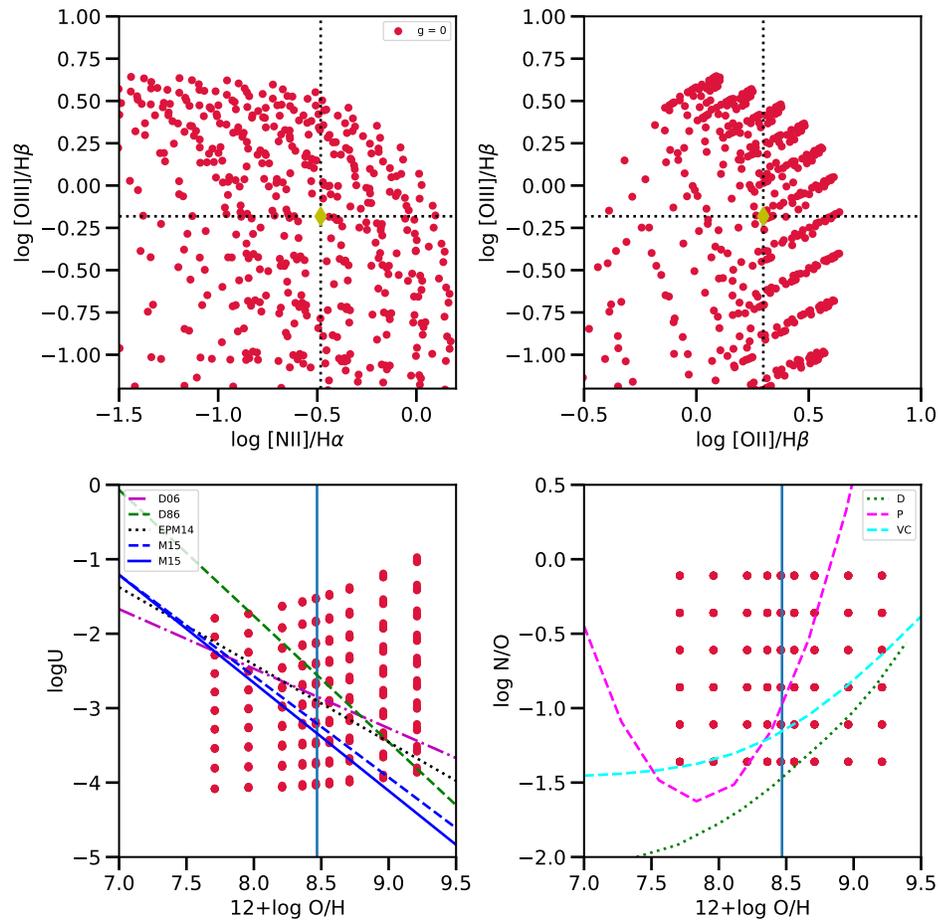


Figura 3.5: Malla inicial de la primera familia de modelos. En la parte superior el diamante representa la observación de CALIFA que se quiere reproducir. En las dos gráficas inferiores se tienen los puntos de la malla inicial. Se presentan algunos ajustes de la literatura: líneas azules (sólida y rayada) son ajustes obtenidos por Morisset et al. (2016). Línea negra punteada es el ajuste obtenido por Pérez-Montero (2014). La línea verde rayada es el ajuste obtenido por Dopita & Evans (1986) y la línea morada de punto y guión es el ajuste echo por Dopita et al. (2006). La línea cyan rayada es el ajuste realizado por Vila Costas & Edmunds (1993). La línea verde punteada es el ajuste hecho por Dopita et al. (2013) y la línea morada rayada es el ajuste hecho por Pilyugin et al. (2012). En ambas gráficas se muestra el ajuste para la abundancia de oxígeno (línea azul sólida) hecha por Marino et al. (2013). Este último ajuste se utilizó en el trabajo anterior por lo que se generaron un mayor número de puntos alrededor de esta línea.

Como se mencionó anteriormente, en cada generación podemos variar la tolerancia en cada aplicación de método genético. Entonces seleccionaremos los modelos que disten una cierta distancia de la observación. Esta distancia es la tolerancia que en esta primera generación fue igual a 1 dex. Una vez seleccionados los modelos dentro de esta tolerancia se escogieron $n = 6$ progenitores los cuales son los más cercanos a la observación. Cada progenitor generará 10 crías. Podemos ver en la Fig. 3.6 la generación 1 de la primera familia de modelos. Los puntos grises es la primera generación y los puntos rojos son las crías generadas por el algoritmo genético. Los asteriscos de color negro representan los seis puntos progenitores en ambos espacios.

En el espacio O vemos cómo los progenitores son efectivamente los puntos más cercanos a la observación. Para cada punto seleccionado como progenitor en el espacio O se identifican los parámetros correspondientes en el espacio P. Estos modelos seleccionados en el espacio P son reproducidos asexualmente y tienen una mutación como ya se mencionó anteriormente. Aunque algunos modelos parecieran ser más cercanos a la observación recordemos que es un espacio tridimensional. Los paneles de esta figura muestran proyecciones en espacios de dos dimensiones por lo que no se alcanza a apreciar la distancia real.

Repetimos este proceso haciendo cada vez mas pequeña la tolerancia para acercarnos a la observación con el paso de las generaciones. En la Fig. 3.7 observamos el resultado después de once generaciones. Al ver que algunos puntos en el espacio O están dentro de la tolerancia dada se incluyen los modelos de todas las generaciones pasadas. Esto para asegurar que si algún valor desde la malla inicial reproducía la observación no se pierda a lo largo del proceso genético.

Podemos ver claramente en el espacio O como tras cada generación los modelos reproducen mejor la observación. Los puntos de la generación final envuelven a la observación en contraste con generaciones tempranas, como la tercera o quinta.

Pero en el espacio P de la Fig. 3.7 notamos cierto comportamiento de los parámetros. La última generación los modelos reproducen la observación, pero sus parámetros correspondientes caen lejos de los ajustes hechos por diferentes autores. En la gráfica $\log U$ vs O/H observamos cómo la última generación está lejos de los ajustes aproximadamente por 0.5 dex. En la gráfica $\log N/O$ vs O/H observamos que la última generación está totalmente apartada de la zona donde se encuentran todos los ajustes de Dopita et al. (2013) y de Pilyugin et al. (2012). Por otra parte, la malla inicial tenía una mayor densidad de puntos alrededor del calibrador de Marino et al. (2013) por lo que podríamos esperar que las

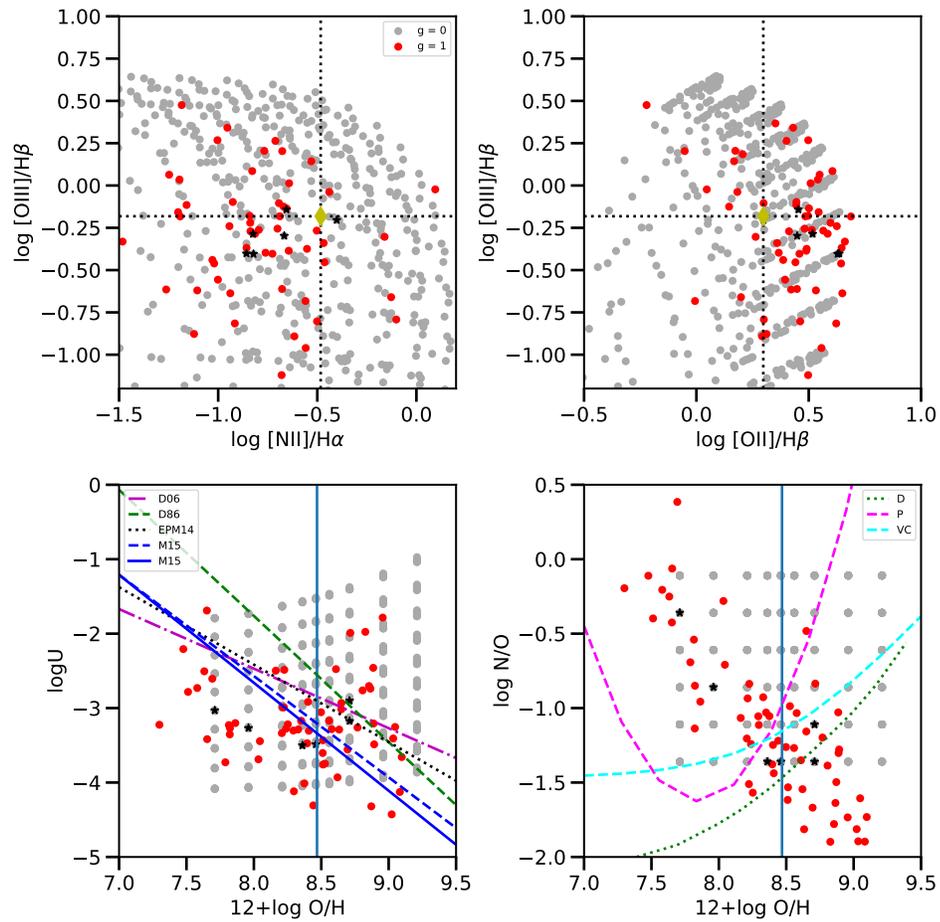


Figura 3.6: Generación 1 de la primera familia de modelos. En puntos grises se representa los puntos de la generación anterior, es decir, el grid inicial de la familia. En puntos negros se muestran los modelos escogidos para ser progenitores. Los puntos rojos representan las crías generadas a partir de los modelos progenitores. Las líneas son las mismas mostradas en la Fig. 3.5.

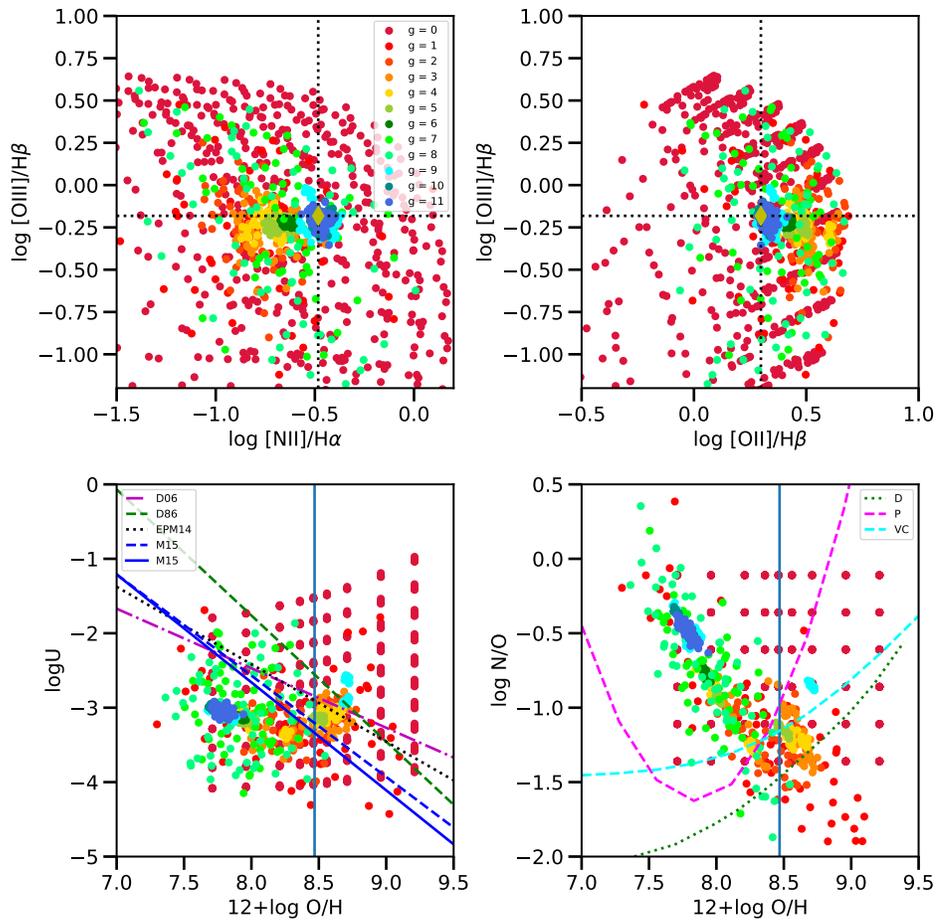


Figura 3.7: Todas las generaciones de la primera familia de modelos de la región 2. Observamos como las primeras generaciones, representadas por tonos rojizos, abarcan una gran zona del espacio de parámetros. Conforme se realizan generaciones observamos como los modelos en el espacio O empiezan a converger hacia la observación (modelos con tonos azules). Pero en el espacio P, sus parámetros correspondientes, igualmente empiezan a converger a una pequeña zona. Esto define un conjunto solución lo cual en este caso solo es un conjunto de los dos esperados. Además el único conjunto solución encontrado se encuentra lejos de los ajustes y resultados de la literatura. Las líneas mostradas son las mismas que en la Fig. 3.5.

soluciones obtenidas fueran cerca de este calibrador. Pero en muchos de los casos se obtiene una solución que difiere mucho del valor predicho del calibrar. Por lo que la distribución inicial de la malla no sesgó de alguna manera las soluciones encontradas.

El algoritmo genético está encontrando una solución que reproduce la observación. Pero por lo mencionado anteriormente al final del cap. I esperamos encontrar al menos dos conjuntos de soluciones en el espacio P que reproduzcan la misma observación en el espacio O. Dado que reproducen la misma observación no podemos discernir entre la que es astrofísicamente correcta de la que no.

En la Fig. 3.8 mostramos una representación de los espacios O y P en una misma gráfica. En el eje horizontal están representados nuestros parámetros y en el eje vertical los datos observacionales disponibles en una sola dimensión cada una. La línea punteada representa los valores actualmente observados. La línea rayada representa un valor de tolerancia dado. Finalmente la línea sólida representa el perfil de la función que estamos tratando de optimizar con el método genético. Es decir, es una representación del perfil de densidad de nuestro espacio P. No sabemos la forma que tiene esta función pero sabemos que hay al menos dos soluciones que cumplen o se acercan al valor observado.

En la Fig. 3.8 los dos *pozos* representan cómo ciertos parámetros reproducen las observaciones. Donde el mínimo de los pozos es la solución buscada. Entonces podemos ver que dependiendo del valor de la tolerancia dada tendremos un conjunto de soluciones diferentes.

Cuando se hace la malla inicial se debe mostrar la línea sólida, es decir, la función que tratamos de optimizar. Cuando se hace el método genético los puntos que estén por debajo de la línea de tolerancia son candidatos a producir una nueva generación.

Los dos pozos que se muestran en la Fig. 3.8 no tienen por qué ser de la misma forma. Uno puede ser más ancho, estrecho, profundo, asimétrico, etc. En la figura ponemos como ejemplo el caso en que un conjunto de soluciones es más ancha que la otra. Dado que se puebla aleatoriamente el espacio es más probable que el pozo ancho tenga más puntos que cumplan con la tolerancia dada. Esto generará que conforme la tolerancia vaya disminuyendo, cada vez sea menos probable que el pozo estrecho tenga puntos con los cuales generar crías .

Es por esto que el algoritmo genético después de cierto número de generaciones puede perder una solución. Para lidiar con este problema necesitaríamos que cada pozo tuviera su propia generación de modelos independientes. Es decir, dividir el espacio P en subespacios donde cada uno de estos pudiera tener sus progenitores y crías independientes.

De la manera que está implementado el algoritmo genético los posibles progenitores

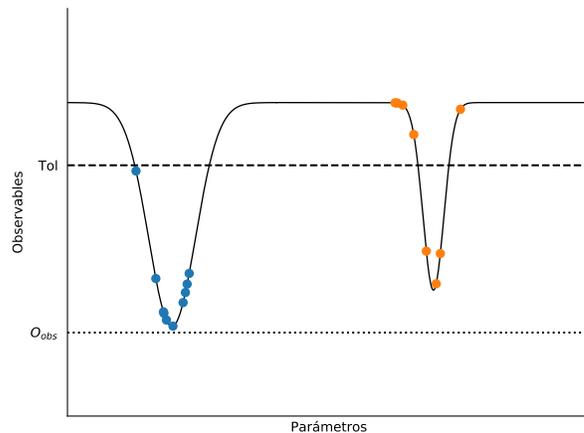


Figura 3.8: Representación de la relación entre el espacio de parámetros y el espacio de observables. Los espacios O y P se representan en una dimensión. Si la tolerancia con la cual se escogen modelos en el método genético es muy pequeña podemos dejar fuera una solución. Esto es representado con los pozos, si la tolerancia pasa uno de los dos pozos entonces se perdería una solución.

son seleccionados en el espacio O. No hay ninguna manera de diferenciar a los progenitores que estén en una solución o en la otra. Esto ocasiona que se combinen ambos conjuntos de soluciones en el espacio P. Al dividir el espacio P en dos o más subespacios podemos clasificar los modelos de acuerdo al subespacio al que pertenezcan, como se muestra en la Fig. 3.9.

Bajo un contexto de método genético a estos subespacios los llamamos *dinastías*. De esta manera, en el espacio P se le asignará su dinastía a cada modelo. Al hacer la selección de progenitores en el espacio O se hará de manera independiente para cada dinastía. Esto significa que por cada dinastía se tendrán progenitores diferentes. Esto causará en el espacio P que en cada dinastía se generen parámetros para la siguiente generación de manera independiente. Con esto esperamos poblar zonas diferentes en el espacio P para que ninguna de las dos soluciones se pierda con el paso de las generaciones.

Pero hacer esta separación automáticamente por medio de una computadora es complicado. El ser humano es capaz de reconocer patrones a simple vista pero para una computadora no es algo trivial. Dado que esperamos que haya dos conjuntos de soluciones entonces en el espacio P existirán dos agrupaciones o cúmulos de puntos. Por lo tanto, podemos utilizar algún algoritmo adecuado de análisis de cúmulos.

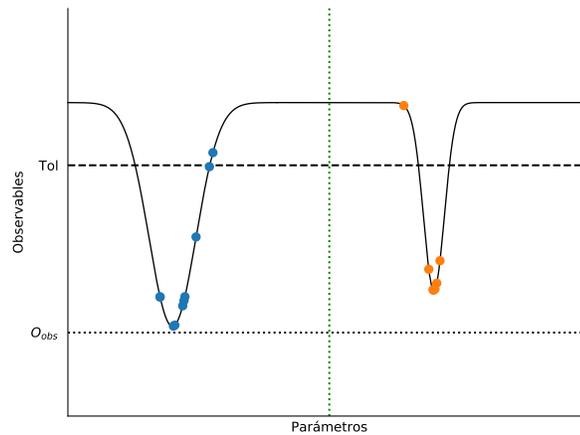


Figura 3.9: Relación entre el espacio de parámetros y el espacio de observables dividido en dos subespacios. Esperamos obtener dos conjuntos de soluciones por lo explicado al final del capítulo 1. Si dividimos el espacio y aplicamos el método genético en cada subespacio entonces obtendríamos los dos conjuntos solución. Esto debido a que la evolución del método genético de cada solución sería independiente una de la otra.

§3.6. Análisis de cúmulos: *Mean Shift*

Introducción

Los algoritmos de análisis de cúmulos se encargan de agrupar un conjunto de datos. Un cúmulo es un conjunto de objetos o datos en este caso en el cual cada objeto está cerca o es muy similar a todos los otros objetos en el cúmulo y muy diferente a los objetos que no están en el cúmulo. Esta agrupación se hace basada solamente en la información que se encuentra en los datos y sus relaciones (Tan et al., 2005). El objetivo es agrupar los datos con características similares y diferenciar diferentes grupos de datos.

Decir cuándo una agrupación de datos es un cúmulo no es nada trivial. En la Fig. 3.10 vemos un ejemplo de diferentes agrupaciones que pueden hacerse. En la Fig. 3.10(a) vemos un conjunto de puntos que para un humano es claramente visible que existen dos agrupaciones de puntos como se observa en Fig. 3.10(b). Pero podemos agruparlos aún más creando subcúmulos de los dos principales. En la Fig. 3.10(c) y Fig. 3.10(d) podemos ver cómo los cúmulos principales se dividen en cuatro y hasta en seis subcúmulos. Por lo que delimitar definir la forma y el número de elementos de un cúmulo de un conjunto de datos no es sencillo. También dependerá del trabajo y el tipo de *clasificación* que se requiera.

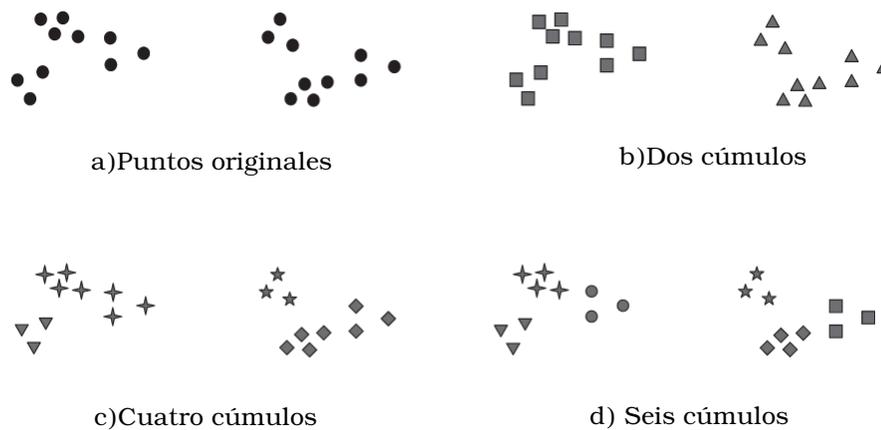


Figura 3.10: Agrupación de grupos de datos en cúmulos. Podemos ver que los puntos originales claramente para un ser humano están agrupados en dos cúmulos como se muestra en el caso b). Pero si lo requiriéramos podríamos formar subcúmulos de estos dos cúmulos. Esto se muestra en el caso c) y d) en el cual se dividen en cuatro y seis cúmulos, correspondientemente. (Tan et al., 2005).

El análisis de cúmulos es llamado a veces como *clasificación sin supervisión*. Esto se debe a que el análisis divide los datos en grupos, crea etiquetas dependiendo de estos grupos y asigna a los datos alguna de las etiquetas creada. En contraste con los métodos catalogados dentro de *clasificación supervisada* donde los datos son etiquetados usando un modelo derivado de los objetos con etiquetas conocidas. Los métodos de *clasificación sin supervisión* derivan las etiquetas a partir de los datos.

Tipos de análisis de cúmulos

Consideremos que hay dos tipos de algoritmos en el análisis de cúmulos: 1. *clustering* particional, 2. *clustering* jerárquico (Kaufman & Rousseeuw, 2009). De aquí en adelante llamaremos *clustering* al proceso de identificar cúmulos realizado por el análisis de cúmulos.

El primer tipo de métodos es el llamado *clustering particional*. Este consiste en una simple división en k cúmulos de los datos. Estos cúmulos deben de tener al menos un elemento y cada elemento debe pertenecer a sólo un cúmulo. Esto causa que los cúmulos dados por este tipo de algoritmos no se superpongan. También que el número total de cúmulos sea menor o igual al número de elementos a clasificar. Ejemplo de este tipo de

clustering esta dado en la Fig. 3.10

En este tipo de algoritmos el número de cúmulos k es dado por el usuario. Por lo que el algoritmo particiona a los datos de manera que se tenga el número de cúmulos deseado. El número de cúmulos k puede ser seleccionado de manera automática probando posibles valores de k y seleccionando aquel que sea mejor con respecto a algún criterio.

El otro tipo es *clustering jerárquico* el cual en principio permite que los cúmulos tengan subcúmulos. Hay dos tipos de técnicas jerárquicas: aglomerativa y divisiva. Los métodos aglomerativos empiezan con un número de cúmulos igual al número de datos. En cada paso dos cúmulos son unidos hasta que solo queda un cúmulo. Por otra parte, los métodos divisivos consisten en tomar un solo cúmulo al inicio y en cada paso siguiente divide el cúmulo hasta tener un número igual al número de datos.

Pareciera que en las dos técnicas obtienes un solo cúmulo o un número grande de cúmulos. Pero depende de la muestra de datos es como se clasificaran los cúmulos. Por ejemplo, para un método aglomerativo puede especificarse que sólo se fusionen aquellos cúmulos que tengan una distancia entre sí menor a una distancia límite dada. Esto causará que al final no todos los cúmulos se combinen. Por otro lado, para un método divisivo de manera análoga podemos especificar que si un cúmulo al separarse da origen a dos cúmulos con una distancia entre ellos es muy grande entonces no se divida. De esta manera se tiene más control sobre el *clustering*.

Además de *clusterings* particionales y jerárquicos podemos clasificarlos en exclusivos, no exclusivos y *fuzzy*. Los *clusterings* exclusivos son aquellos que a cada dato le asignan un único cúmulo. Pero en un caso general, un punto podría pertenecer a dos o más cúmulos al mismo tiempo. Los *clusterings* que asignan a cada dato más de un cúmulo son llamados *clusterings* no exclusivos.

Un *clustering fuzzy* a diferencia de los tipos de *clustering* anteriores no asigna a cada dato un cúmulo como tal. Lo que calcula es una probabilidad o peso de que un punto pertenezca a un cúmulo o a otro. Esto puede ser útil cuando hay una ambigüedad en la decisión si un dato pertenece a uno o a otro cúmulo. En la práctica un *clustering fuzzy* es transformado a un *clustering* exclusivo al asignar a un cúmulo solo aquellos datos que cumplan con una alta probabilidad de pertenencia.

Por último, podemos clasificar los *clusterings* por completos o parciales. Los *clustering* completos asignarán todos los puntos a un cúmulo. En los *clusterings* parciales podemos omitir algunos puntos y no asignarlos a ningún cúmulo. Esto puede causar que los cúmulos resultantes estén mejor definidos ya que podemos considerar que algunos puntos pueden

ser ruido. Sin embargo, los *clusterings* completos pueden ser útiles si queremos que todos los puntos de la muestra sean clasificados.

Diferentes tipos de cúmulos

Además de los diferentes tipos de *clustering* que existen también hay diferentes tipos de cúmulos. Esto es importante ya que dependiendo del tipo de cúmulos que tengamos escogeremos el *clustering* más apropiado.

Supongamos que tenemos una muestra de datos u objetos que dan como resultado dos cúmulos como en la Fig. 3.10a). Los cúmulos están bien definidos de manera que la distancia entre dos puntos en diferentes cúmulos es más grande que la distancia entre dos puntos dentro del mismo cúmulo.

Los cúmulos pueden ser construidos con base en un prototipo, es decir, que los puntos que constituyen un cúmulo son los que se encuentran más cercanos al prototipo de su cúmulo que al de cualquier otro cúmulo. Un prototipo común es el centroide de los puntos que en este caso es el punto más representativo de este cúmulo.

Otro tipo de cúmulos es el definido basado en la densidad de puntos. Este tipo de cúmulos se definen en un área donde hay una alta densidad de objetos y son delimitados por un área de baja densidad. Es conveniente definir un cúmulo por densidad cuando estos son irregulares o cuando se tienen cúmulos entrelazados.

Existe otro tipo de cúmulos como aquellos donde los datos son representados por una gráfica o aquellos donde los datos comparten alguna propiedad creando cúmulos conceptuales (Kaufman & Rousseeuw, 2009).

Eligiendo un algoritmo de *clustering*

Para escoger un método de *clustering* tenemos que pensar en qué tipo de cúmulos pueden formar nuestra muestra de datos. Por otra parte, tenemos que elegir si queremos dar un número de cúmulos o es mejor que el algoritmo determine el número de cúmulos.

Recordemos que al espacio al que aplicaremos el *clustering* es nuestro espacio P el cual es un espacio tridimensional. Si retomamos la Fig. 3.7 y la Fig. 3.9 nos interesa identificar las dos soluciones. Pero queremos dividir el espacio P en dos zonas para que el algoritmo genético pueda ser independiente en cada zona.

No podemos asegurar que las dos soluciones de nuestros datos formen cúmulos bien definidos. En la Fig. 3.11 podemos ver distintos casos que podemos obtener. El de la parte

superior izquierda es el mejor caso posible ya que estamos lejos del punto de máximo de la relación entre [OIII] y O/H. Esto causará, como se ve en la parte superior derecha, que los dos conjuntos de soluciones estén bien separados y definidos. En este caso nuestros cúmulos estarían bien definidos.

En la Fig. 3.11 parte media izquierda se muestra qué pasa si estamos cerca del punto máximo. En la parte media derecha se muestra cómo en el espacio de parámetros empiezan a sobreponerse las soluciones. Por lo tanto, los cúmulos que pueden formar ya no son bien definidos. Por lo que nuestro método de *clustering* debe ser capaz de lidiar con cúmulos sobrepuestos.

Como último caso tenemos el de la Fig. 3.11 parte baja izquierda cuando estamos exactamente en el punto máximo. En el espacio P sólo tendremos un conjunto de datos. En este caso sólo se debería formar un cúmulo, por lo tanto, el método de *clustering* que escojamos debe de determinar automáticamente el número de cúmulos.

El primer algoritmo de *clustering* que probamos fue *K-means* (Hartigan, 1975). Está basado en calcular distancias de los puntos a un centro de prueba. Una vez que los puntos están asignados a un centro se calcula un centroide y éste será el nuevo centro prueba. Este proceso se realiza de manera iterativa hasta la convergencia. A pesar de ser un buen algoritmo y fácil de implementar necesita un número inicial de cúmulos para funcionar. Por lo que se mencionó anteriormente no es conveniente usarlo.

Otros algoritmos que se probaron fueron:

- *Gaussian Mixture*: este método es parte de los algoritmos de *Mixture models*. A diferencia de los tipos de *clustering* descritos arriba (particionales y jerárquicos), este es un método que utiliza funciones de densidad de probabilidad para determinar los elementos de los cúmulos. En este caso utiliza funciones gaussianas para ajustar los datos (Reynolds, 2015). Pero este método se le deben de proporcionar el número de componentes, es decir, el número de cúmulos por lo que se descartó.
- *DBSCAN*: El algoritmo *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) se basa en encontrar cúmulos por densidades (Ester et al., 1996). Este método se basa en muestras de núcleos, es decir, un cúmulo es un conjunto de núcleos que están cerca unos con los otros. Este método tiene la ventaja de que los cúmulos pueden tener una forma arbitraria.
- *Mean Shift*: Mean shift es un algoritmo basado en centroides que en cada iteración

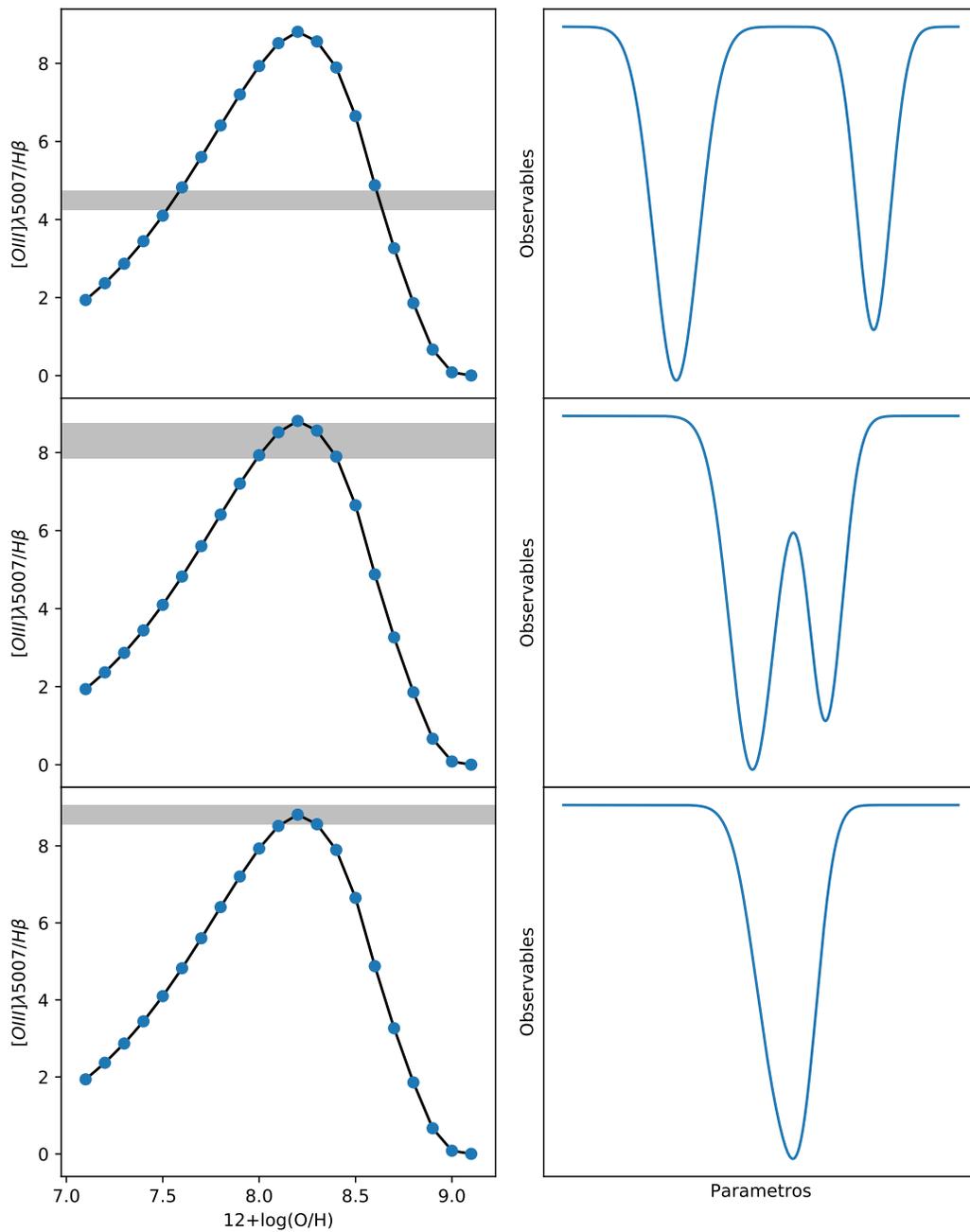


Figura 3.11: Diferentes casos en la función bi valuada visto en el espacio P vs espacio O. En el panel superior observamos que cuando la solución está lejos del punto máximo de la relación entre O/H y [OIII] tenemos dos soluciones bien definidas en el espacio de parámetros. En el panel medio se muestra que si estamos cerca del punto máximo en el espacio de parámetros las dos soluciones pueden superponerse una con la otra. El panel inferior muestra el caso en que estamos en el punto máximo donde solo tendremos un conjunto de soluciones.

actualiza los candidatos a centroides por el promedio de puntos dentro de una región dada. Si hay centroides que sean muy cercanos, casi duplicados, se eliminan y se obtienen el conjunto de centroides finales.

Como se puede observar en la Fig. 3.12 se usó la paquetería de Python Scikit-learn (Pedregosa et al., 2011) para diferentes casos junto con sus tiempos de ejecución. En la figura se muestran diversos tipos de cúmulos demostrando el poder de *clustering* de cada uno de los algoritmos.

A pesar de que podemos ver que en ciertos casos algunos algoritmos realizan una mejor identificación de cúmulos debemos de pensar que es lo más conveniente para nuestro espacio P. Debido a lo discutido anteriormente necesitamos un algoritmo que identifique el número de cúmulos por él mismo. También esperamos en algunos casos cúmulos bien definidos por lo que debe de poder identificarlos a la perfección. Es conveniente que cuando se tenga una muestra homogénea de datos como la malla inicial el algoritmo sólo identifique un cúmulo. Por último, nos interesa más que divida en dos o más partes el espacio de parámetros.

El *clustering* que se adecua más a nuestro problema es el algoritmo *Mean Shift*. El algoritmo DBSCAN a pesar de que muestra un mejor *clustering* es un método parcial mientras que *Mean shift* es un método completo. Esto quiere decir que DBSCAN puede no clasificar todo el conjunto de datos mientras que *Mean Shift* sí clasifica a todos los puntos dentro de un cúmulo. Para nuestro caso nos conviene un método que sea completo. Por otra parte, el método DBSCAN depende de dos parámetros (número mínimo de elementos de un núcleo y una distancia mínima) mientras que el Mean Shift solo dependerá de un parámetro que se explicará en la siguiente sección.

En la Fig. 3.12 también se muestran los tiempos de ejecución. Notamos como Mean Shift es el que más tarda de los tres métodos probados. Sin embargo, es el método que mejor características ofrece para nuestro problema.

Mean Shift

El algoritmo *Mean Shift* es una poderosa herramienta. Es un proceso iterativo no paramétrico que puede ser usado para un gran número de tareas. Fue introducido por Fukunaga (1975) para reconocimiento de patrones.

Mean Shift considera al espacio como una función de densidad de probabilidad empírica. Si se cuenta con un conjunto de datos, el método los considera como una muestra discreta de la función de densidad de probabilidad relacionada. Si en este espacio hay regiones densas,

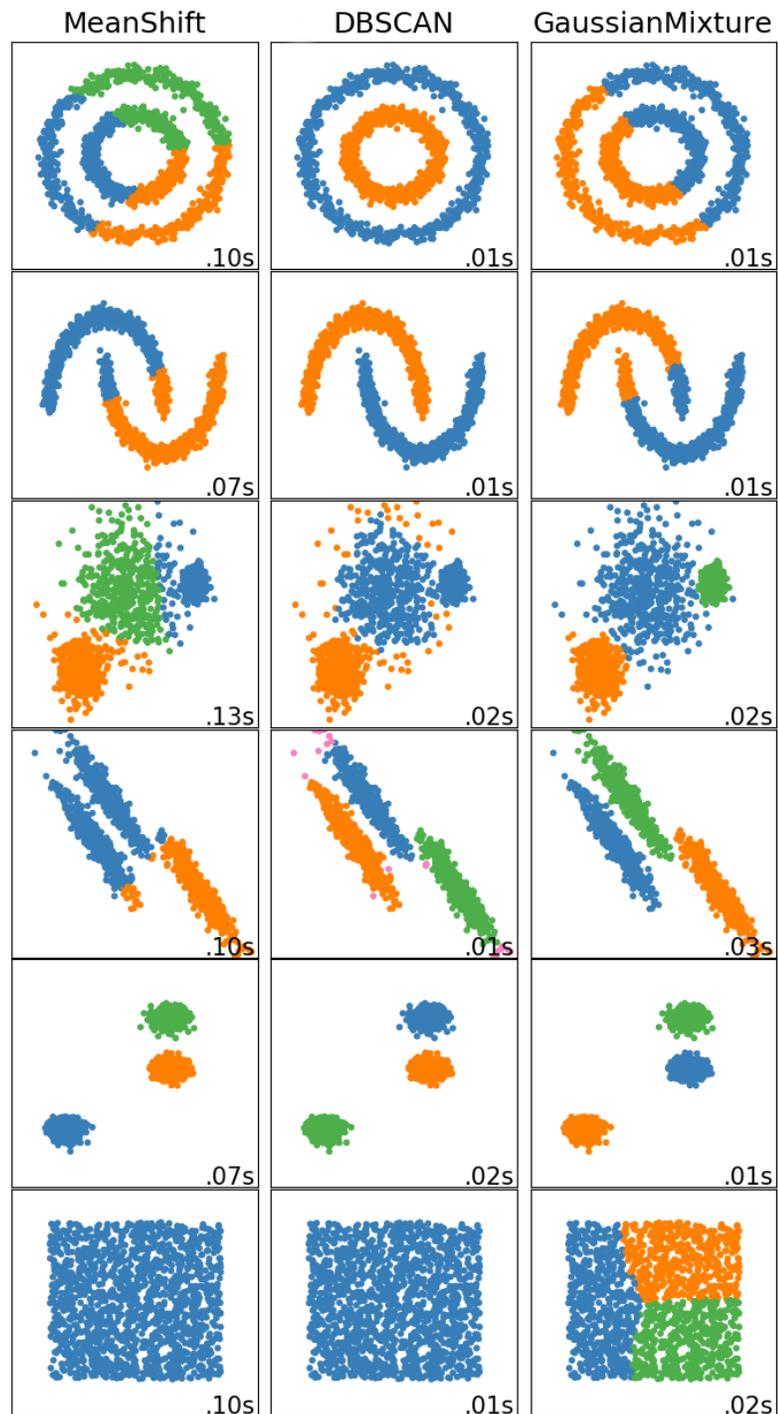


Figura 3.12: Comparación de diferentes métodos de *clusterings* (Pedregosa et al., 2011). A pesar de que solo se representan los tres métodos considerados para nuestro proceso existen un gran número de algoritmos de identificación de cúmulos. Notemos que en cada uno de los paneles se muestra el tiempo de ejecución. Este tiempo varía de acuerdo a la cantidad de puntos como de la distribución de datos. En general vemos que el algoritmo *Mean Shift* es el que tarda más en identificar los cúmulos.

entonces estas corresponden a un máximo local de la función de densidad de probabilidad.

Mean Shift crea una función de densidad de probabilidad para el espacio donde se encuentran los datos. Después asocia cada punto de los datos con el pico más cercano de esta función. Posteriormente define una ventana alrededor del punto y calcula el promedio de los datos en ella. Entonces mueve el centro de la ventana al valor promedio y repite el algoritmo hasta converger. Después de cada iteración, se puede considerar que la ventana se mueve a una región más densa del conjunto de datos.

Básicamente *Mean Shift* tiene los siguientes pasos:

- Fija una ventana alrededor de cada punto de los datos
- Calcula el promedio de los datos dentro de la ventana
- Mueve la ventana a el promedio y lo repite hasta converger.

Para ilustrar, esto veamos la Fig. 3.13. En una muestra de datos, donde podemos observar claramente tres zonas densas de puntos. El círculo verde está centrado en un dato inicial. La línea roja es cómo se mueve el centro del círculo en cada iteración moviéndose hacia una zona más densa de puntos. El último punto de su recorrido es en el centro del cúmulo. Esto sucede con todos los demás puntos.

Mean Shift utiliza una estimación de la densidad del *kernel* (KDE por sus siglas en ingles, Kernel Density Estimation). Antes de definir KDE, definiremos que es un *kernel*. Un *kernel* es un tipo especial de función de densidad de probabilidad que debe de cumplir las siguientes propiedades (Epanechnikov, 1969):

- No negativa y valuada en los reales

$$0 \leq \phi(x) < \infty; \quad (3.8)$$

- Par

$$\phi(-x) = \phi(x); \quad (3.9)$$

- Estar normalizada

$$\int_{-\infty}^{\infty} \phi(x) dx = 1; \quad (3.10)$$

Algunas funciones de densidad de probabilidad comunes son kernels:

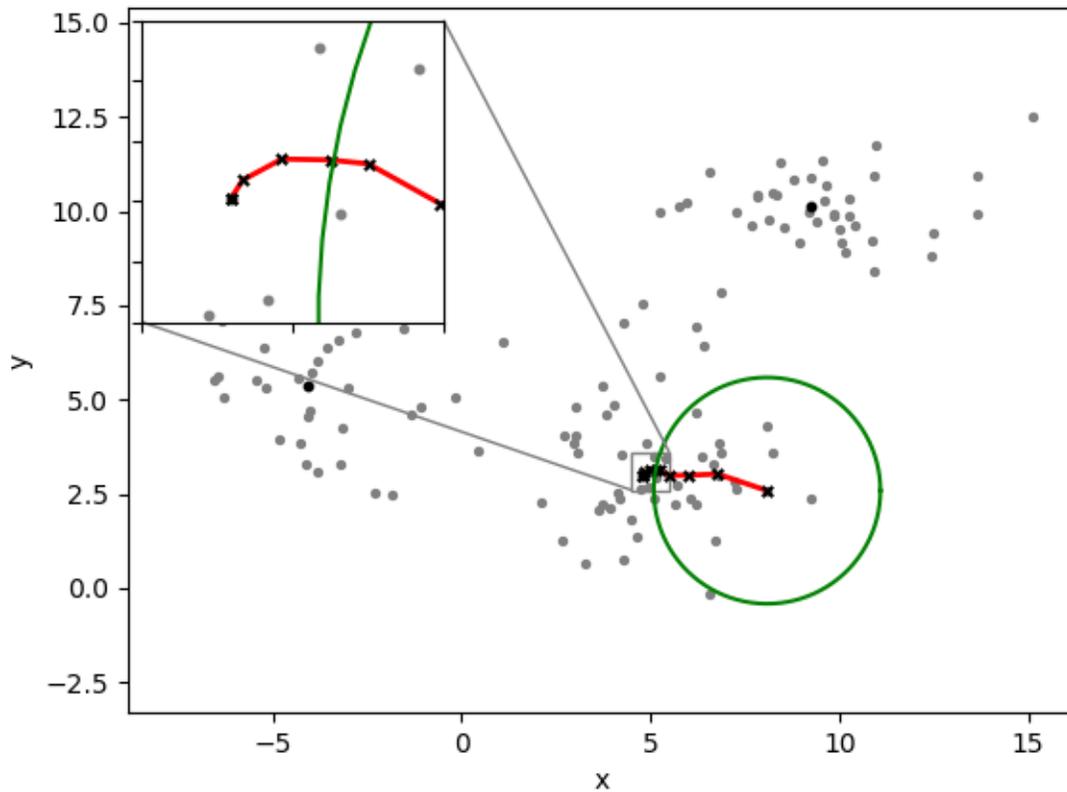


Figura 3.13: Proceso llevado a cabo por *Mean Shift*. Para cada punto de los datos define una ventana, representada por el círculo verde. Dentro de esta ventana se calculará donde se encuentra la mayor densidad de puntos. Se moverá la ventana, y por lo tanto el punto relacionado a ella hacia donde se encontrará la mayor densidad de puntos. Este movimiento se representa por la línea roja. Este proceso se repite hasta que dentro de la ventana la densidad de puntos es constante. En este punto lo definirá como el centro del cúmulo. Todos los puntos que lleguen al mismo centro o estén muy cercanos a este centro se clasificarán como miembros del cúmulo definido.

- Rectangular

$$\phi(x) = \begin{cases} 1 & a \leq x \leq b, \\ 0 & \text{Otro caso;} \end{cases} \quad (3.11)$$

- Gaussiana

$$\phi(x) = e^{-\frac{x^2}{2\sigma^2}}. \quad (3.12)$$

KDE es un método no paramétrico para estimar la función de densidad de probabilidad de una variable aleatoria continua. No es paramétrico porque no asume ninguna distribución relacionada para la variable.

Esencialmente, para cada dato, una función *kernel* es creada con el dato en su centro, esto asegura que el *kernel* sea simétrico con respecto al dato. La PDF (Probability Density Function) es estimada al sumar todas las funciones *kernel* y dividiendo por el número de total datos, n , (Comaniciu, 2002):

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (3.13)$$

Esto es usualmente llamado la técnica de ventanas de Parzen (Richard O. Duda, 2001). La constante d es la dimensión de nuestros datos y \mathbf{x} es el dato sobre el cual estamos calculando el KDE. El parámetro h es llamado ancho de banda, es decir, el ancho de la ventana. Uno de los pasos con mayor dificultad es la elección del ancho de banda que capture la distribución asociada a la variable.

Cuando se tiene un valor pequeño h se tendrá una desviación estándar pequeña por lo que el *kernel* colocará gran parte de la probabilidad en el dato. Este caso se usa cuando el tamaño de la muestra es grande y los datos están altamente aglomerados. Por el contrario, un valor grande de h resultará en una gran desviación estándar, el *kernel* esparcirá la probabilidad del dato a sus vecinos. Es recomendable usar estos valores de h cuando el tamaño de la muestra es pequeño y la muestra está dispersa.

Una vez que se tiene la ventana, el KDE, se procede a moverla. Para esto, se utiliza un algoritmo de optimización de gradiente ascendente. Este tipo de algoritmos es una manera de maximizar una función objetivo, es decir, se buscan los valores de los parámetros de una función los cuales la maximizan. Se puede considerar que *Mean Shift* se basa en un ascenso de gradiente sobre los contornos de densidad.

Para encontrar máximos, tenemos que encontrar los puntos donde $\nabla f(\mathbf{x}) = \bar{0}$. Si la

aplicamos a nuestro KDE, ec.3.13:

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^n K' \left(\frac{x - x_i}{h} \right). \quad (3.14)$$

Si suponemos que la derivada de k existe para todo x , definimos

$$g(x) = -K'(x). \quad (3.15)$$

Sustituyendo en la ec. 3.14

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}) g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right). \quad (3.16)$$

Dividimos y multiplicamos por $\sum_{i=1}^n g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)$ obtenemos:

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \left[\sum_{i=1}^n g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right) \right] \left[\frac{\sum_{i=1}^n \mathbf{x}_i g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)}{\sum_{i=1}^n g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)} - \mathbf{x} \right]. \quad (3.17)$$

El primer término es igual o proporcional (Comaniciu, 2002) al estimado de densidad en \mathbf{x} con el *kernel* . El segundo término es llamado *mean shift*:

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{x}_i g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)}{\sum_{i=1}^n g \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)} - \mathbf{x}. \quad (3.18)$$

Este término es la diferencia entre la media pesada por el *kernel* g , y \mathbf{x} , que es el centro de la ventana.

El proceso *mean shift* puede ser resumido como: Para cada punto \mathbf{x}_i :

1. Calcular el vector *mean shift* $m(x_i)$
2. Mover la ventana de estimación de densidad una distancia $m(x_i)$
3. Repetir hasta la convergencia

En la Fig. 3.14 observamos la misma muestra de datos que en la Fig. 3.13. En un inicio, cada punto es tomado como el centro de su propia ventana. Al calcular el vector *mean shift* se mueve el centro de acuerdo a $m(x_i)$. Esto se repite hasta que todos los puntos alcanzan un punto en común. Este punto será propio de cada cúmulo. Como se puede apreciar en la Fig. 3.14, se llega a tres centros, es decir, existen tres cúmulos.

Esto presenta una gran ventaja sobre algunos métodos de análisis de cúmulos, *mean shift* no necesita un número de cúmulos para iniciar el análisis. *Mean shift* calcula de manera automática el número de cúmulos existentes.

Uno de los problemas de *Mean shift* es que no funciona bien para altas dimensiones. En altas dimensiones el número de máximos locales son muy altos por lo que puede dar resultados no correctos.

Aplicando Mean Shift a nuestros modelos

El algoritmo *Mean Shift* fue implementado usando la librería de Python Scikit-learn (Pedregosa et al., 2011). El método Mean Shift implementado en esta paquetería puede depender de un solo parámetro llamado *quantile*.

Como se habló en la sección anterior el ancho de banda es la dimensión de la ventana que usará Mean Shift. Al no dar un valor de ancho de banda el algoritmo implementado por *Scikit-learn* lo calcula automáticamente dado un parámetro. Este parámetro es *quantile* el cual en estadística y la teoría de probabilidad son puntos que dividen el rango de una distribución de probabilidad en intervalos contiguos con probabilidades iguales. También pueden dividir un conjunto de observaciones en muestras de la misma manera (Hyndman & Fan, 1996). En nuestro caso determina donde se *corta* la muestra que se pasará al *kernel* de Mean Shift (Drabas, 2016).

El valor de *quantile* tiene un efecto directo en el número de cúmulos resultantes pudiendo ir de 0 a 1 donde con este último valor clasificará a los datos en un solo cúmulo. Para un valor de *quantile* igual a 0.5 corresponde a la mediana de la distribución. En este punto en el que se empezaría a explorar el método de mean shift se decidió que el valor de *quantile* se ajustaría manualmente cada vez que se hiciera el *clustering*. Ajustamos el valor de *quantile* con el objetivo de dividir en dos nuestro espacio de parámetros como se muestra en la fig. 3.15.

Por lo tanto, una vez generada la malla inicial de modelos se aplicaría el *clustering*. El valor de *quantile* sería tal que hubiera dos o más cúmulos bien definidos. En este punto no

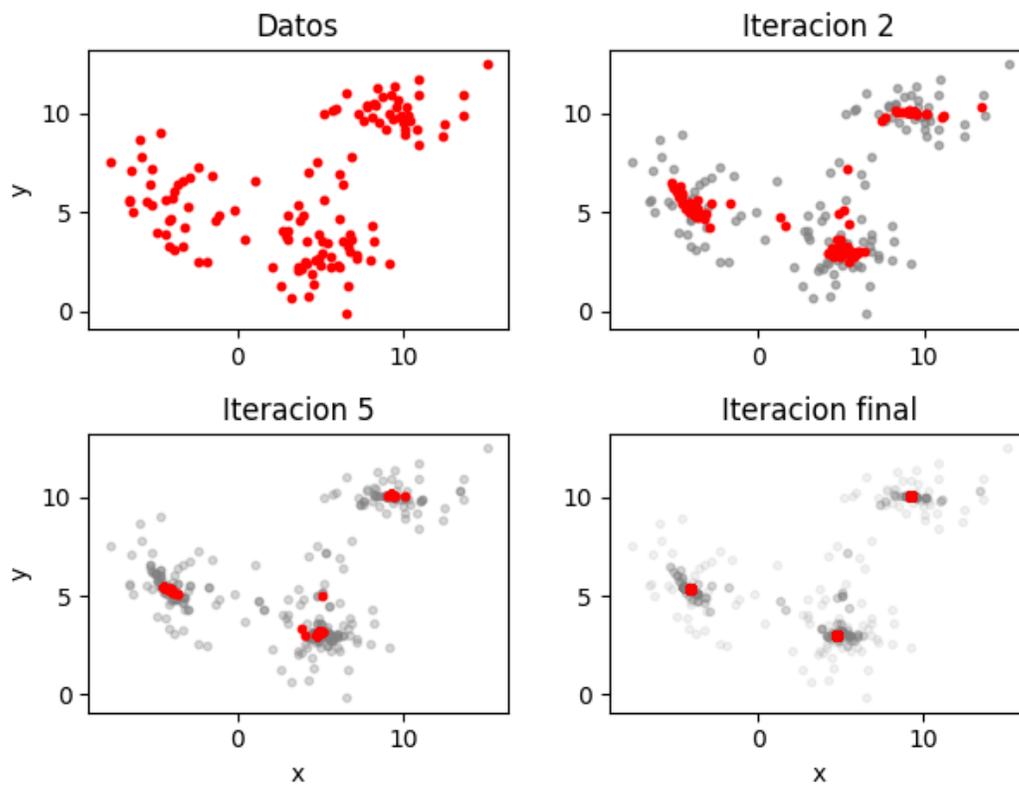


Figura 3.14: Encontrando centro de los cúmulos con el algoritmo *Mean Shift*. Observamos la distribución de datos en el panel superior izquierdo. En cada paso del algoritmo los puntos se mueven hacia el centro de su respectivo cúmulo. En la iteración final vemos como todos los puntos caen en el centro de su respectivo cúmulo. Todos los puntos que hayan convergido al mismo centro se clasificarán en el mismo cúmulo.

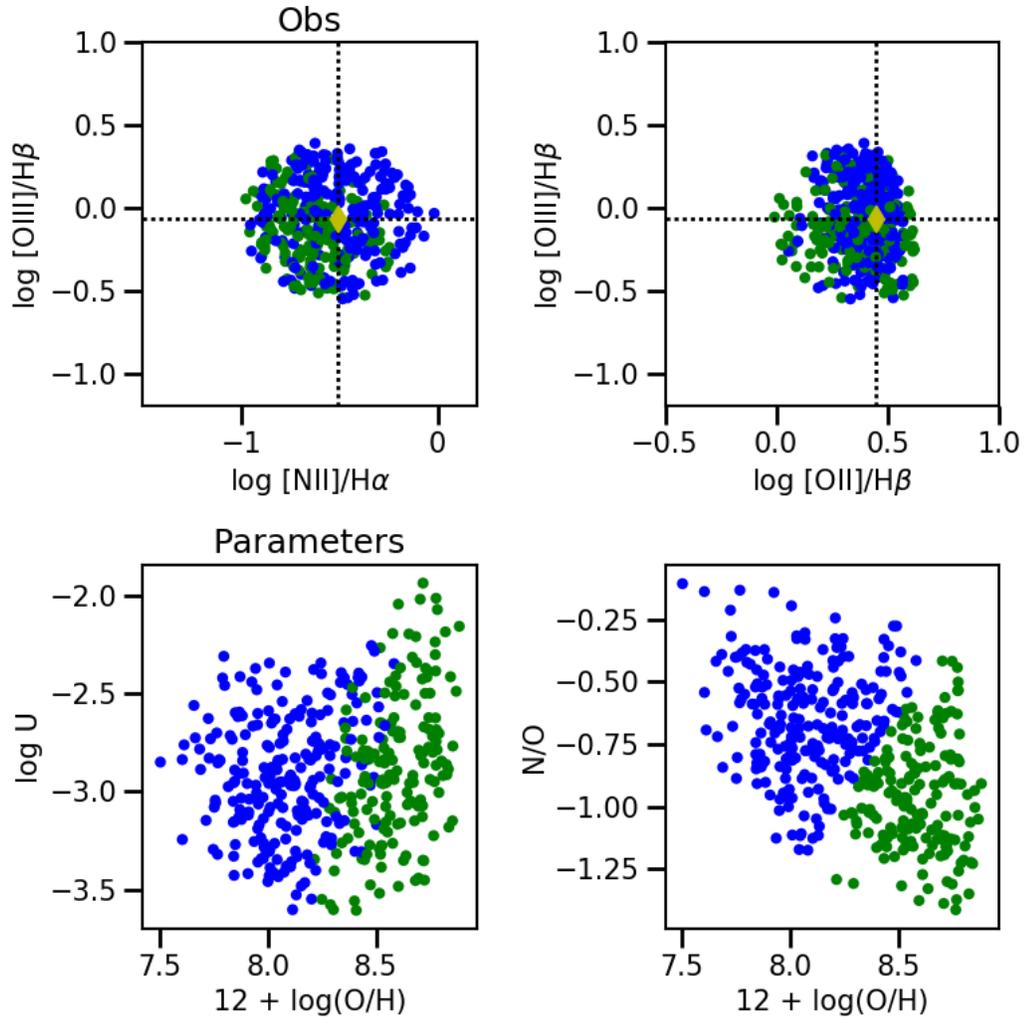


Figura 3.15: Algoritmo *Mean Shift* aplicado a nuestros modelos de la región 2. Observamos que todos los modelos reproducen la observación con una diferencia de 0.5 dex en el espacio O. En el espacio P el *clustering* nos divide el espacio en dos cúmulos. A pesar de que en el panel inferior izquierdo se sobreponen los cúmulos recordemos que el espacio P es un espacio en tres dimensiones y una proyección en dos dimensiones hace que encimen los cúmulos identificados. Con esta división podemos aplicar el método genético en cada cúmulo de manera independiente.

Rama Genealógicas	Tolerancia	Quantile	No de crías
A	0.8	0.4	40
B	0.8	0.2	10
C	0.5	0.4	20
D	0.8	0.3	20

Tabla 3.2: Valores utilizados en las distintas dinastías en el primer paso genético.

se tomó un valor de *quantile* diferente para todas las regiones. Se trató de que el valor de *quantile* obtuviera resultados óptimos de todas regiones en general.

Después del *clustering* se procedería en cada dinastía con un método genético como en la primera familia de modelos. Este proceso se repetiría hasta que los modelos generados por nuestros parámetros difirieran de las observaciones por una tolerancia igual o menor a 0.05.

Segunda familia de modelos

Para la segunda familia de modelos se implementó el algoritmo de *clustering*. En esta familia a diferencia de la primera familia al malla inicial se le nombró generación 1. También la malla inicial fue generado de manera totalmente aleatorio (Fig. 3.16) a diferencia de la malla inicial de la primera familia que fue regular (Fig. 3.5). De esta manera, la malla inicial no generará sesgo alguno. Finalmente, el número de modelos en esta familia fue reducida a 100 puntos en el espacio P.

La construcción de esta segunda familia siguió el siguiente proceso: Se escogen los modelos que ajusten mejor las observaciones. A estos modelos se aplica el *clustering*. Una vez identificados los diferentes cúmulos se aplica el método genético. Por último, se realizan los modelos Cloudy con base en los parámetros generados.

Para ver el efecto del parámetro *quantile* se realizaron ramas genealógicas. A cada rama se le asignaron diferentes valores como se puede ver en la tabla 3.2. Se realizaron cuatro ramas genealógicas con parámetros diferentes. En todas las ramas se conservó el mismo número de progenitores ($n = 6$).

Se realizó una comparación de los resultados de cada rama genealógica. Se eligió la rama en la que consideramos que mejor se realizaba el *clustering*. Esta elección se basó en la calidad de selección e identificación de cúmulos. Esta selección se hizo de forma cualitativa, es decir, en cada iteración se buscaba el valor óptimo de *quantile* comparando

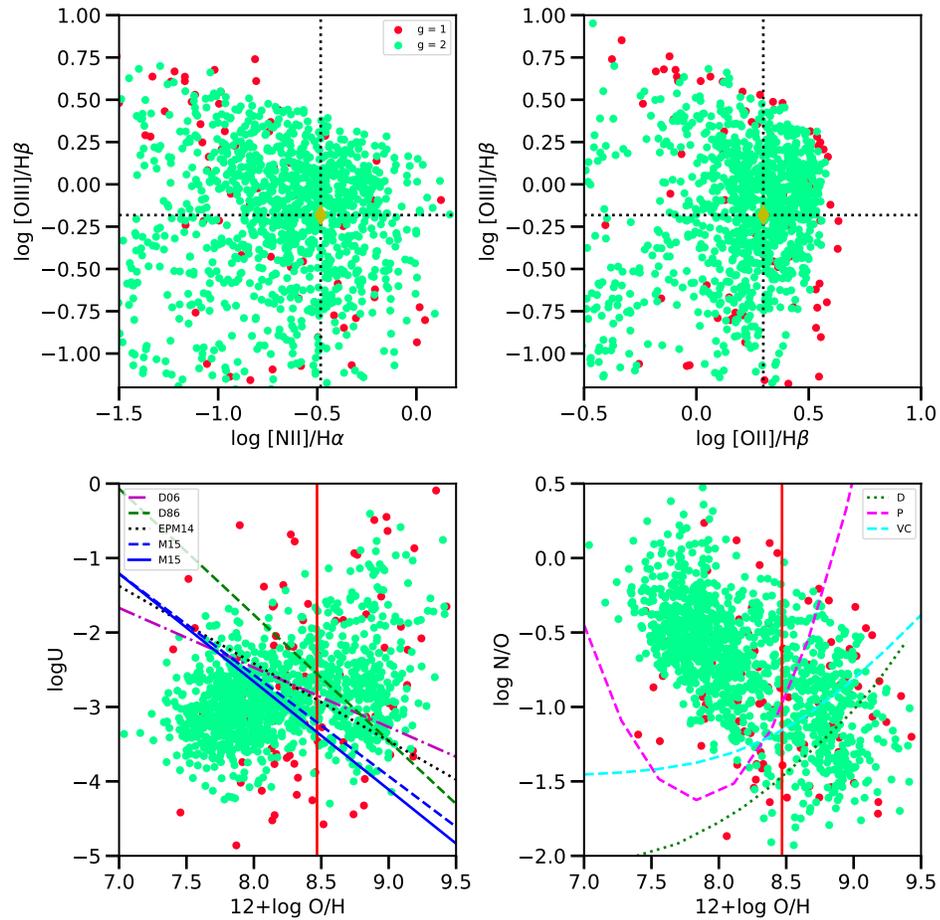


Figura 3.16: Generación 1 y 2 de la segunda familia de modelos para la región 2. La generación 1 (puntos rojos) es el grid inicial generado de forma aleatoria. La segunda generación (puntos turquesa) son generados a partir de los modelos de la primera generación que mejor ajustan las observaciones.

a ojo los cúmulos resultantes para diferentes valores del parámetro en el espacio P (ver Fig. 3.15).

La producción de la tercera se realizó de la misma manera. Se generaron diversas ramas genealógicas variando los valores de la tolerancia y *quantile* mientras el número de crías se mantuvo constante $n_{crias} = 20$.

A partir de la cuarta generación se realizó sólo una rama genealógica. De esta manera se hizo hasta una octava generación probando distintos valores de *quantile* y disminuyendo la tolerancia. Con esto se estudio qué valores de *quantile* eran los más adecuados para generaciones avanzadas donde los conjuntos de soluciones son más definidos. Se observó que si el número de datos es pequeño y el valor de *quantile* es menor a 0.20 ocasiona que el *clustering* identifique cúmulos muy pequeños. Esto causa que a pesar de que nosotros visualmente identifiquemos dos cúmulos principales el *clustering* divide estos cúmulos principales en subcúmulos.

En la Fig. 3.17 y Fig. 3.18 vemos el resultado final para la región 2 y región 25 respectivamente. Podemos ver que en la región 2 se tienen los dos conjuntos de soluciones pero en la región 25 sólo se tiene un conjunto de soluciones. En otras regiones igualmente se pierde un conjunto de soluciones y el que permanece es aquel que está en la zona que no es astrofísicamente posible.

Teniendo una idea de los valores adecuados para el parámetro *quantile* finalizamos la segunda familia que consistió aproximadamente de 262000 modelos de fotoionización. La primera familia consistió en cerca de 97000 modelos de fotoionización. Son una gran cantidad de modelos que revisar para asegurarnos de que el *clustering* se realiza de manera correcta. Es por esto que ideamos una manera de poder automatizar el proceso hasta ahora explicado. En la tercera familia se trató de automatizar el proceso de *clustering* de tal manera que se buscara de manera automática el valor óptimo de *quantile*.

Tercera familia de modelos

La búsqueda automática del parámetro *quantile* es la principal característica de la tercera familia. Se definió una lista de posibles valores de *quantile* para encontrar el valor óptimo para el parámetro *quantile* dado un número deseado de cúmulos. El proceso fue el siguiente:

1. Se toma un valor de la lista de valores de *quantile*
2. Realización del *clustering*

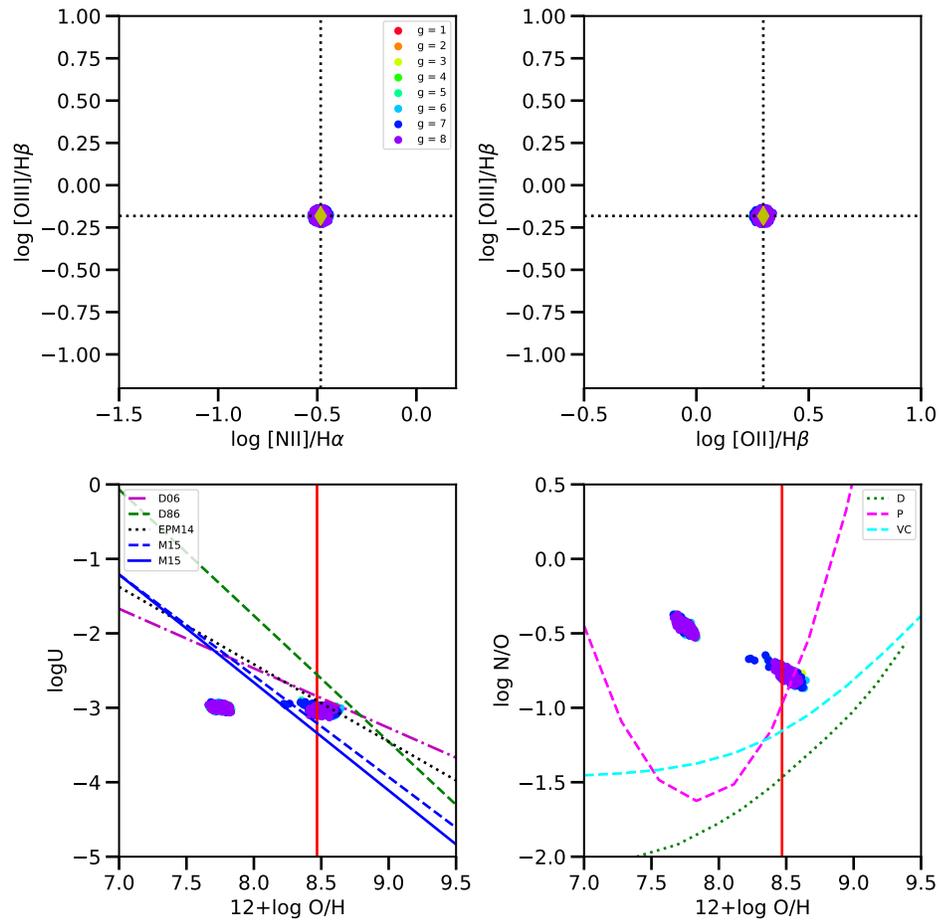


Figura 3.17: Resultado final de la segunda familia con una tolerancia de 0.05 para la región 2. Observamos como en esta región obtenemos los dos conjuntos de soluciones esperadas. Las líneas son las mismas mostradas en la Fig. 3.5.

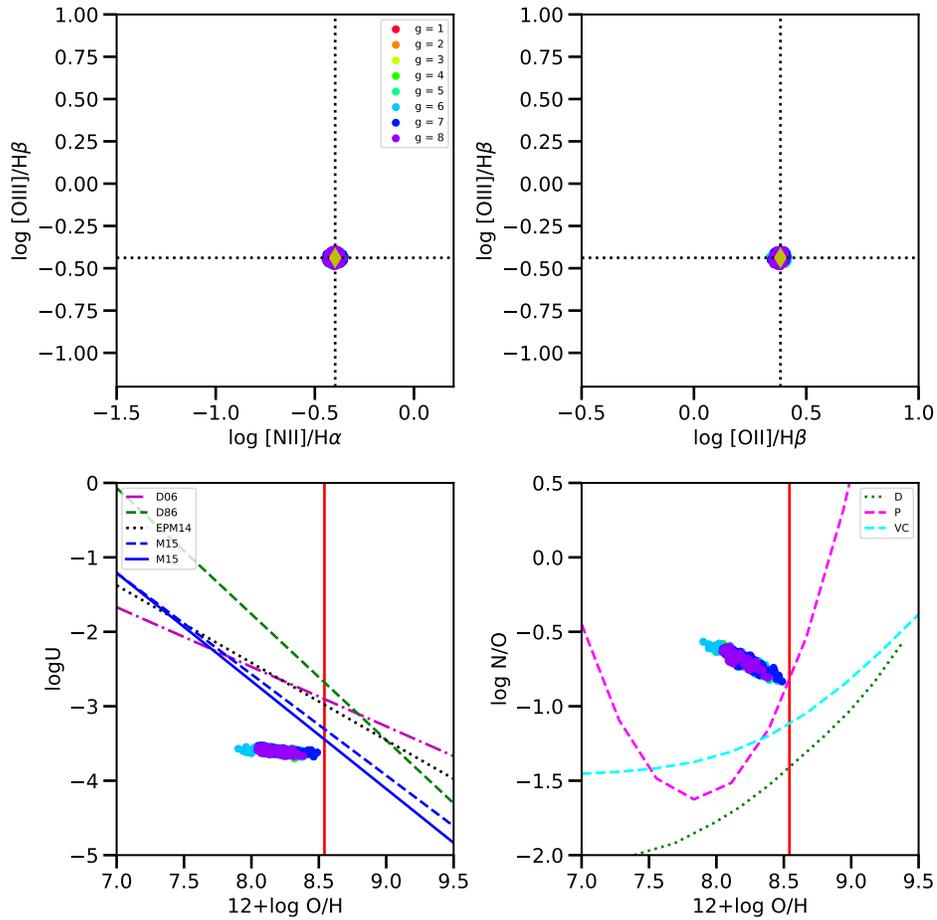


Figura 3.18: Resultado final de la segunda familia con una tolerancia de 0.05 para la región 25. A diferencia de la región 2, en esta región solo se obtiene un conjunto solución. Esto se puede deber a que aún perdemos una solución o estemos en el punto máximo de relación entre O/H vs $[\text{OIII}]\lambda 5007$. Todos los modelos de estos parámetros reproducen la observación con una discrepancia máxima del 0.05 dex. Las líneas son las mismas mostradas en la Fig. 3.5.

3. Comparación del número de cúmulos obtenidos y el número de cúmulos deseados
4. Repetición del proceso con el siguiente valor de la lista de valores de *quantile* hasta que se tenga el número de cúmulos deseados.

La lista de valores de *quantile* va desde 1.0 hasta 0.05 con una separación de 0.1. Para valores menores de 0.2 se hace una malla de valores más finos con una separación de 0.02.

El flujo del nuestro algoritmo es el mismo que para la segunda familia. La malla inicial es tomado de la familia anterior, es decir, la primera y segunda generación de la segunda familia son usadas para generar la tercera generación de la tercera familia de modelos.

A partir de estas generaciones se realizaron hasta la séptima generación siempre buscando el valor de *quantile* que generara dos cúmulos de datos. En la Fig. 3.19 vemos el resultado después de siete generaciones. Se tomó una tolerancia igual a 0.2 la cual es cuatro veces más grande que la tolerancia ocupada para las Fig. 3.17 o Fig. 3.18. Por lo tanto, nuestro proceso automatizado de *clustering* no funciona correctamente.

En la fig 3.20 mostramos el proceso para nuestro proceso automático de obtención de *quantile*. Empezamos con valores de *quantile* alto de tal forma que tenemos la Fig. 3.20a. Para estos valores sólo se obtiene un solo cúmulo. Si seguimos disminuyendo el valor de *quantile* obtendremos en algún momento dos cúmulos. Sin embargo, estos dos cúmulos tendrán un tamaño desigual y, por lo tanto, una densidad de puntos diferentes, lo que provoca que un cúmulo sea mucho más grande que el otro como se ejemplifica en la Fig. 3.20b. Este caso es lo que está ocurriendo con nuestro proceso automatizado ya que se detiene al encontrar cualquier valor de *quantile* que de como resultado dos cúmulos o el número de cúmulos que se desee. El valor encontrado de esta manera no es precisamente el que da el mejor *clustering*.

En la segunda familia se tuvo un mejor resultado dado que la calidad del *clustering* se realizó manualmente. En cada generación se escogieron los parámetros que tuvieran dos cúmulos con un tamaño equiparables como en la Fig. 3.20c. Por lo tanto, nuestro proceso automático de *clustering* y el proceso de obtención del valor de *quantile* debe de incluir una manera de cuantificar la calidad del *clustering*.

Calidad del *clustering*: Coeficiente de Calinski Harabaz y coeficiente de Silhouette

Existen diferentes maneras de cuantificar qué tan bueno es un algoritmo de análisis de cúmulos. En un sentido general los métodos de *clustering* tienen diversos usos y depen-

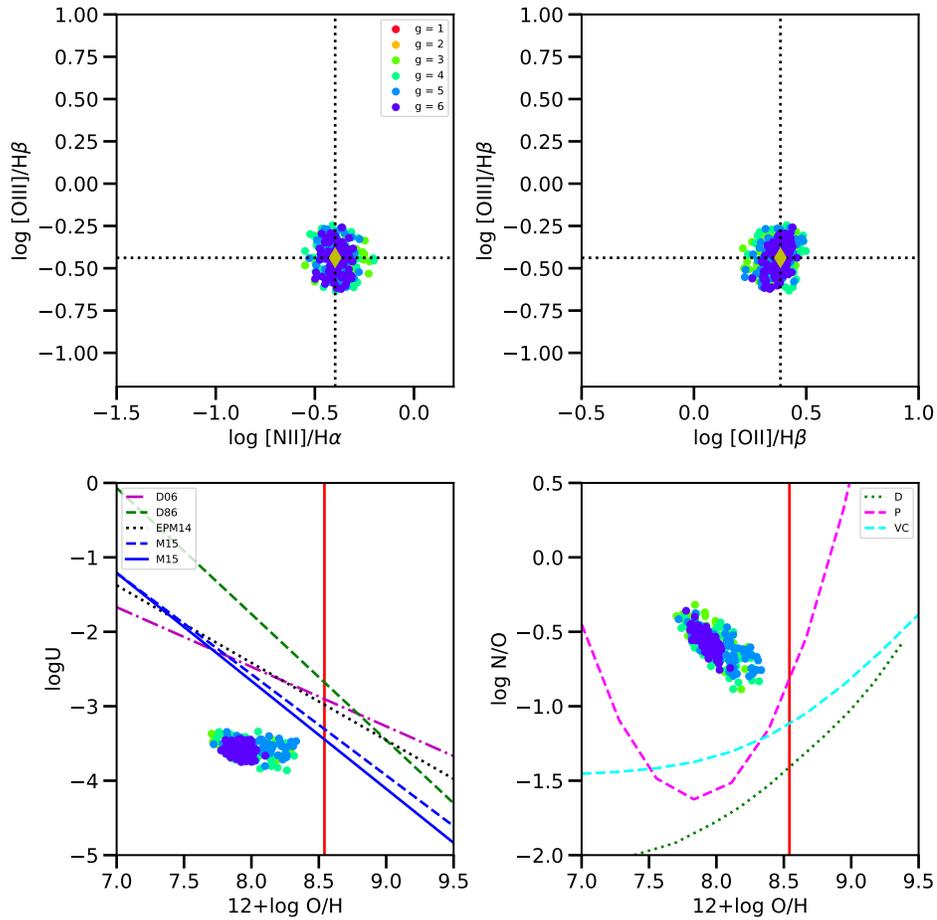


Figura 3.19: Resultado después de siete generaciones de la tercera familia con una tolerancia de 0.2 para la región 25. A diferencia de la familia anterior donde los parámetros se ajustaban en cada generación a ojo. En esta generación que se trató de automatizar el proceso el conjunto solución que se obtiene está alejado de los resultados de la literatura. Esto se debe posiblemente a la automatización de la búsqueda del parámetro *quantile* para el método *Mean Shift*. Las líneas son las mismas mostradas en la Fig. 3.5.

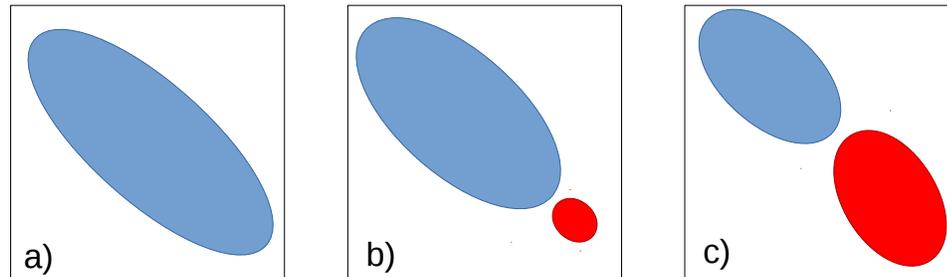


Figura 3.20: Posibles cúmulos obtenidos para diferentes valores del parámetro *quantile*. Dependiendo del valor que se utilice podemos obtener diferentes tipos de cúmulos. Para un valor igual a $quantile = 1$ es el caso a) tomará todos los elementos y solo identificará un cúmulo. Para valores muy cercanos a 0 para el parámetro *quantile* el método puede identificar cúmulos muy pequeños. El caso ideal para algún valor de *quantile* que queremos obtener es el del caso c) donde los dos cúmulos son de tamaño comparable.

diendo de la aplicación puede requerirse diferentes tipos de cúmulos. Para nuestro caso necesitamos dividir nuestro espacio de parámetros en al menos dos regiones. Por lo tanto, los cúmulos que se obtengan por medio del *clustering* deben de ser comparables como los mostrados en la Fig. 3.20).

Existen algunos índices o coeficientes para cuantificar la calidad del *clustering*. En cierto sentido los métodos de *clustering* asignan etiquetas a los datos. Estas etiquetas corresponden a los cúmulos identificados por el algoritmo utilizado. Muchos de estos índices para cuantificar la calidad de *clustering* suponen que se tienen las etiquetas correctas. Por lo que algunos se basan en hacer una comparación de las etiquetas encontradas por el *clustering* y las etiquetas reales.

Algunos de estos son:

- Índice *Adjusted Rand*
- Coeficiente Fowlkes-Mallows
- Coeficiente *V-measure*

En nuestro estudio no sabemos las etiquetas reales por lo que necesitamos una manera de medición independiente de las etiquetas reales. Estos coeficientes que cumplen con esto son: Coeficiente de Calinski Harabaz y coeficiente de Silhouette.

Coefficiente de Calinski Harabaz

El índice Calinski Harabaz (Caliński & Harabasz, 1974; de Amorim & Hennig, 2015) se usa cuando no se conocen las etiquetas del modelo. Este coeficiente nos mide qué tan bien definidos son los cúmulos. Su valor será mucho mayor cuando estos sean densos y bien separados. Además es rápido de calcular .

Tengamos k cúmulos. Entonces el coeficiente Calinski Harabaz (s) está dado por el cociente de la dispersión promedio entre los cúmulos y la dispersión interna de los cúmulos:

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N - k}{k - 1}. \quad (3.19)$$

Donde Tr significa la traza, B_k es la matriz de dispersión entre grupos:

$$B_k = \sum_q n_q (c_q - c) (c_q - c)^T. \quad (3.20)$$

Y W_k es la matriz de dispersión dentro del cúmulo:

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q) (x - c_q)^T. \quad (3.21)$$

Con N el número de datos, C_q el conjunto de puntos en el cúmulo q , c_q el centro del cúmulo q , c el centro de todos los datos y n_q es el número de puntos en el cúmulo q .

La desventaja de este coeficiente es que es generalmente más alto en cúmulos convexos que en otro tipos de cúmulos definidos de diferente manera. Por ejemplo, los cúmulos obtenidos por el método DBSCAN.

Coefficiente de Silhouette

Al igual que el coeficiente de Calinski Harabaz, el coeficiente de Silhouette evalúa el *clustering* con el modelo mismo. Un valor alto del coeficiente de Silhouette relaciona a un modelo de *clustering* con cúmulos bien definidos. El coeficiente de Silhouette se basa en dos medidas: a) la distancia media entre un dato i y todos los demás puntos en el mismo cúmulo y b) la distancia media entre la dato i y todos los demás puntos del cúmulo más cercano (de Amorim & Hennig, 2015; Rousseeuw, 1987).

El coeficiente de Silhouette s está dado por:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \quad (3.22)$$

Donde $a(i)$ y $b(i)$ son las dos distancias medias antes descritas. Esta expresión es aplicada a cada dato i .

Este coeficiente toma valores desde -1 a 1 donde -1 indica que nuestro dato i debe de estar asociado a otro cúmulo vecino. Un valor cercano a 1 indica que nuestro dato i está apropiadamente asociado a su cúmulo. En dado caso que este coeficiente sea igual a 0 nos indica que hay cúmulos superpuestos, es decir, que un dato i se encuentra en la frontera de dos cúmulos. Si calculamos el promedio de todos los coeficientes de Silhouette de todos los datos obtendremos una cuantificación del *clustering* realizado.

El máximo valor que puede alcanzar es 1 por lo que un coeficiente cercano o igual a este valor indica que los cúmulos están bien separados y son densos. En la Fig. 3.21 mostramos un ejemplo para una muestra genérica de datos. En esta figura vemos que dependiendo de los cúmulos encontrados y la distribución de la muestra el coeficiente de Silhouette, promedio y de cada dato, cambiará.

Al igual que el coeficiente de Calinski Harabaz, el coeficiente de Silhouette es generalmente más alto en cúmulos convexos que en otro tipos de cúmulos definidos de diferente manera. Además en algunas ocasiones toma mucho tiempo el poder calcular este coeficiente a comparación del coeficiente de Calinski Harabaz. Para un conjunto de 150 puntos el cálculo del coeficiente de Silhouette toma alrededor de $3.62\text{ms} \pm 254\mu\text{s}$ mientras que el coeficiente de Calinski Harabaz tarda $1.1\text{ms} \pm 48.9\mu\text{s}$.

Con estos dos coeficientes podemos medir la calidad del *clustering* y los usaremos en la cuarta familia de nuestros modelos. Ambos coeficientes fueron utilizados usando el paquete *Scikit-learn* (Pedregosa et al., 2011).

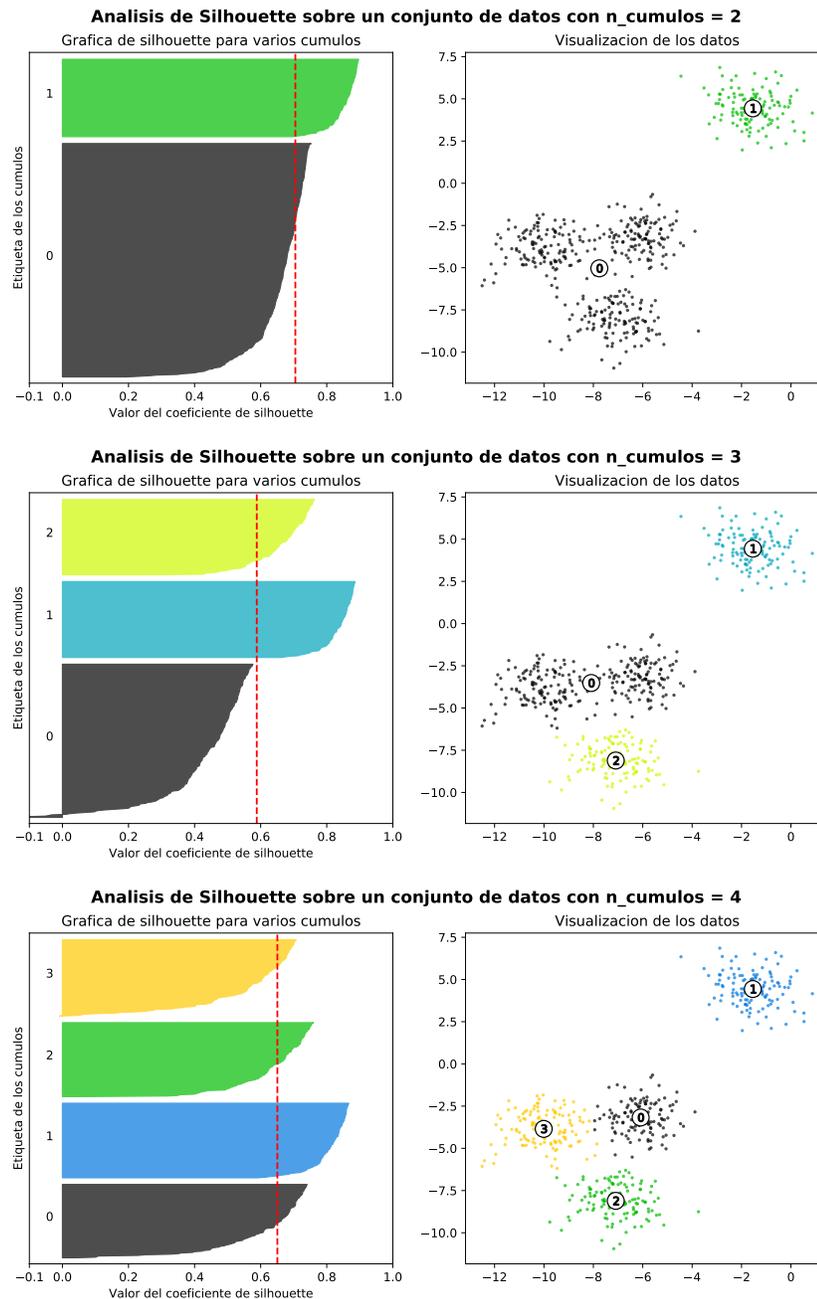


Figura 3.21: Ejemplo de un análisis de la calidad de *clustering* usando el coeficiente de Silhouette. Tenemos una muestra de datos en los paneles derechos. Cada uno se le ha aplicado un *clustering* usando el método *Kmeans* para que identificara 2, 3 y 4 cúmulos. El coeficiente de Silhouette puede ser calculado para cada cúmulo identificado, mostrado en los paneles izquierdos de diferente color y su respectiva etiqueta. El coeficiente promedio de Silhouette es mostrado como una línea discontinua roja, el cual es calculado a partir de todos los coeficientes de Silhouette. Según la muestra de datos y del número de cúmulos identificados el coeficiente tendrá un mayor o menor valor. Basado en un ejemplo de *Scikit-learn* (Pedregosa et al., 2011)

Capítulo 4

Resultados

La última familia de modelos realizada se creó aplicando el algoritmo genético y el método Mean Shift. Por parte del algoritmo genético se tienen dos parámetros: tolerancia (llamado *tol* de aquí en adelante) y σ . Como se explicó en el capítulo anterior, el parámetro *tol* debe ir decreciendo para garantizar que nuestros modelos se acercan a el valor observado. Por otra parte, el parámetro σ depende del número de generaciones como se muestra en la fig.3.4 con un valor inicial $\sigma_0 = 0.5$.

Por parte del método Mean Shift se tiene el parámetro *quantile*. En las familias anteriores donde se aplicó Mean Shift nos dimos cuenta de lo sensible que es el clustering a este parámetro. Y dado que nuestra generación depende del clustering se utilizaron los coeficientes de Silhouette (de Amorim & Hennig, 2015; Rousseeuw, 1987) y el coeficiente de Calinski-Harabaz (Caliński & Harabasz, 1974; de Amorim & Hennig, 2015). Como se explicó en la sección anterior estos dos coeficientes sirven para cuantificar qué tan bien está hecho el clustering. Con estos dos coeficientes se ideó un proceso automático para determinar el parámetro *quantile* óptimo entre dos y cuatro cúmulos. De esta manera garantizamos que el clustering y la elección del valor de *quantile* sea un proceso automatizado.

Por lo tanto, el flujo de nuestro programa en cada generación se muestra en la fig. 4.1. Con un grid de inicio (ya sea el grid inicial de toda la familia o el grid final de la generación anterior) se seleccionan modelos que estén dentro de una tolerancia dada generalmente menor con respecto a la tolerancia de la generación anterior. Con estos puntos seleccionados se realiza el método Mean Shift. Una vez identificadas las dinastías en cada una se realiza un código genético seleccionando los mejores modelos como progenitores generando un número de crías. Y una vez que se obtuvieron los nuevos parámetros en el espacio P se

corren modelos de fotoionización en Cloudy para generar los puntos en el espacio O y se inicia el proceso nuevamente.

En la tabla 4.1 podemos ver todos los parámetros usados en cada generación de esta familia. Para todas las regiones estudiadas se hizo el mismo procedimiento con base en la experiencia de las familias anteriores. La primera generación fue generada como se explicó en el capítulo anterior de forma totalmente aleatoria. En la segunda generación cabe remarcar que no se hizo el algoritmo Mean Shift. Dado que es la generación siguiente al grid inicial se consideró que aún no hay una tendencia clara de las dos soluciones. Por este motivo se omitió el clustering en esta generación. El valor de la tolerancia fue 100 ya que se tomaron todos los puntos del grid inicial. Con estos puntos se escogieron una gran cantidad de progenitores con respecto a otras generaciones para poblar más las zonas cercanas a las soluciones.

A partir de la tercera generación hasta la undécima generación se siguió el patrón establecido: la generación siguiente mantiene o reduce la tolerancia. El cálculo de σ en estas generaciones fue automático siguiendo la ecuación 3.7. En estas generaciones se tomaron entre 10 y 15 progenitores que generaban veinte crías. Estos números se establecieron después de experimentar con distintos valores en las anteriores familias.

Este proceso se repite hasta la generación 12. La generación 13 no se realizó como una generación de la familia final. Esto se debe a que en una familia anterior se consideró hacer una generación 13 con un gran número de progenitores y crías. Se utilizó un valor de $\sigma = 0.3$ para poblar las zonas donde se encuentran las soluciones en el espacio P. Se pensó que sería conveniente, ya que algunas zonas de una de las soluciones tenía muy pocos puntos en el espacio P.

Pero los progenitores eran más numerosos en la zona de la solución que tenía una mayor densidad de puntos. En la otra zona donde se quería que se generarán crías los progenitores seleccionados fueron muy pocos por lo que el efecto no fue el esperado. Por esta razón se descartó esta generación y dado que este tipo de evolución de la familia se usaría en posteriores familias para tener una base se tomó que la generación 13 no se tomará en ninguna familia siguiente.

Para lograr poblar las zonas despobladas se ideó otra manera. En la imagen 4.2 observamos las soluciones en el espacio P de la generación 12. En la gráfica O/H vs N/O los dos conjuntos de soluciones pueden unirse por una línea recta. Revisando todas las regiones disponibles en general podemos unir los dos conjuntos de soluciones por una línea recta.

Se ajustaron de forma lineal los dos conjuntos de soluciones en el espacio O/H vs N/O.

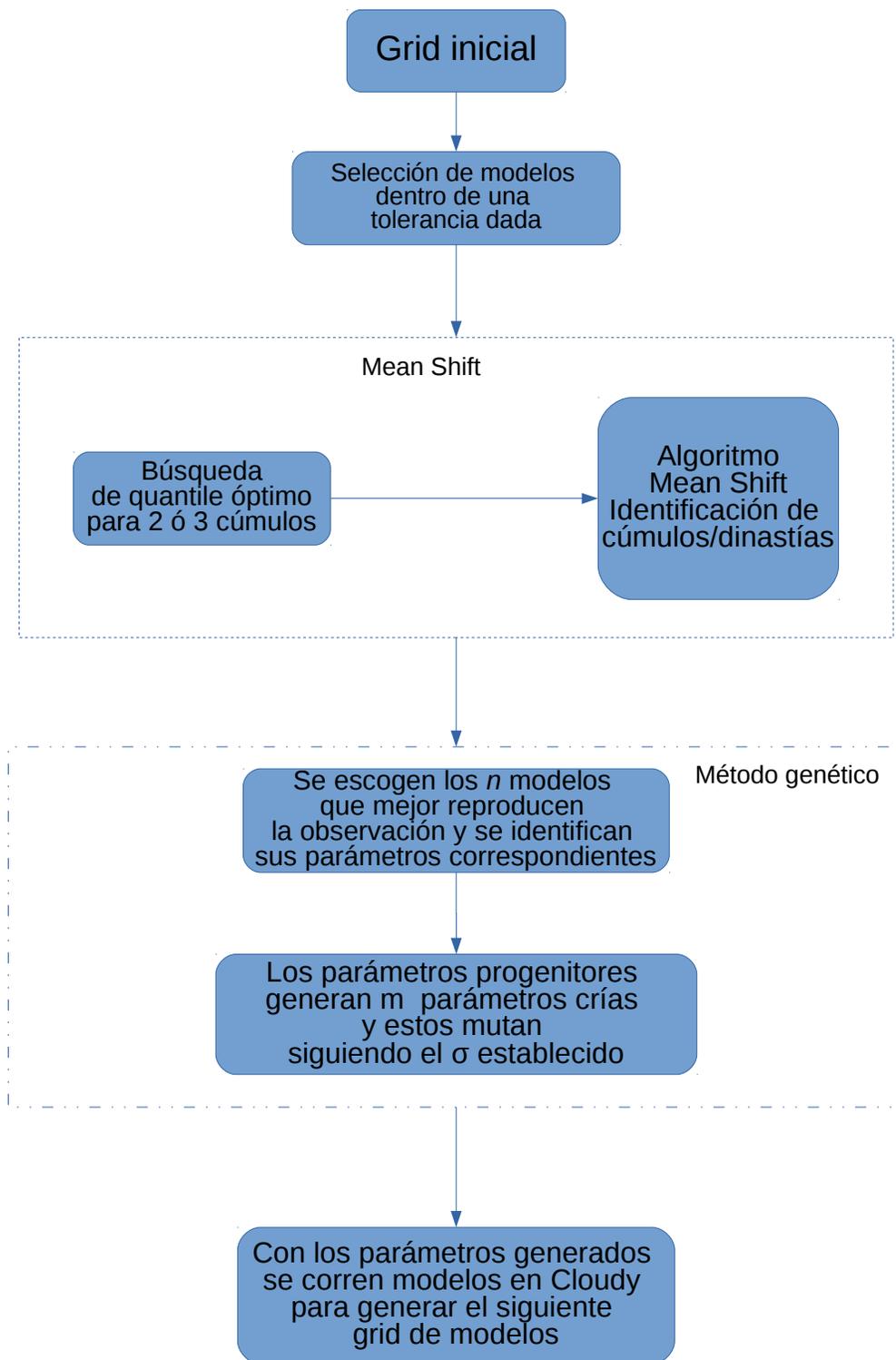


Figura 4.1: Esquema de una generación de nuestro método final. A partir de un grid inicial seleccionan los modelos que mejor reproducen las observaciones dentro de una tolerancia dada. A estos modelos se les aplica el algoritmo *mean shift* buscando el valor de *quantile* que identificará de dos a tres cúmulos. Se aplica el método genético seleccionando los mejores modelos para ser progenitores y generar m crías. Con estos modelos se corren modelos en Cloudy e inicia el proceso nuevamente.

Generación	Tolerancia	σ ($\sigma_0 = 0.5$)	No de Progenitores	No de Crías
1	Creación del grid inicial			
2*	100	Automático	30	10
3	0.40	Automático	10	20
4	0.30	Automático	10	20
5	0.30	Automático	10	20
6	0.20	Automático	10	20
7	0.20	Automático	10	20
8	0.15	Automático	10	20
9	0.10	Automático	10	20
10	0.05	Automático	15	20
11	0.05	Automático	15	20
12	0.1	Automático	10	50
13	Generación no válida			
14	Generación por ajuste lineal Nrand=50, Ntab=20			
15	Generación por ajuste lineal Nrand=50, Ntab=20			
16	0.1	0.3	15	50
17	0.1	0.3	15	50

Tabla 4.1: Parámetros utilizados en cada generación de la última familia. Este mismo proceso fue realizado en cada una de las regiones estudiadas. En la segunda familia no se realizó el clustering. Se tomó todos los puntos como una sola familia y se aplicó el método genético. Lo anterior se hizo para poblar aún más el espacio P. La generación 13 no fue realizada dado que cuando se realizaron pruebas para poblar más el espacio P está generación no se obtuvo el resultado esperado. Por lo cual se omitió en todas las demás familias para comparar el mismo proceso en todas. En las generaciones 14 y 15 se realizó una generación alrededor de un ajuste lineal que se hizo entre los dos conjuntos soluciones. A partir de la generación 16 se siguió con el procedimiento usual de la Fig. 4.1.

Para el valor de $\log U$ se tomó la mediana directamente. Desde un valor de $O/H = -1.3$ a $O/H = 0.5$ se generaron 50 puntos alrededor de esta línea ajustada (en la tabla 4.1 $N_{\text{rand}}=50$). El número $N_{\text{tab}}=20$ en la tabla 4.1 se refiere a cuántos puntos intermedios se tomaron entre $O/H = -1.3$ a $O/H = 0.5$.

Para estos puntos se cambiaron los valores de $\sigma_{O/H} = 0.2$, $\sigma_{N/O} = 0.1$ y $\sigma_{\log U} = 0.05$ (mencionados en el capítulo 3). Este procedimiento fue realizado para generar la generación 14 y 15 en esta y posteriores familias. Cabe mencionar que la generación 13 en esta familia contiene puntos generados aleatoriamente alrededor de un ajuste lineal justo como se acaba de explicar. Se aprovechó esta generación para probar esta producción de puntos alrededor del ajuste lineal.

Por último se realizaron dos generaciones más. Ambas comparten varios parámetros: quince progenitores, cincuenta crías, 0.1 de tolerancia y la $\sigma = 0.3$. A diferencia de todas las anteriores generaciones estas dos generaciones tuvieron un σ fijo. Como se ve en la fig.3.4 el valor mínimo con el modo automático es de 0.05. En estas generaciones el valor se aumentó dado que se quería poblar aun más las zonas y se pensó que el valor de sigma podría afectar al ser tan pequeño.

Después de todo este proceso se obtuvieron para todas las regiones las dos soluciones esperadas. Podemos ver en las figs. 4.3 y 4.4 las soluciones encontradas para las regiones 2 y 25, respectivamente. En el espacio O nuestros modelos difieren un máximo de 0.05 del valor observado. Mientras que en el espacio P se observa claramente las dos soluciones en ambos diagramas: O/H vs $\log U$ y O/H vs N/O .

Una de las dos soluciones está en una zona donde no se encuentra ningún resultado anterior de la literatura. A pesar de que es una solución físicamente posible no parece ser astrofísicamente viable. La otra solución se encuentra justo en la zona donde se localizan diversos ajustes realizados de manera empírica.

Recordando la fig.3.11 teníamos tres posibilidades. En la región 2 (fig.4.3) tenemos las dos soluciones bien definidas. Por otra parte, en la fig.4.4 tenemos a la región 25 donde estamos casi en el punto de quiebre. Es por esto que las dos soluciones que obtenemos están casi juntas. En este caso, se nota una pequeña separación entre ambos grupos de soluciones.

Estos modelos fueron tomados con una tolerancia de 0.05. Si se da una tolerancia mayor entonces el número de modelos que obtendremos será mas grande. Una discrepancia de máximo 0.05 entre los modelos y el observable es bastante bueno.

La solución que es astrofísicamente posible en ambas regiones está cerca de ajuste hecho

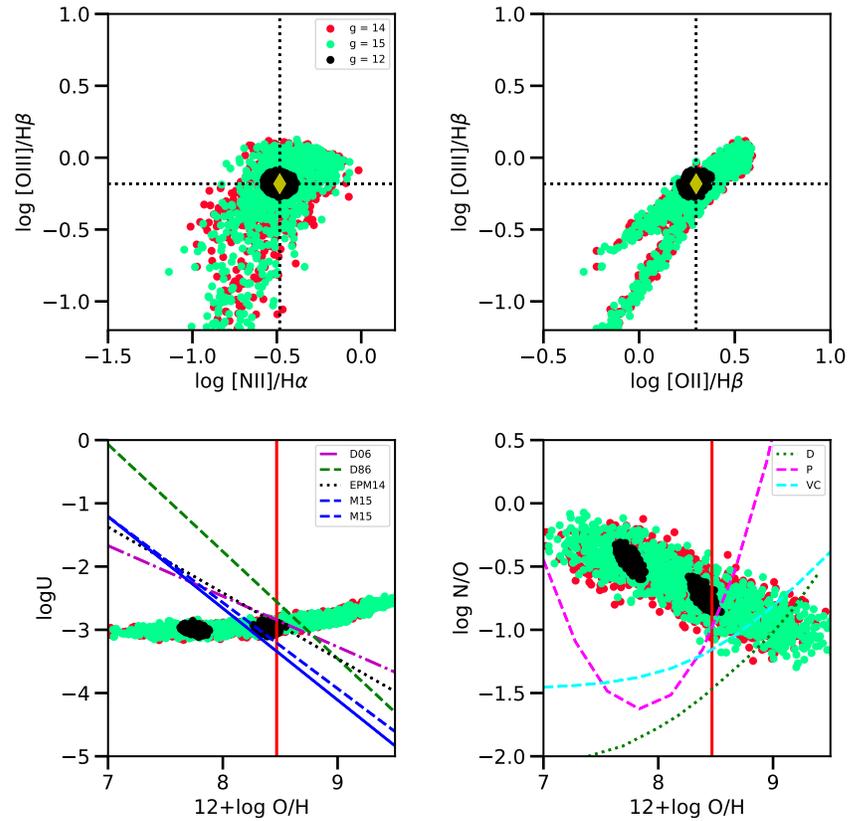


Figura 4.2: Generación de puntos por ajuste lineal de las generaciones 14 y 15. La generación 12 (puntos negros) son tomados como bases para hacer un ajuste lineal. Alrededor de esta línea se generan los puntos para la generación 14 y 15. Se generan puntos desde un valor de O/H igual a 7 hasta aproximadamente 9. Esto se debe al σ utilizado para generar estos puntos. Este proceso se realizó con el fin de poblar aún más las zonas donde no había una alta densidad de puntos.

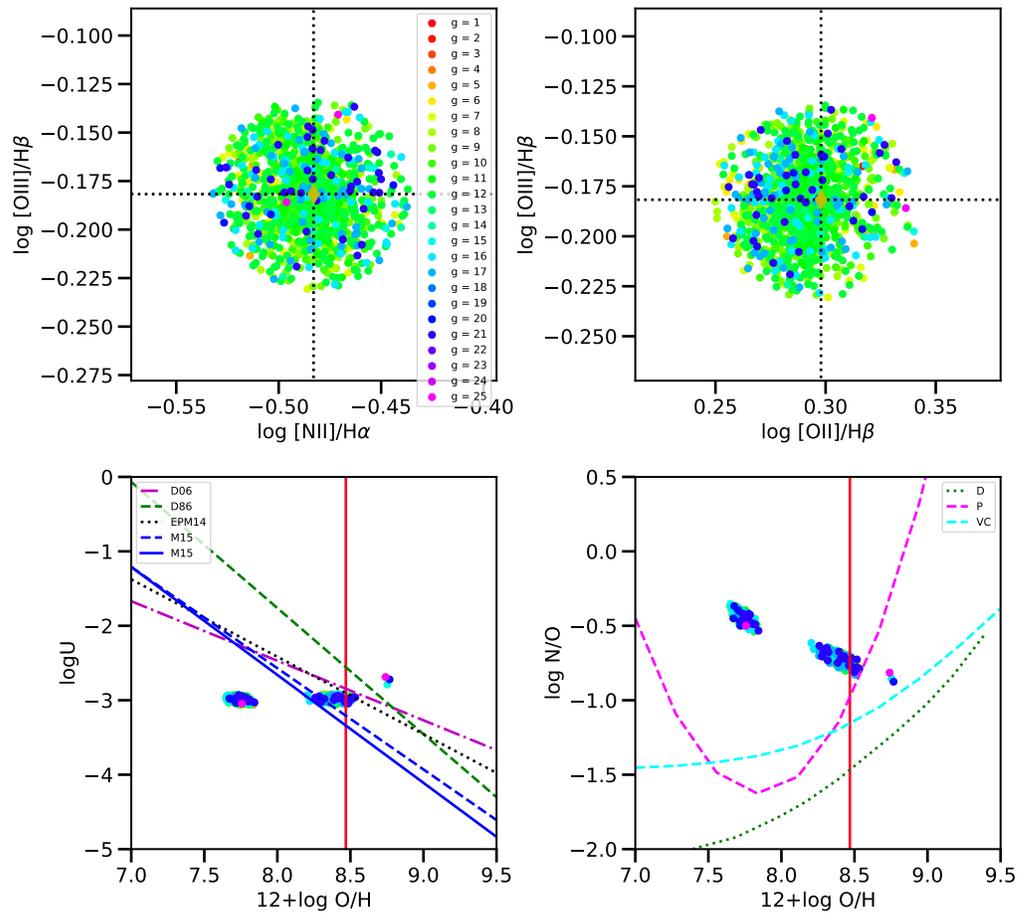


Figura 4.3: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la última familia de modelos de la región 2. Observamos como nuestros modelos reproducen la observación con una diferencia máxima de 0.05 dex en el espacio O. En el espacio P se encontraron los dos conjuntos solución principales más un tercer conjunto muy pequeño el cual corresponde a lo explicado al final del capítulo 1 (ver Fig. 1.11) Las líneas mostradas son las mismas que las de la Fig. 3.5

por Pilyugin et al. (2012) (línea discontinua) el cual es un ajuste empírico a diferencia del ajuste de Dopita et al. (2013) (línea punteada) que es hecho por medio de modelos de fotoionización del cual nuestros modelos difieren.

Por otra parte, en el artículo previo (Morisset et al., 2016) se utilizó el calibrador de Marino et al. (2013). Este calibrador está marcado como la línea sólida en nuestras figuras. Notemos cómo nuestros modelos también pasan por la zona de este ajuste el cual también es un ajuste empírico.

En la fig. 4.4 tenemos además cuatro gráficas más:

- La primera es del espacio $[\text{OIII}]4363/5007$ vs $[\text{OIII}]\text{H}\beta$
- El segundo es el espacio $[\text{NII}]5755/6584$ vs $[\text{OIII}]\text{H}\beta$
- El tercero $[\text{ArIII}]7135/[\text{NeIII}]3869$ vs $[\text{OIII}]\text{H}\beta$.

En cada una de estas gráficas, podemos apreciar cómo existen los dos conjuntos de soluciones. Recordemos que estas líneas son importantes ya que los cocientes $[\text{OIII}]4363/5007$ y $[\text{NII}]5755/6584$ son indicadores de temperatura electrónica. Si estas líneas son observadas se puede discriminar entre los dos conjuntos de soluciones que tenemos e indicar cuál es la solución correcta. Pero estas líneas casi nunca se ven en el survey CALIFA. Por otra parte, el cociente $[\text{ArIII}]7135/[\text{NeIII}]3869$ nos sirve para conocer las abundancias relativas y también dependen de la T_e . Pero en nuestro caso las usamos ya que fueron propuestas por Vale Asari et al. (2016). En este trabajo usan el cociente de $[\text{ArIII}]7135/[\text{NeIII}]3869$ para separar los modelos de alta abundancia de oxígeno de los de baja abundancia como se mencionó en el Cap. 1. En nuestro trabajo queremos ver si hay un comportamiento claro para separar las soluciones de nuestros modelos.

En la gráfica se alcanzan a distinguir los dos conjuntos de soluciones por lo cual se tiene el mismo problema que con los otros cocientes. Si no se tiene alguna observación que indique cuál de las dos soluciones es la correcta no podemos asegurar que estemos calculando de manera correcta los parámetros correspondientes.

En el apéndice A se pueden encontrar todos los modelos finales para todas las regiones disponibles para la galaxia NGC5947 en CALIFA. Recordar que estas regiones son solamente aquellas regiones que son ionizadas por estrellas jóvenes.

Tenemos en total 24 regiones cada una modelada con dos valores para $fr = 0.03$ y $fr = 3.0$. Como hemos visto en las figuras de los modelos finales encontramos diferentes conjuntos de soluciones. Calculamos los centroides de cada conjunto de soluciones para ambos valores

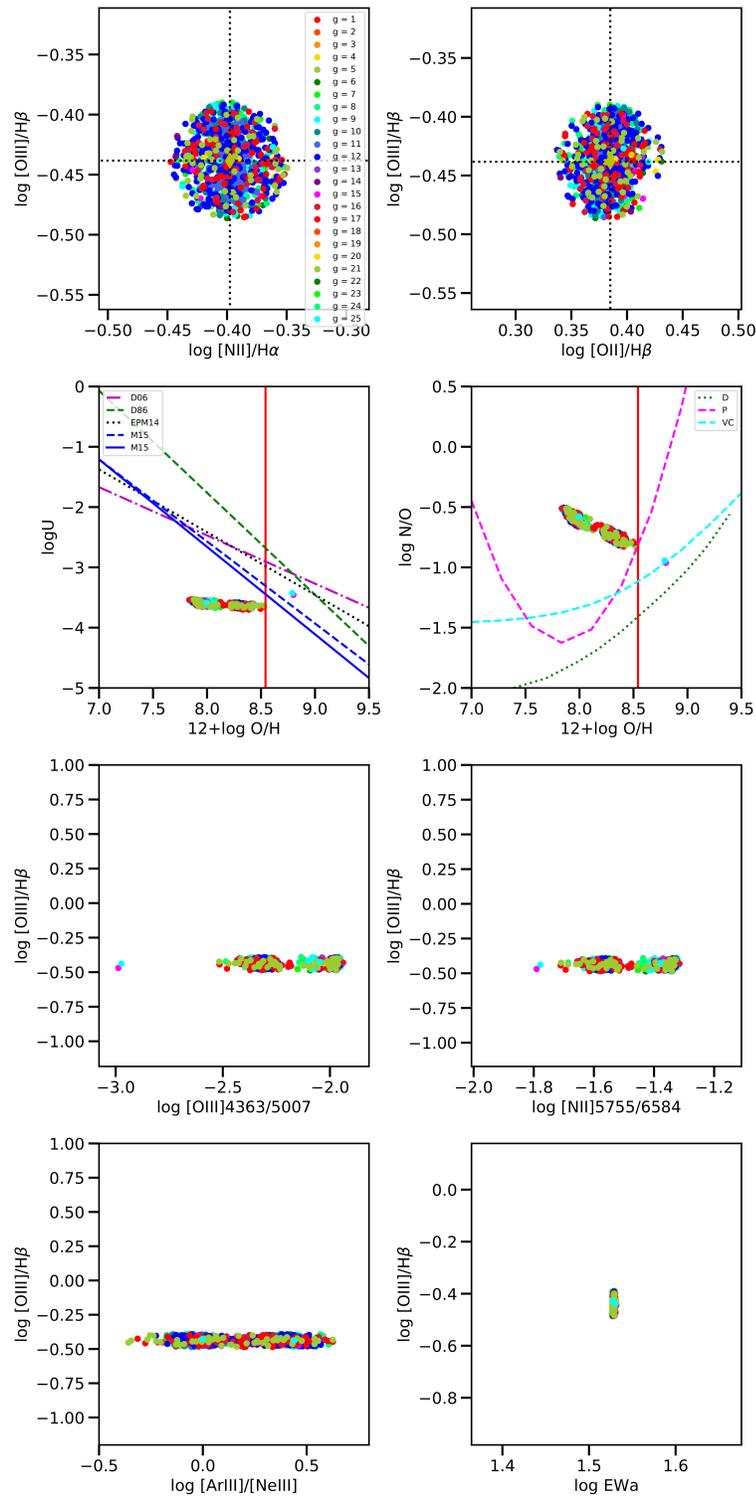


Figura 4.4: Modelos finales que difieren de la observación un máximo de 0.05 de la última familia de modelos de la región 25. Las líneas mostradas son las mismas que las de la Fig. 3.5. Tenemos un comportamiento similar que en la región 2 en los cuatro primeros paneles. En la tercera fila se muestran el valor de $[\text{OIII}]4363/5007$ vs $[\text{OIII}]$ y $[\text{NII}]5755/6584$ vs $[\text{OIII}]$ en el cual vemos que de conocerse el valor de $[\text{OIII}]4363$ ó $[\text{NII}]5755/6584$ podríamos diferenciar qué conjunto solución es correcto. Esto también se observa en el panel que muestra $[\text{ArIII}]/[\text{NeIII}]$. Por último se muestra EWa vs $[\text{OIII}]$. EWa está relacionado con el cociente entre el número de fotones procesados por el gas y el número de estrellas ionizantes, pero no hay diferencia alguna entre los dos conjuntos de soluciones.

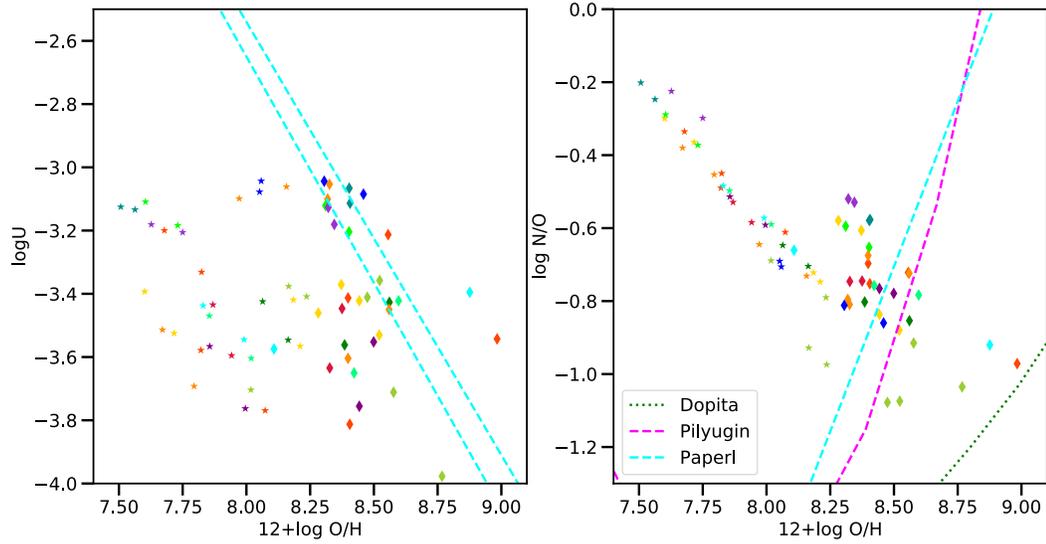


Figura 4.5: Valores promedio de modelos finales para las dos geometrías ($fr = 0.03$ y 3.0). Los puntos mostrados son el promedio de todos los modelos que reproducen las observaciones para cada región. Se muestran con diamantes los valores promedio de los modelos que son astrofísicamente posibles. Los valores representados con asteriscos son aquellos que se encuentran apartados de las zonas de los calibradores de la literatura. Estos modelos y sus respectivos valores promedio los consideramos astrofísicamente improbables.

de fr y para todas las regiones los cuales se muestran en la Fig. 4.5. En esta figura los asteriscos representan las soluciones están alejadas de los ajustes de la literatura. Estas soluciones consideramos que no son astrofísicamente posibles. Los diamantes representan los modelos que son astrofísicamente posibles, es decir, son las que se encuentran en la zona de los ajustes de la literatura. Los asteriscos nos indican un valor de O/H bajo y los diamantes un valor mayor de O/H. Todos los diamantes coinciden con el ajuste hecho por Morisset et al. (2016) (líneas cyan discontinuas). Por otra parte, también coinciden con el ajuste de Pilyugin et al. (2012) que está hecho a partir de observaciones. Por lo que en general nuestros modelos son consistentes con los resultados obtenidos por los ajustes empíricos.

Cambio de la edad de la población estelar subyacente

En capítulos anteriores se explicó que la población estelar subyacente que ioniza a las regiones HII está ajustada por medio de FIT3D/POPSTAR. Se toma el espectro observado,

se ajusta y a partir de éste se calculan las poblaciones estelares. Como resultado de este proceso se obtiene la fracción de estrellas jóvenes y estrellas viejas que ionizan el medio (ver capítulo 2). Pero las estrellas que ionizan las regiones HII son las estrellas jóvenes.

Las estrellas jóvenes no emiten de manera importante en el óptico. Si tomamos los resultados de POPSTAR veremos que las estrellas jóvenes contribuyen menos del 6% del espectro observado en el rango visible. Pero si tomamos la cantidad de fotones que ionizan el hidrógeno neutro, es decir Q_0 (capítulo 1) encontraremos que las estrellas ionizantes más importantes son las estrellas jóvenes. Estas estrellas llegan a producir cerca del 99% de fotones ionizantes en las regiones HII.

FIT3D nos entrega una lista de estrellas con diferentes edades y metalicidades con sus respectivos pesos para reproducir el espectro observado. Pero las estrellas jóvenes tienen poca contribución en el rango visible. Por lo que es probable que en la derivación de la edad de la población estelar existan errores.

Para buscar si hay algún efecto de la edad de la población estelar ionizante se realizaron distintas familias donde se cambiaron los pesos de las estrellas jóvenes. Es decir, en una familia se impuso que la población estelar 100% dominante fuera de 1 Myr de edad, en otra de 3 Myr, en otra más de 4 Myr, etc.

Cada familia siguió una evolución igual a la mostrada en la tabla 4.1. Se utilizaron los mismos parámetros: tolerancia, σ , número de progenitores, etc. para cada generación. De esta forma podemos comparar cada familia dado que el proceso hasta la última generación es idéntico.

En la Fig. 4.6 podemos ver las distintas familias y sus diferentes edades de la población estelar joven que se realizaron. En total fueron seis familias con distintas edades:

- 1 Myr
- 2 Myr
- 2.5 Myr
- 2.75 Myr
- 3 Myr
- 4 Myr

Como se ve en la figura todos nuestros modelos difieren de la observación un máximo de 0.05. En el espacio de parámetros es donde tenemos el efecto del cambio en la edad de la

población estelar. En el espacio O/H vs $\log U$ tenemos cómo va variando el valor de $\log U$ conforme cambia la edad. El parámetro de ionización tiene que cambiar con el cambio de edad ya que $\log U$ nos dice qué tan ionizada está la nebulosa. Podemos ver cómo parte de un valor aproximadamente de $\log U = -3.15$ con una población de 1 Myr. Alcanza un máximo cuando se llega a una edad estelar de 2.5 Myr. Pero da la vuelta y empieza a disminuir conforme la edad crece. Podemos ver que la edad de 4Myr, que fue la edad más alta que se realizó, corresponde al $\log U$ más bajo.

Por otra parte, en el espacio O/H vs N/O observamos que el cambio es menor que el que se tuvo para $\log U$. Podemos observar que aunque se notan las dos soluciones el cambio en cada una de estas es menor de 0.1 dex. Por lo que el cambio en la edad de la población estelar no nos produce un cambio significativo en el cálculo de nuestras soluciones.

Por lo que un error en la derivación de las edades de la población estelar ionizante no tendrá un efecto significativo en los conjuntos de soluciones que obtenemos para O/H y N/O. Como hemos visto el valor del parámetro de ionización cambiará conforme cambie la edad de las estrellas.

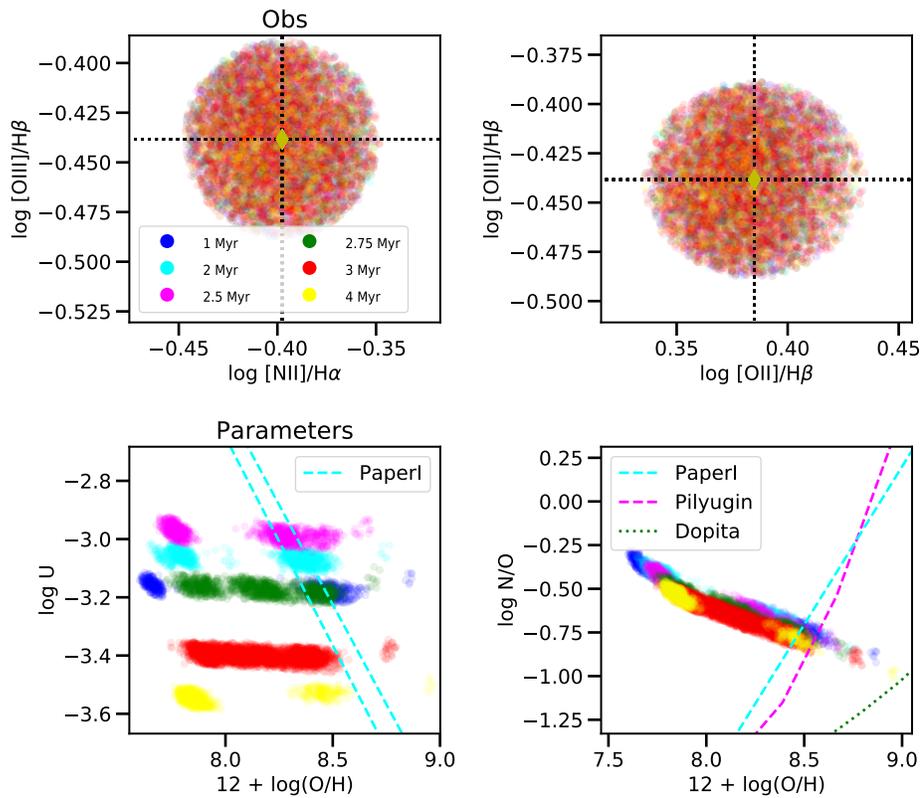


Figura 4.6: Modelos con diferentes edades para la población estelar ionizante sólo realizados para la región 25. En los paneles superiores se muestran que todos nuestros modelos para todas las edades utilizadas reproducen las observaciones con una discrepancia máxima de 0.05 dex. En los paneles inferiores mostramos nuestro espacio P. Observamos que el parámetro de ionización U cambia respecto a la edad de la población estelar. Notemos cómo el valor de U da la vuelta alrededor de una edad de 2.5 Myr. Para los valores de N/O y O/H observamos que no hay una gran diferencia entre los parámetros calculados con distintas edades.

Capítulo 5

Discusión de resultados y trabajo a futuro

Como se había discutido anteriormente existe una relación bi-valuada entre la abundancia de oxígeno (O/H) y la emisión $[OIII]5007/H\beta$. Esto implica que en nuestro espacio de parámetros existirán al menos dos soluciones, una de baja y otra de alta abundancia de O, que reproducen las observaciones. Si no podemos decir si estamos en la zona de abundancia alta o baja no sabremos si estamos obteniendo soluciones astrofísicamente posibles.

En la fig.4.3 se muestra la región 2 de nuestros modelos. Como podemos observar en el espacio de parámetros (O/H vs N/O y O/H vs $\log U$) hay dos conjuntos de soluciones. En general para todas las regiones estudiadas existen uno o dos conjuntos de soluciones como se mostró en la fig.3.11.

Estos conjuntos de soluciones generan modelos en el espacio O que reproducen las observaciones de $[OIII]$, $[NII]$ y $[OII]$ con una discrepancia de 0.05 dex del valor observado. La discrepancia está dentro de los errores observacionales de CALIFA. En el espacio de observables no podemos diferenciar de ninguna forma los modelos con parámetros astrofísicamente posibles de los que no lo son. Por lo tanto, no podemos asegurar que los parámetros que se obtienen al ajustar las observaciones sean los correctos sin ninguna información extra que nos indique en qué régimen de metalicidad nos encontramos.

Nuestros modelos son completamente teóricos sin hacer uso de una relación N/O vs O/H para su producción. Para obtener las dos soluciones se tuvieron que usar distintos métodos para lograr que ambas permanecieran hasta el final del proceso. Como hemos visto, no es sencillo obtener las dos soluciones dado que ambas reproducen las observaciones

tomadas. Además, nuestro método genético tiende a eliminar una de las dos soluciones durante el proceso evolutivo. Se tuvo que implementar la combinación del método genético y el algoritmo Mean Shift de análisis de cúmulos para poder obtener ambas soluciones al mismo tiempo. Esta combinación de métodos nos permitió que durante todo el proceso no perdiéramos en ningún momento alguna de las dos soluciones.

En todas las regiones se encontraron dos conjuntos solución principales. Pero como se mencionó en el capítulo 1 pueden existir más de dos soluciones si se aumenta la resolución y el número de parámetros libres. En algunas regiones se observa que existe un tercer conjunto de soluciones que están cerca de los ajustes de Dopita et al. (2013). Estas soluciones también reproducen las observaciones sólo difiriendo del valor observado 0.05 dex. Posiblemente si disminuyéramos la tolerancia y aumentáramos el número de modelos podríamos obtener más subconjuntos de soluciones a partir de los conjuntos de soluciones principales.

En las Fig. 5.1 y Fig. 5.2 se muestran los mismos puntos de la Fig. 4.5 separados por valores de $f_r = 0.03$ y $f_r = 3.0$, respectivamente. En ambas figuras se muestran las 24 regiones y los modelos que reproducen las observaciones de cada región (puntos en gris claro). Se muestran los valores promedio de todos los modelos por región (puntos de colores representados por diamantes y estrellas). Los puntos diamante son aquellos modelos que tienen una alta probabilidad de representar regiones HII reales. Los puntos con forma de estrella son los modelos que pueden ser objetos no astrofísicos. Esta clasificación se hizo por la cercanía de los modelos a los calibradores y resultados de la literatura. También se muestran contornos que representan los conjuntos solución. Toda el área que encierran estos contornos reproducen las observaciones de las regiones HII. Los contornos sólidos son aquellos tienen una alta probabilidad de representar regiones HII reales. Los contornos discontinuos son para los modelos que tienen una baja probabilidad de ser objetos astrofísicos reales.

La clasificación que se hace entre modelos que representan objetos astrofísicos reales y los que no está hecha dado que conocemos trabajos y resultados anteriores. Sin saber estos resultados no podríamos distinguir entre los conjuntos de parámetros que reproducen regiones HII reales y los que no son regiones reales.

En el trabajo de Dors et al. (2017) para una muestra de galaxias Seyfert 2 se determinan abundancias de nitrógeno. El criterio para seleccionar los modelos son los que mejor reproduzcan las líneas de emisión consideradas: $[\text{OII}]\lambda 3726$, $[\text{OIII}]\lambda 5007$, $[\text{NII}]\lambda 6584$ y $[\text{SII}]\lambda 6716$ con un error del 20%. Realizan una malla de modelos con Cloudy variando la metalicidad Z , N/H y S/H y en otros modelos varían N_e , R_{in} y el exponente α de $F_\nu \sim \nu^\alpha$ que describe

la SED.

A pesar de ajustar los cocientes de línea pueden estar derivando los parámetros que no son astrofísicamente correctos. Sin calibrar antes el método o decir cómo acotan sus soluciones en su correspondiente espacio P no se puede asegurar que todos los valores obtenidos sean correctos

En la Fig. 5.3 y Fig. 5.4 mostramos los resultados de los conjuntos solución de O/H vs el cociente O3N2 para un de $f_r = 3.0$ y $f_r = 0.03$, respectivamente. Se muestra el calibrador realizado por Marino et al. (2013) el cual utiliza el cociente O3N2 para estimar el valor de O/H. Se muestran todos los valores de O/H de nuestros modelos que representan una región HII para cada una de las regiones modeladas con diferentes colores. En general vemos que hay dos conjuntos solución, uno de ellos concuerda con el calibrador de Marino y, a su vez, es el mismo conjunto solución que concuerda con los calibradores empíricos.

También se observan pequeños conjuntos solución a un valor más alto de O/H. Estos conjuntos solución concuerdan con los calibradores realizados por modelos de fotoionización. Por lo que se obtienen soluciones igualmente posibles para ambos calibradores. Aunque un mayor número de nuestros modelos caen en la zona de los calibradores empíricos.

Sabemos que nuestros modelos dependen del ajuste de la población estelar. Sin embargo, no podemos asegurar completamente que el ajuste obtenido es correcto. Como hemos visto, se buscó cuál era el efecto si cambiábamos la edad de la población estelar ionizante. En nuestra Fig. 4.6 observamos cómo en el espacio O/H vs N/O las diferentes edades de la población estelar ionizante no presenta un gran cambio. En el espacio O/H vs logU se presenta un cambio en el valor de logU dependiendo de la edad estelar.

Conclusiones

En este trabajo se realizaron modelos de fotoionización basados en las regiones HII del catálogo de CALIFA. Se utilizó solo una galaxia, NGC 5947, por ser una galaxia con muchas regiones HII. El espectro ionizante usado en estos modelos es derivado a partir del análisis de FIT3D (Sánchez et al., 2016) y la población estelar modelada por POPSTAR (Mollá et al., 2009). Se utilizaron 25 regiones de la galaxia NGC 5947 dado que estas son ionizadas por poblaciones estelares jóvenes. Las regiones descartadas son aquellas en que la población estelar ionizante dominante es vieja.

Los modelos realizados anteriormente por Morisset et al. (2016) consideraban cómo

parámetros libres $\log U$ y N/O . Para la abundancia de oxígeno se usó el ajuste hecho por Marino et al. (2013). Nuestros modelos ya no usan el ajuste para la abundancia de oxígeno y lo toman como un parámetro libre. Nuestros modelos, por lo tanto, buscan los valores de $\log U$, O/H y N/O que ajustan las observaciones de los cocientes de líneas $[NII]/H\alpha$ y $[OIII]/H\beta$. Se usaron dos morfologías en nuestros modelos: esfera llena y un cascarón de esfera.

Sabemos que existe una degeneración en los métodos de líneas fuertes ampliamente comentada en la literatura (Stasińska, 2002). Pero al tener tres parámetros libres ($\log U$, O/H y N/O) es posible que no sólo tengamos dos valores de parámetros que ajusten nuestras observaciones sino un número mayor de soluciones para nuestros parámetros que ajusten las observaciones.

Los conjuntos de soluciones obtenidas por medio del método genético y el algoritmo Mean Shift reproducen las observaciones con una diferencia máxima de 0.05 dex. Estos conjuntos son un grupo de parámetros que reproducen los datos observados. De los dos conjuntos de soluciones principales obtenidos uno de ellos es astrofísicamente compatible con la relación N/O vs O/H y el otro no. Las soluciones astrofísicamente posibles que se derivaron son compatibles con los ajustes hechos por Pilyugin et al. (2012, basados en observaciones) y en menor medida con los calibradores realizados por Dopita et al. (2013, basados en modelos de fotoionización).

Llegamos a una degeneración en la determinación de los parámetros que esperábamos (Fig. 1.10). Se puede resolver la degeneración usando la relación N/O vs O/H o tomando en cuenta otra línea de emisión (por ejemplo $[OIII]\lambda 4363$). Esto será parte del trabajo a futuro que se puede realizar en el doctorado.

Trabajo a futuro

Dado que sólo se realizó este proceso para una sola galaxia del catálogo de CALIFA, en un futuro trabajo se planea realizar el mismo proceso para todas las galaxias y regiones HII de CALIFA. Además podemos explorar otro tipo de métodos de clustering que posiblemente mejoren aun más nuestro proceso genético. En este trabajo se ajustaron tres parámetros pero podríamos explorar la posibilidad de ajustar otros cocientes de líneas. En otras regiones se tiene la línea de $[OIII]\lambda 4363$ lo cual nos permitirá saber en qué régimen de metalicidad nos encontramos. Por lo que podremos comparar nuestros modelos con los modelos realizados para estas regiones donde sí se tiene información observacional.

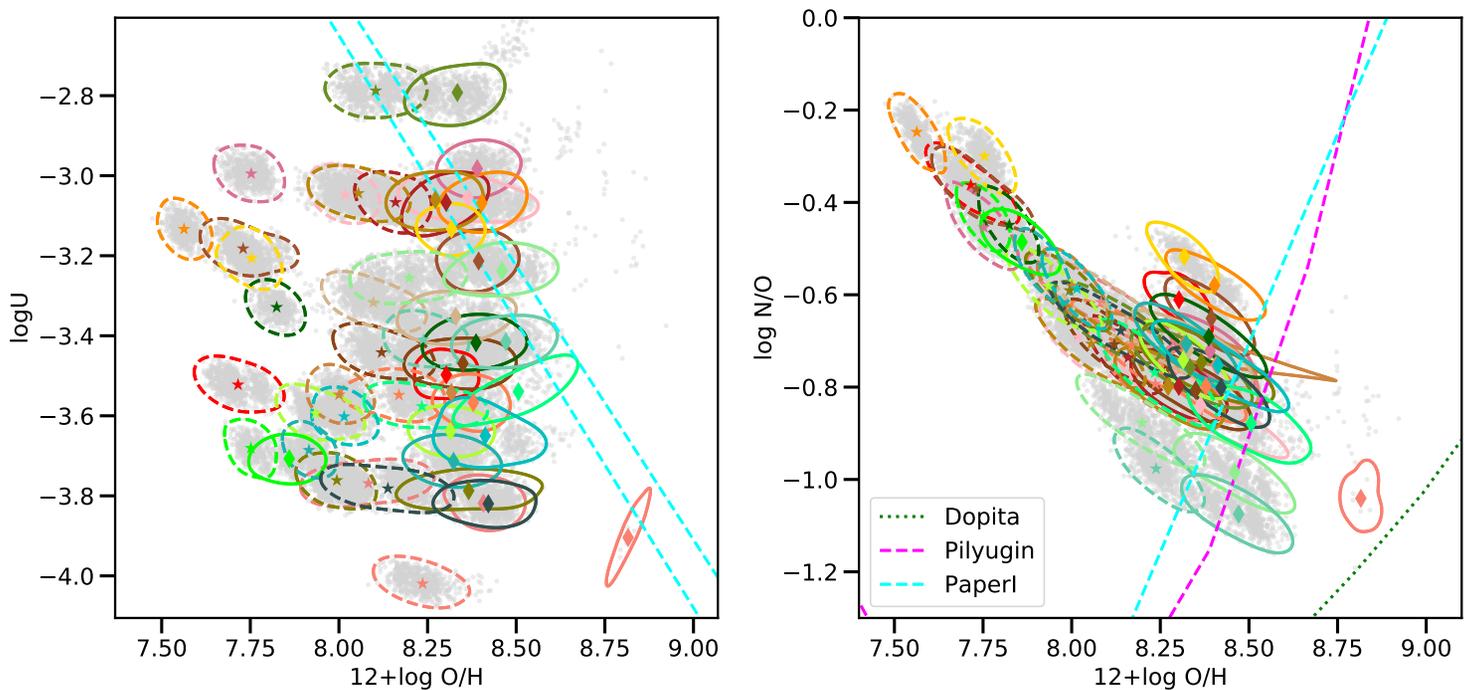


Figura 5.1: Conjuntos de soluciones de todas las regiones modeladas para $f_r = 0.03$. Se muestran todos los modelos que reproducen la observación en color gris claro. También se muestran los valores promedio por región de todos los modelos correspondientes. Cada color representa una región diferente. Los diamantes representan soluciones que tienen una gran probabilidad de representar regiones HII reales. Mientras que las estrellas representan modelos que son físicamente correctos pero hay una baja probabilidad de que sean objetos astrofísicos reales. El criterio para clasificarlos de esta forma es la cercanía que tienen con las líneas mostradas. Las líneas discontinuas es el ajuste de Morisset et al. (2016) llamado PaperI. La línea morada discontinua representa el calibrador de Pilyugin et al. (2012) y la línea verde punteada el calibrador de Dopita et al. (2013). Junto a cada punto de valor medio es mostrado un contorno. Este contorno representa dónde se concentran la mayor cantidad de modelos que reproducen la observación para la región correspondiente. Es decir, el área que delimita el contorno es un conjunto solución de la región. Toda esa área puede reproducir la observación correspondiente. Los contornos sólidos son aquellos que corresponden a una región HII real y los contornos discontinuos representarán los modelos que tienen una alta probabilidad de no representar un objeto astrofísico real.

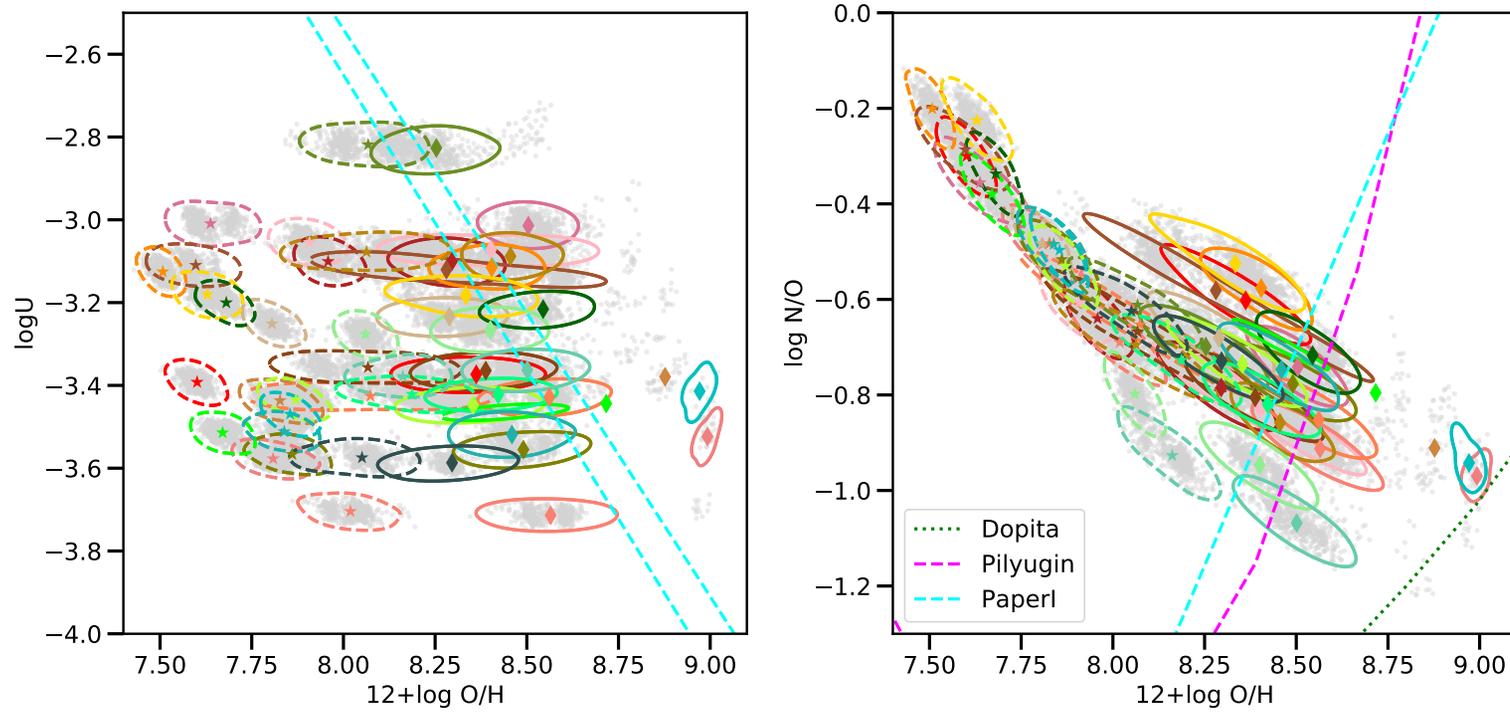


Figura 5.2: Conjuntos de soluciones de todas las regiones modeladas para $f_r = 3.0$ con sus respectivos contornos. Está figura es análoga a la Fig. 5.1.

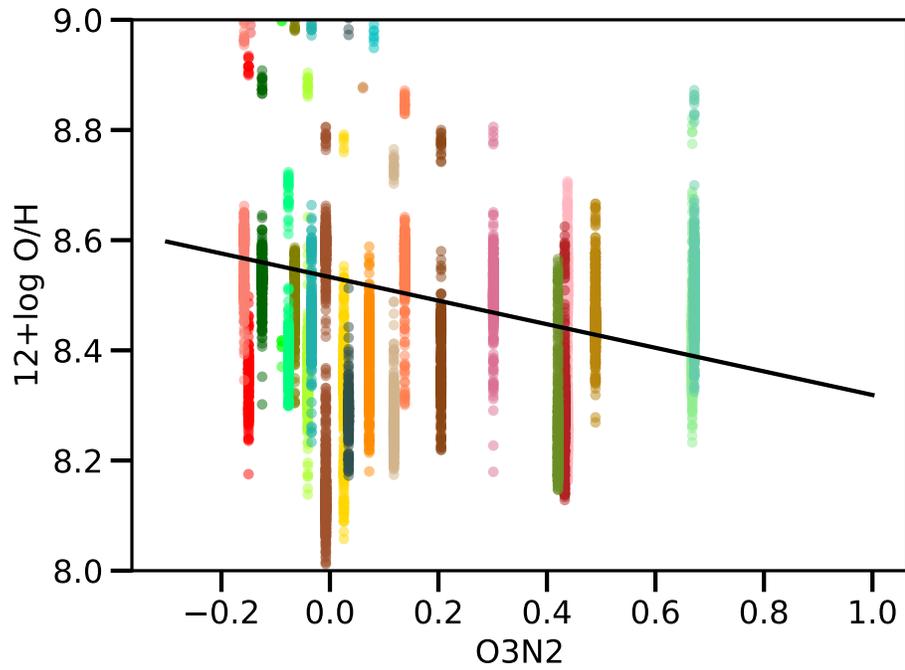


Figura 5.3: Abundancia de oxígeno vs el coeficiente O3N2. Cada color corresponde a la solución astrofísicamente posible de una región modelada con un $f_r = 3.0$. Se muestran los valores de la abundancia de O para cada región. Sólo se muestran los modelos que tienen una alta probabilidad de representar un objeto astrofísico real. La Línea negra representa el calibrador realizado por Marino et al. (2013). Las líneas definidas por nuestros modelos representan los conjuntos solución que son compatibles con el calibrador de Marino.

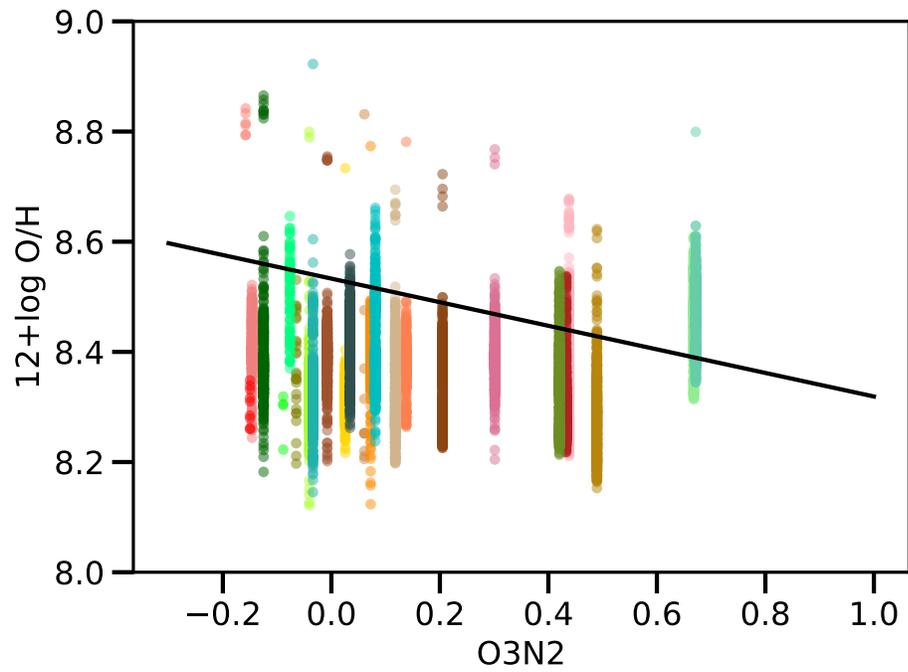


Figura 5.4: Abundancia de oxígeno vs el coeficiente O3N2. Cada color corresponde a la solución astrofísicamente posible de una región modelada con un $f_r = 0.03$. Esta imagen es análoga a la Fig. 5.3

Apéndices

Apéndice A

Modelos finales para todas las regiones

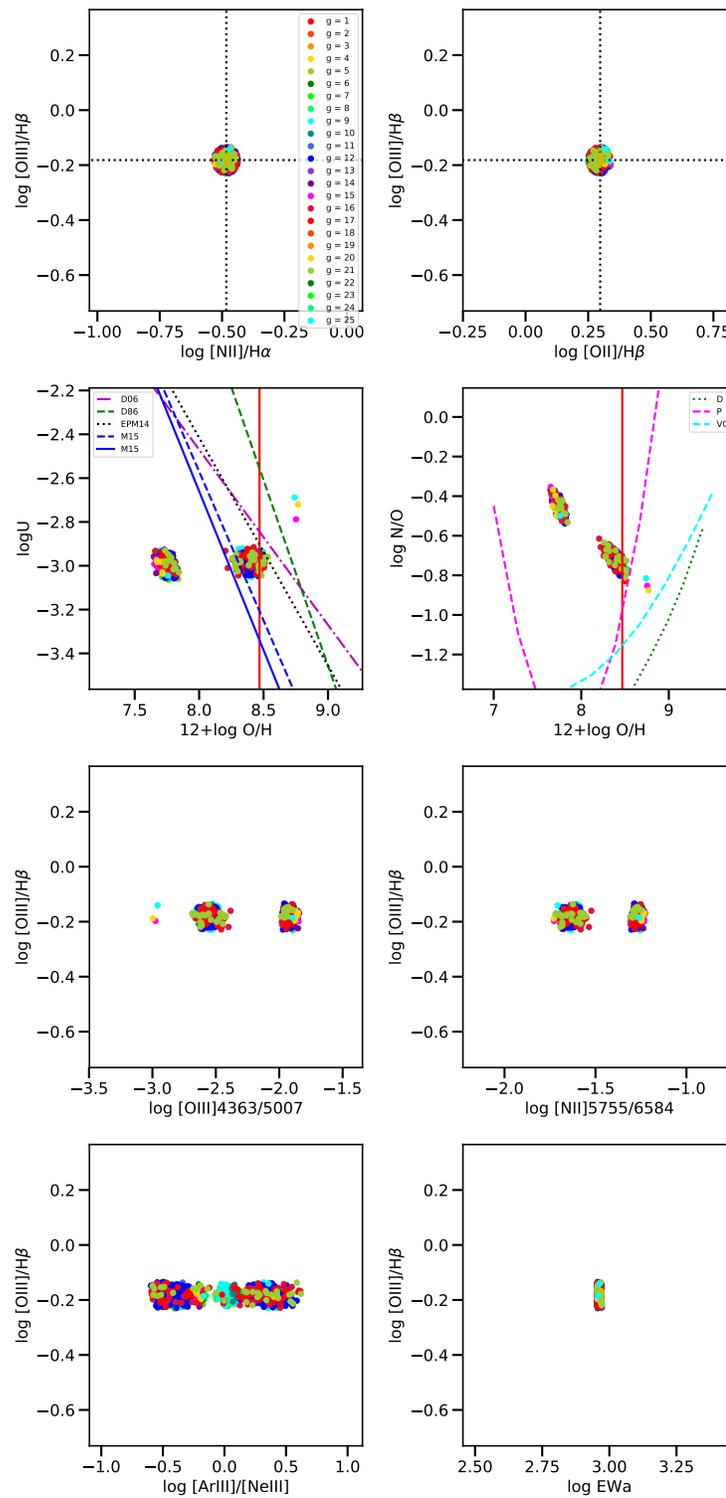


Figura A.1: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 2

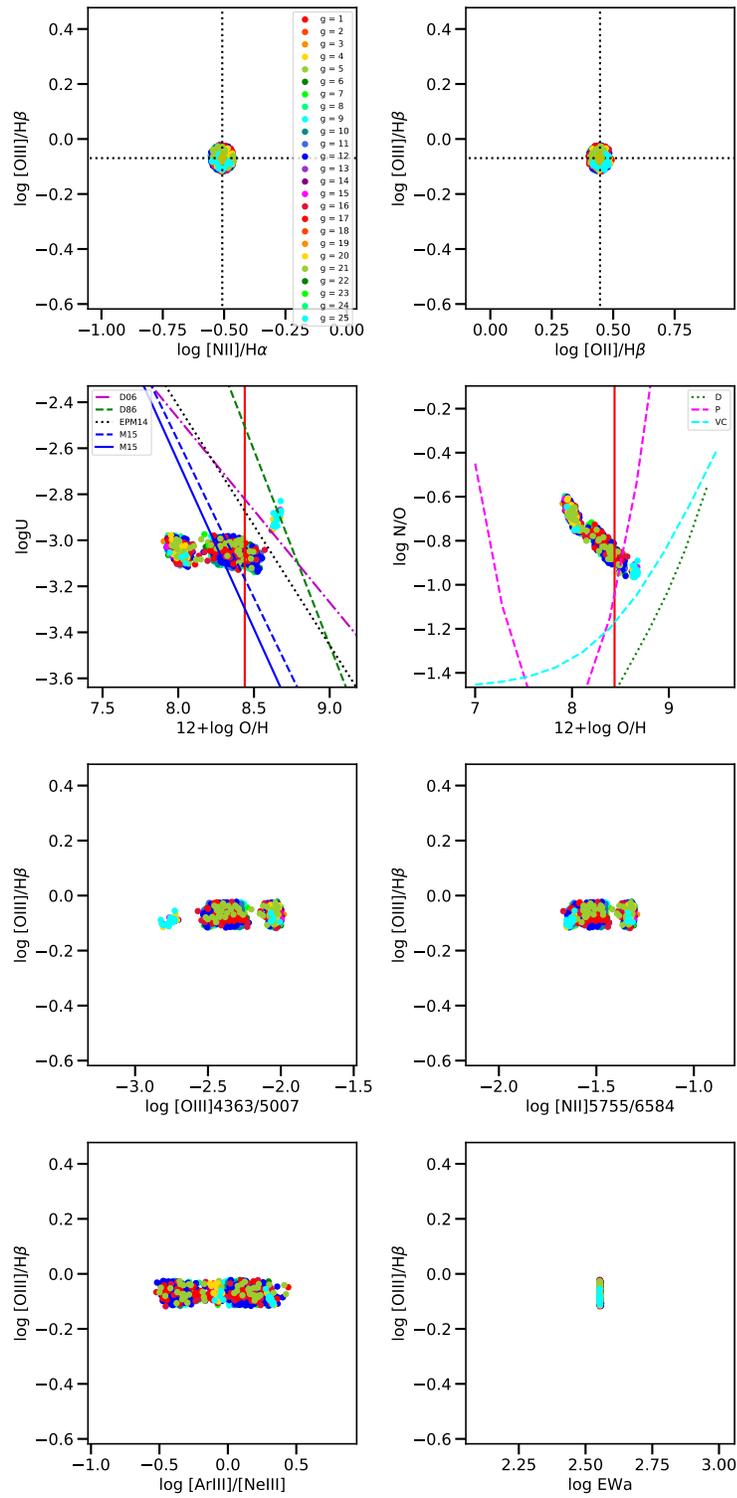


Figura A.2: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 3

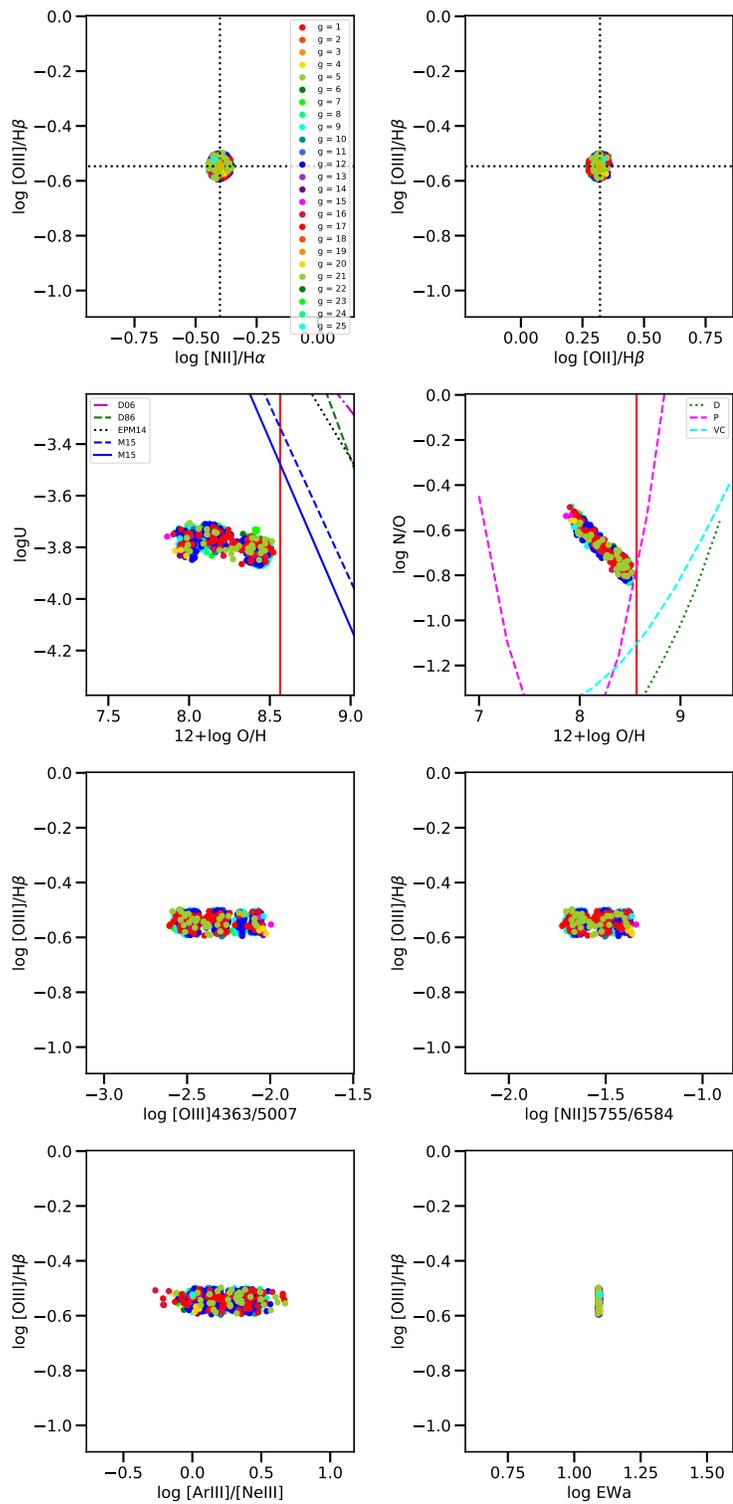


Figura A.3: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 5

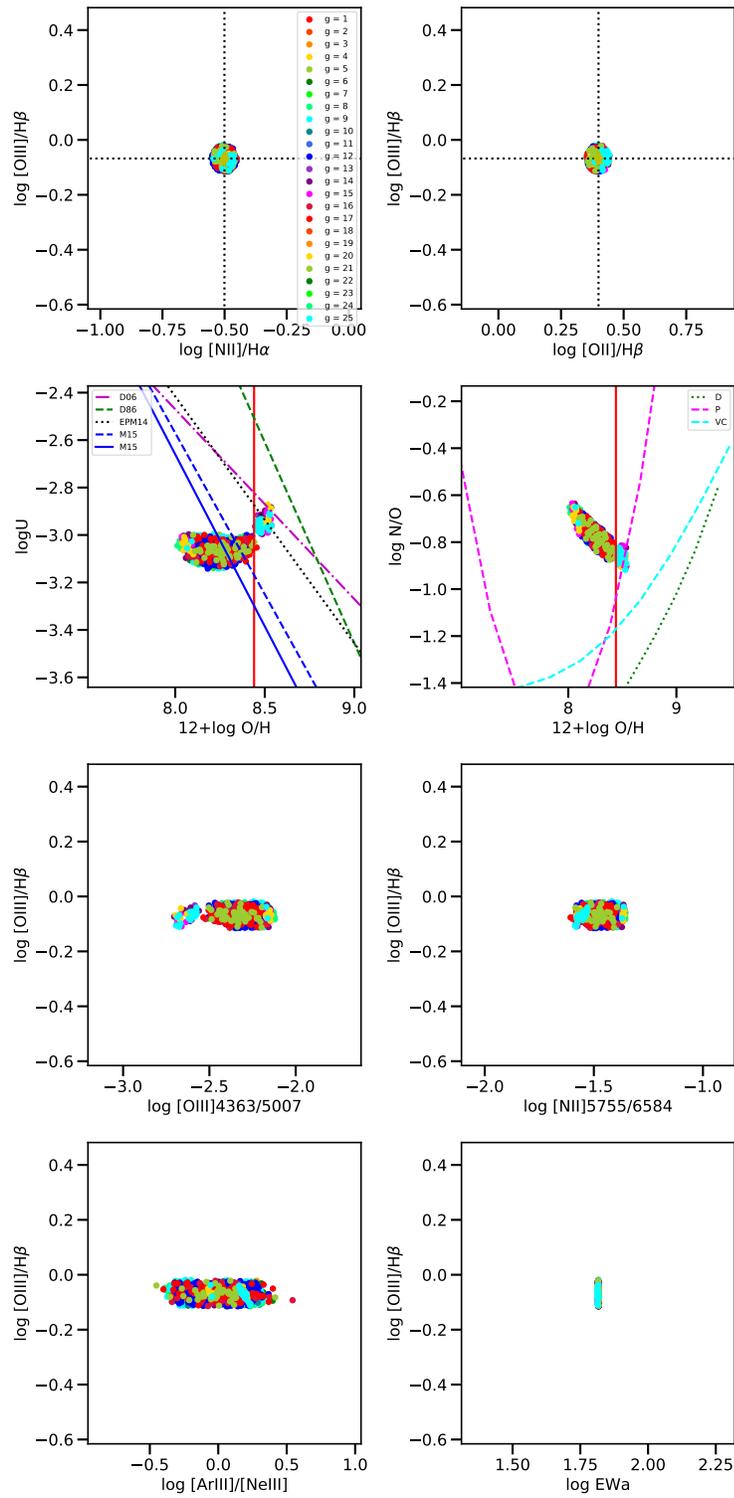


Figura A.4: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 6

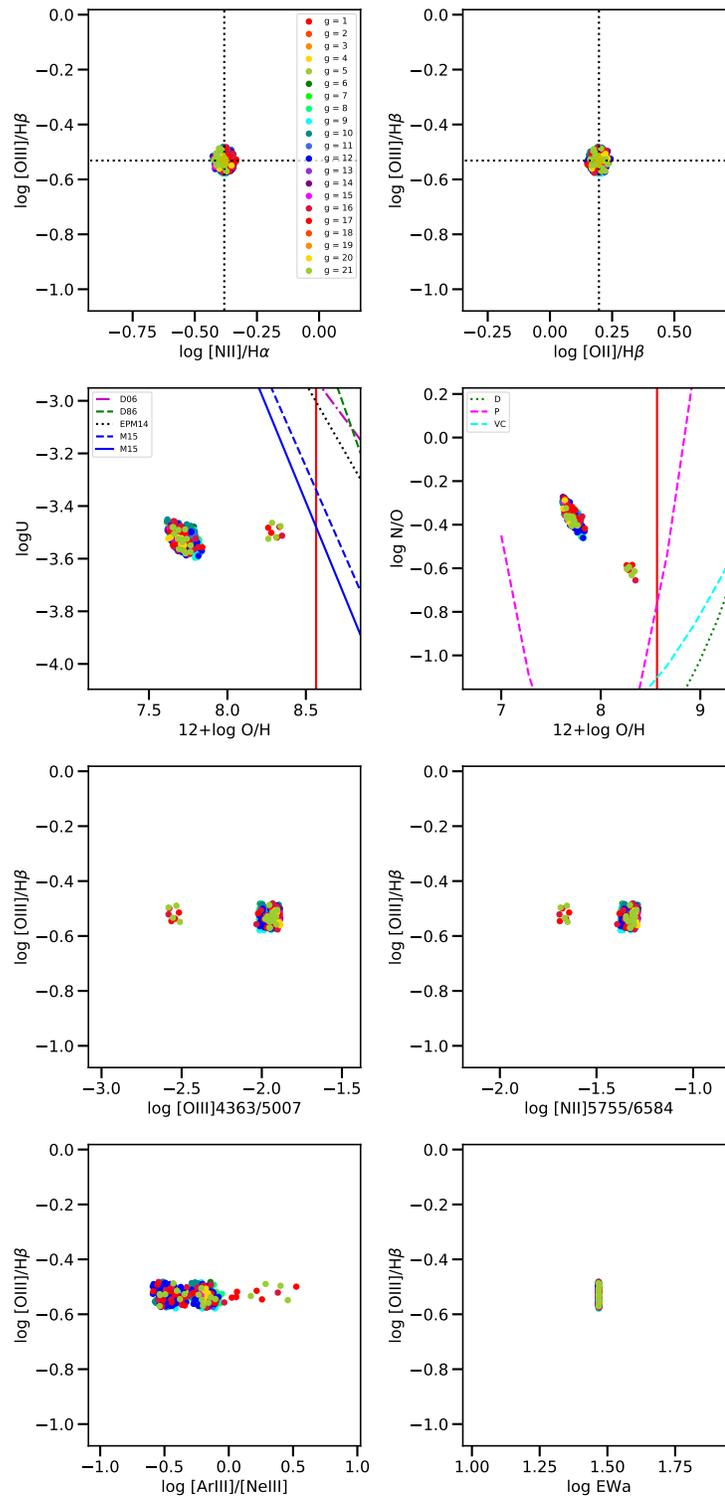


Figura A.5: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 7

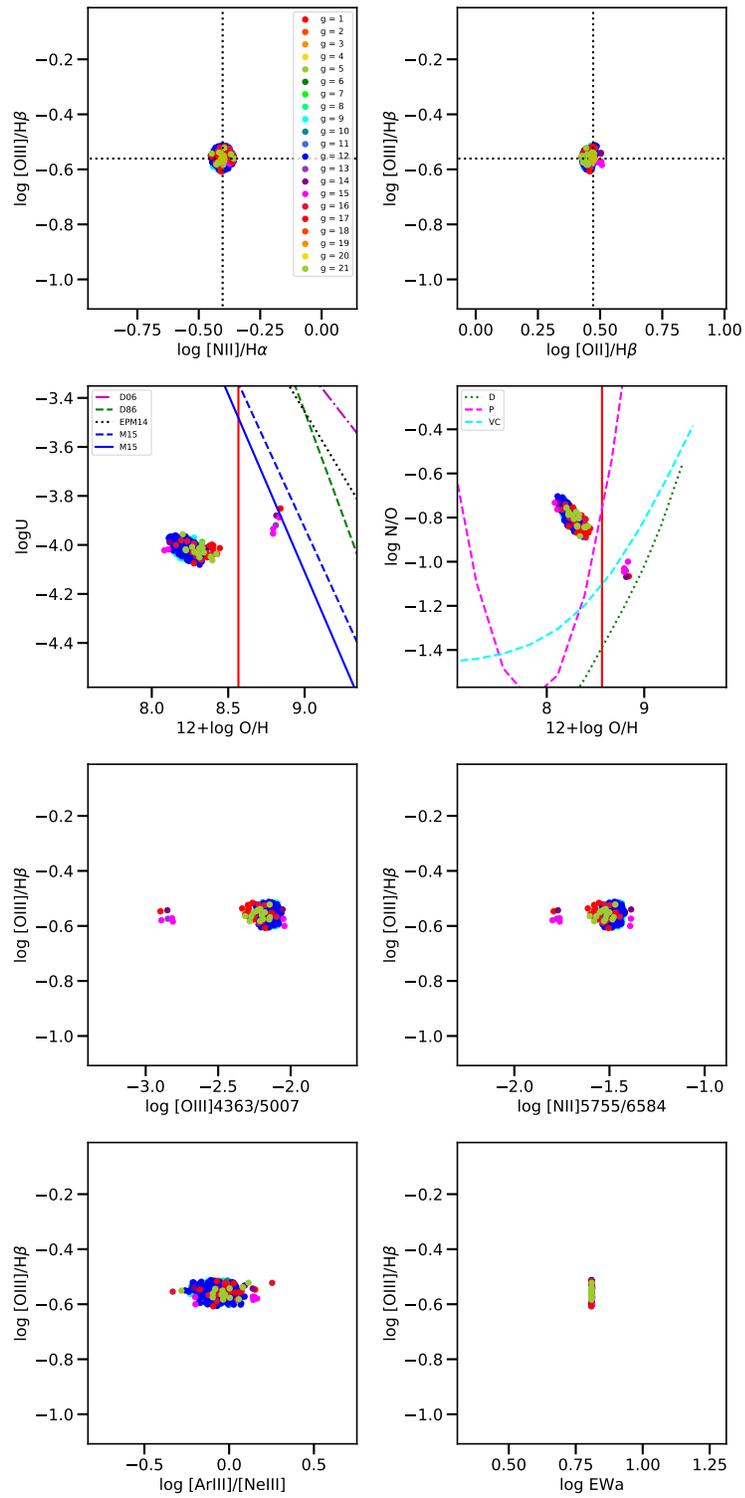


Figura A.6: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 9

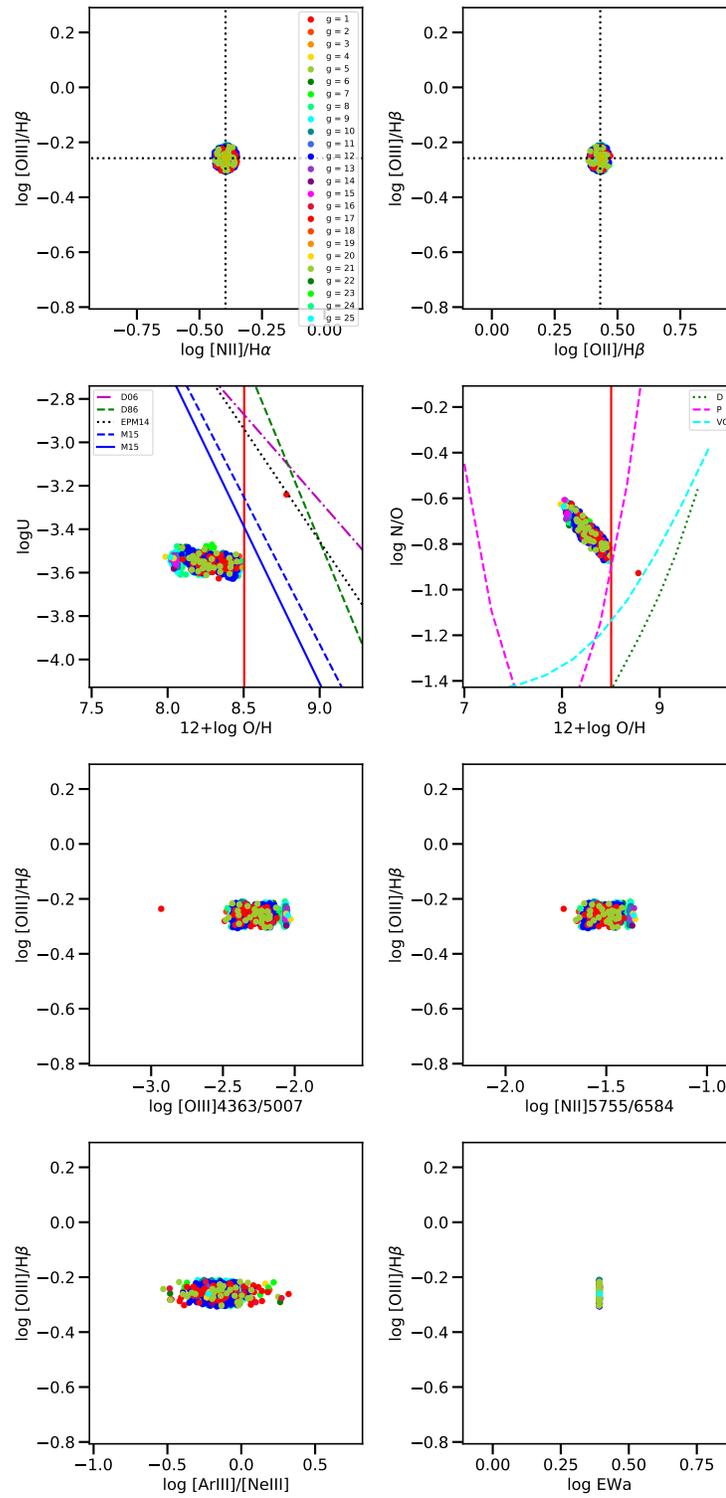


Figura A.7: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 10

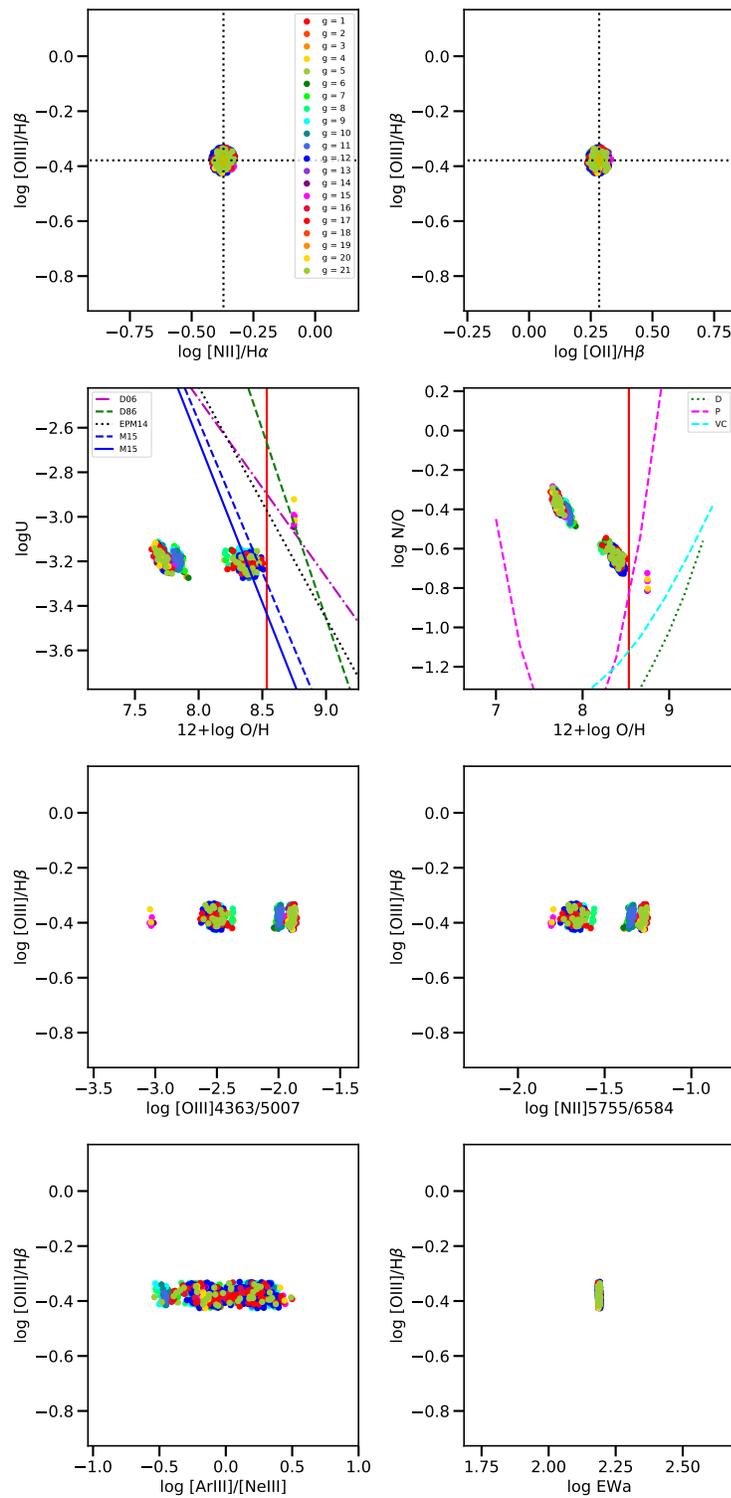


Figura A.8: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 12

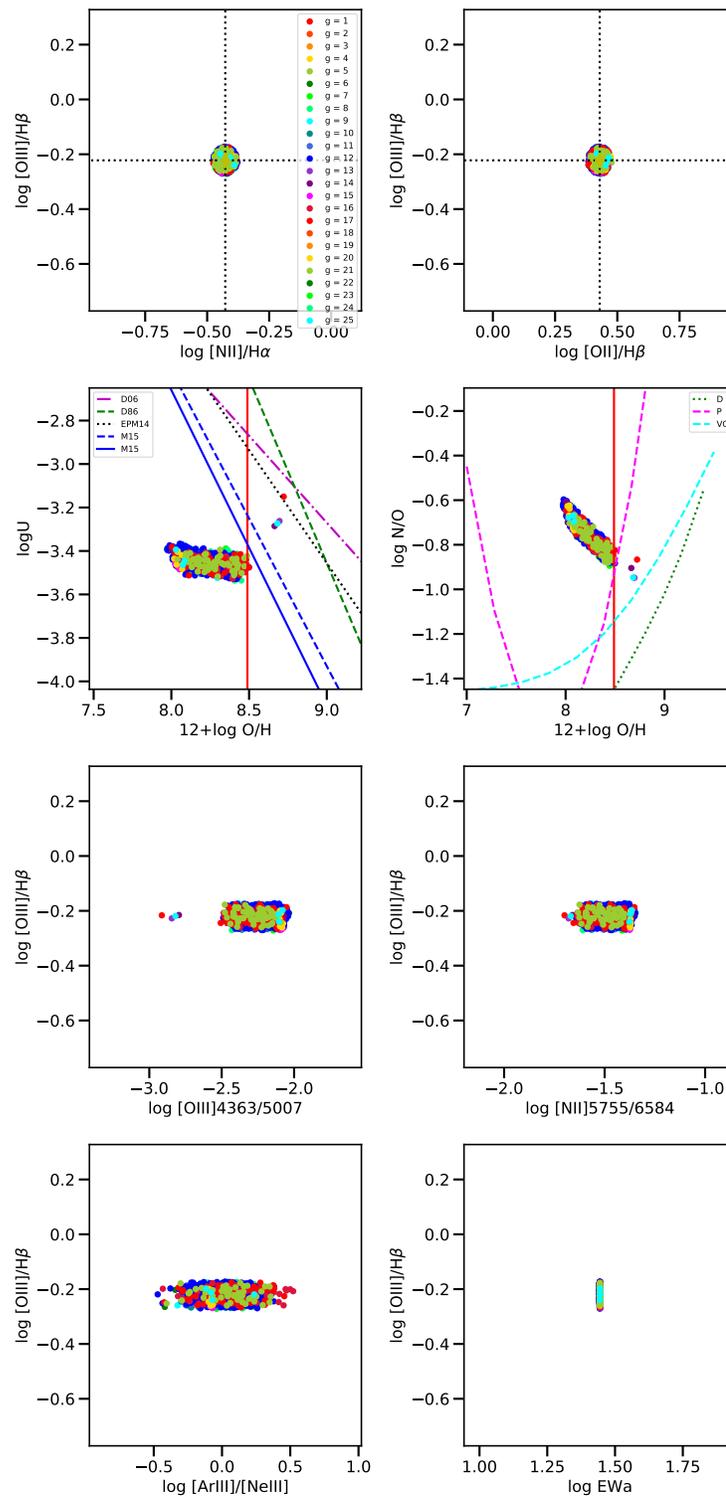


Figura A.9: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 13

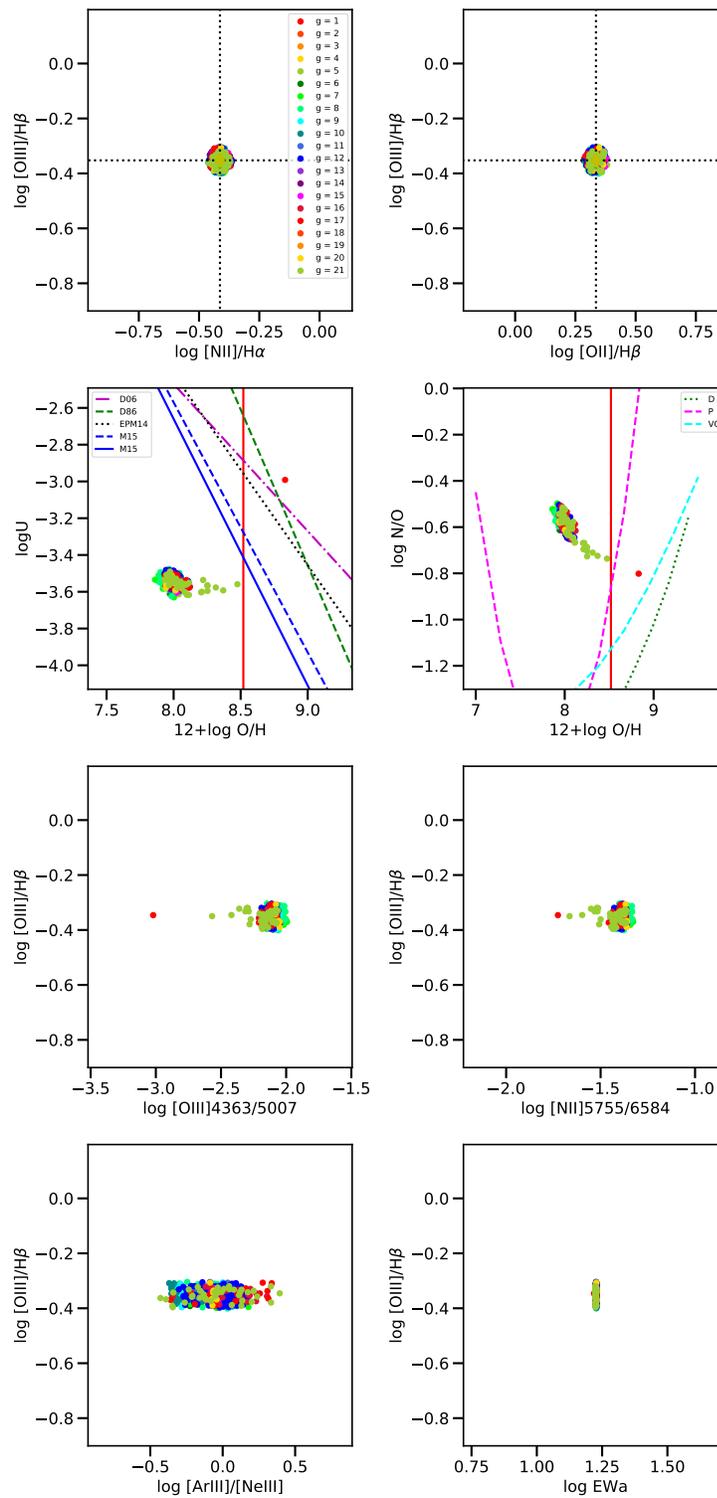


Figura A.10: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 14

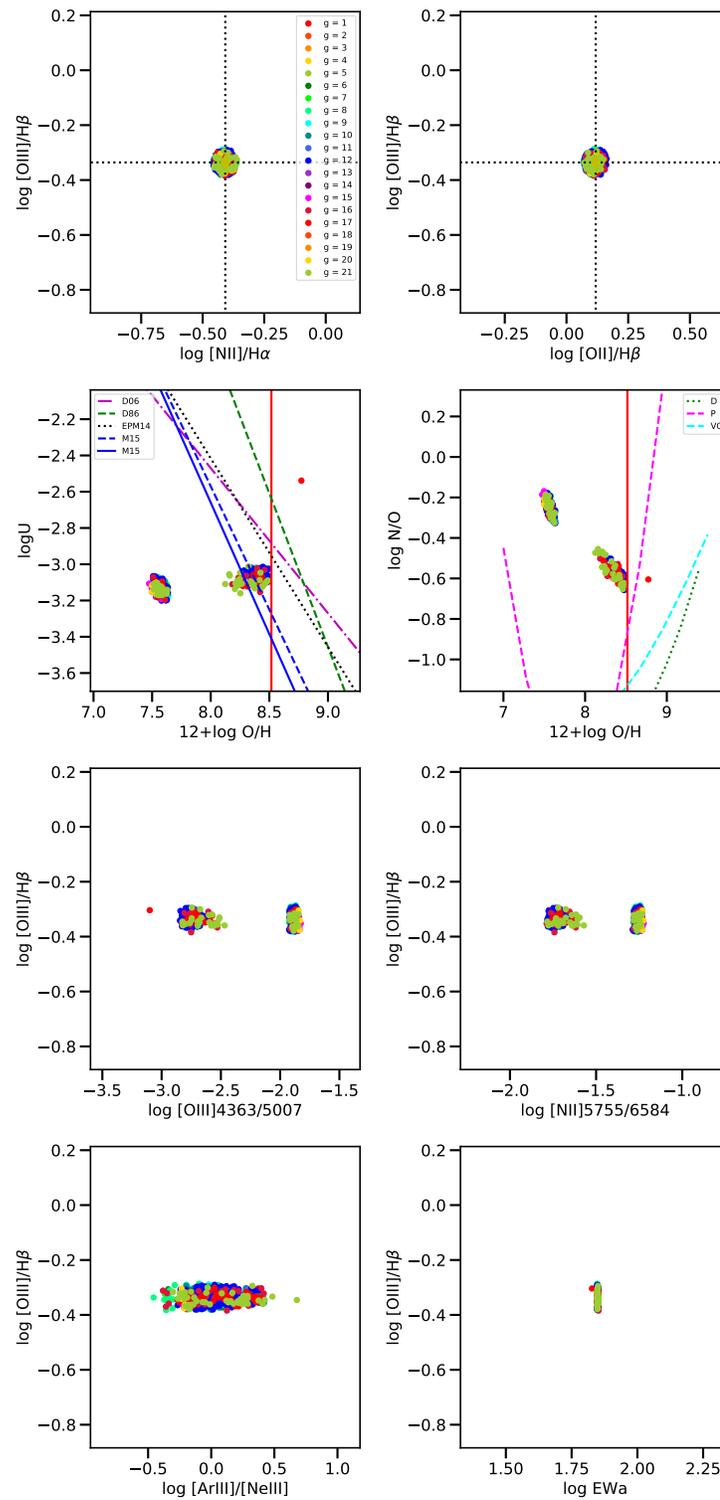


Figura A.11: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 16

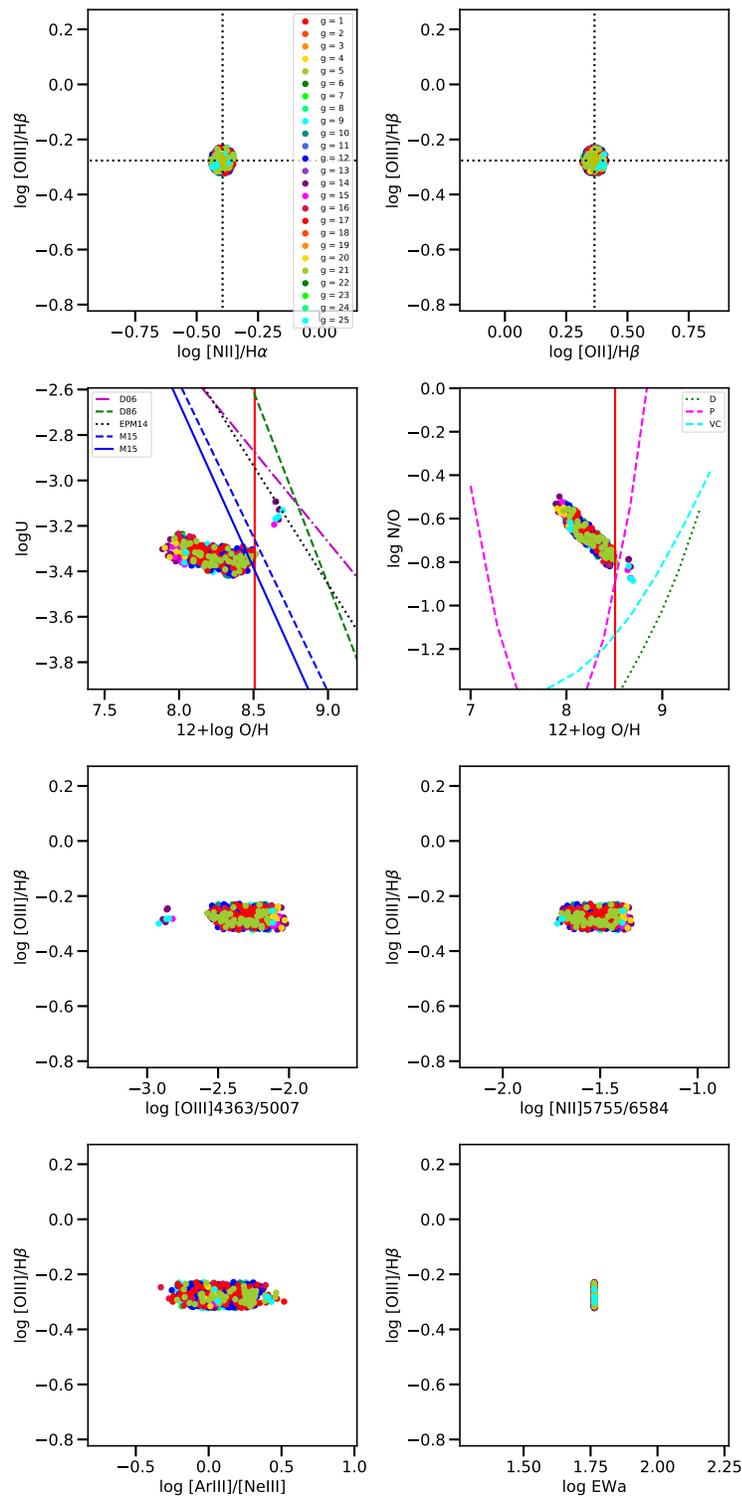


Figura A.12: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 18

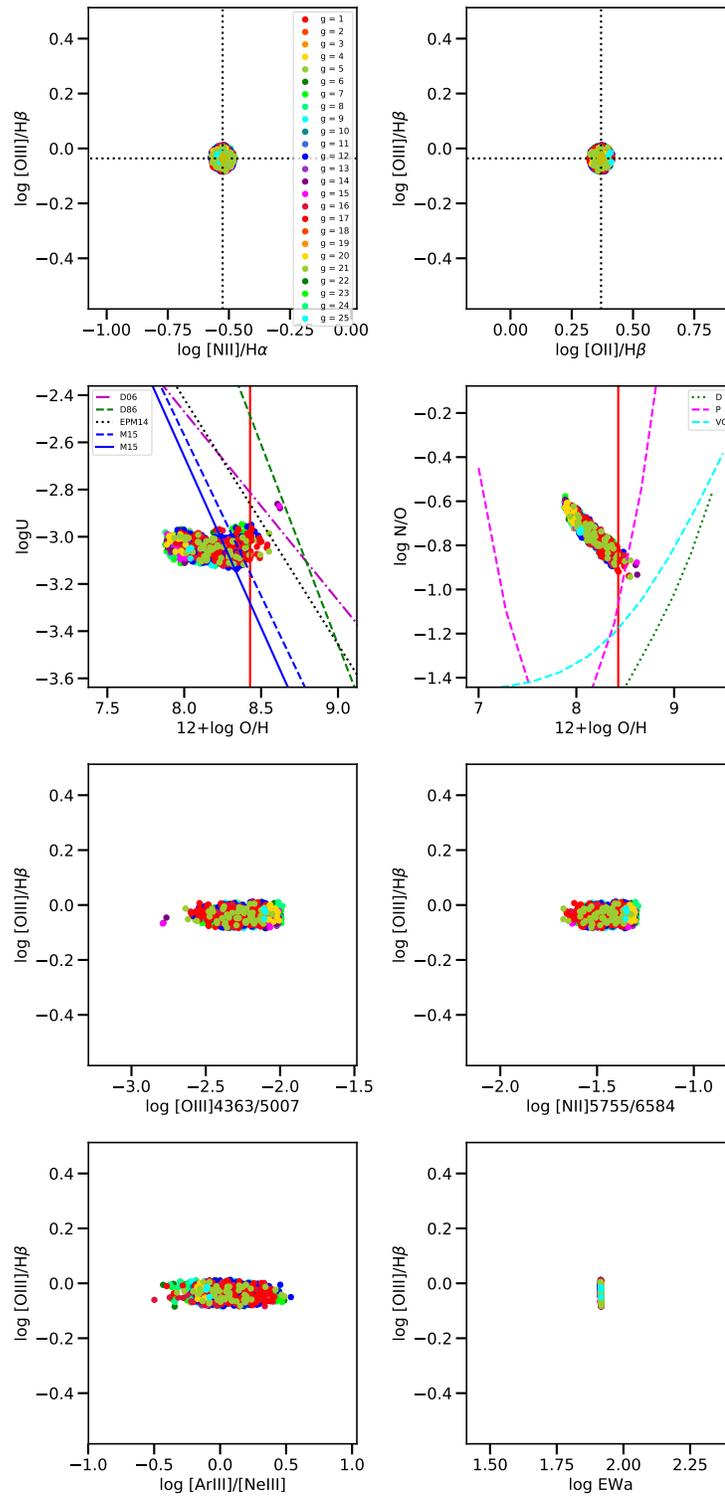


Figura A.13: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 19

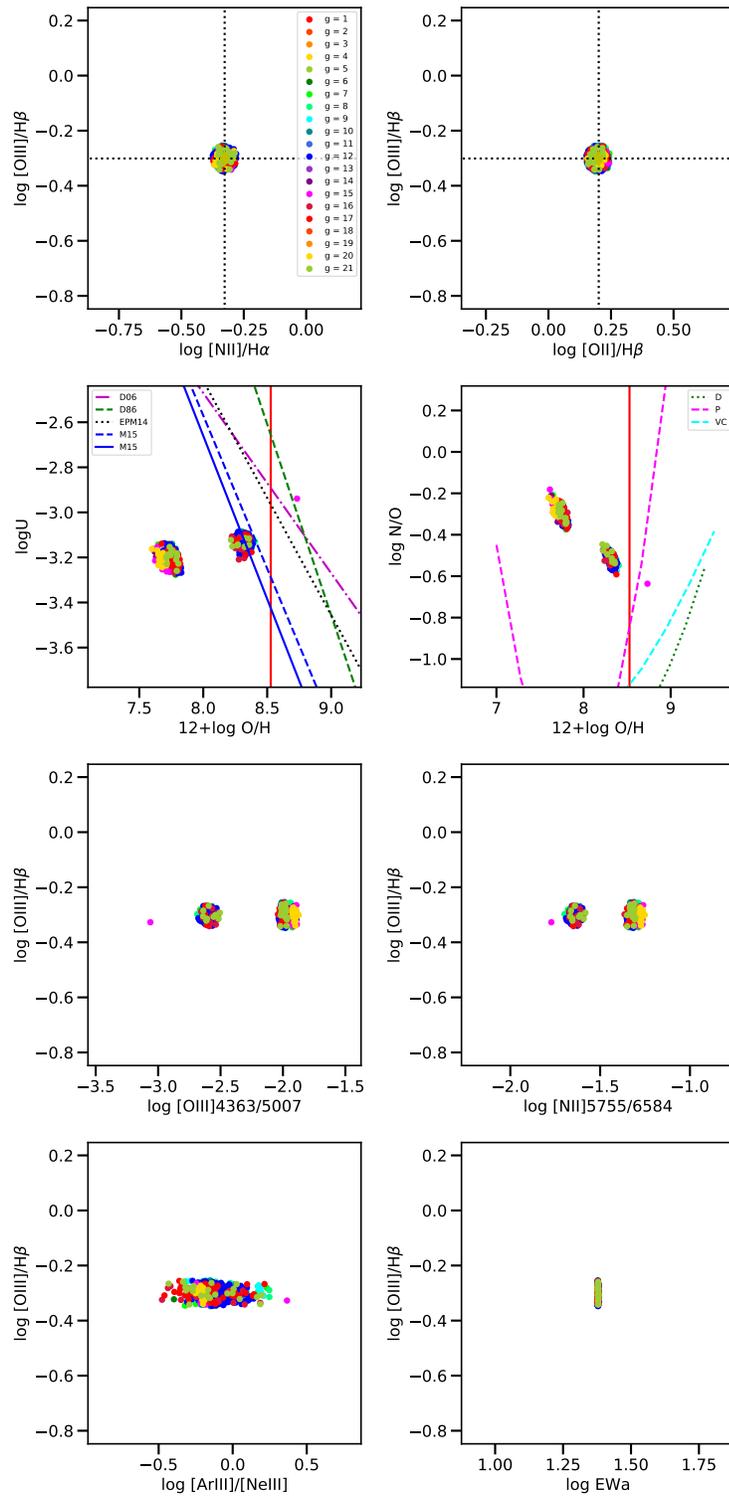


Figura A.14: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 20

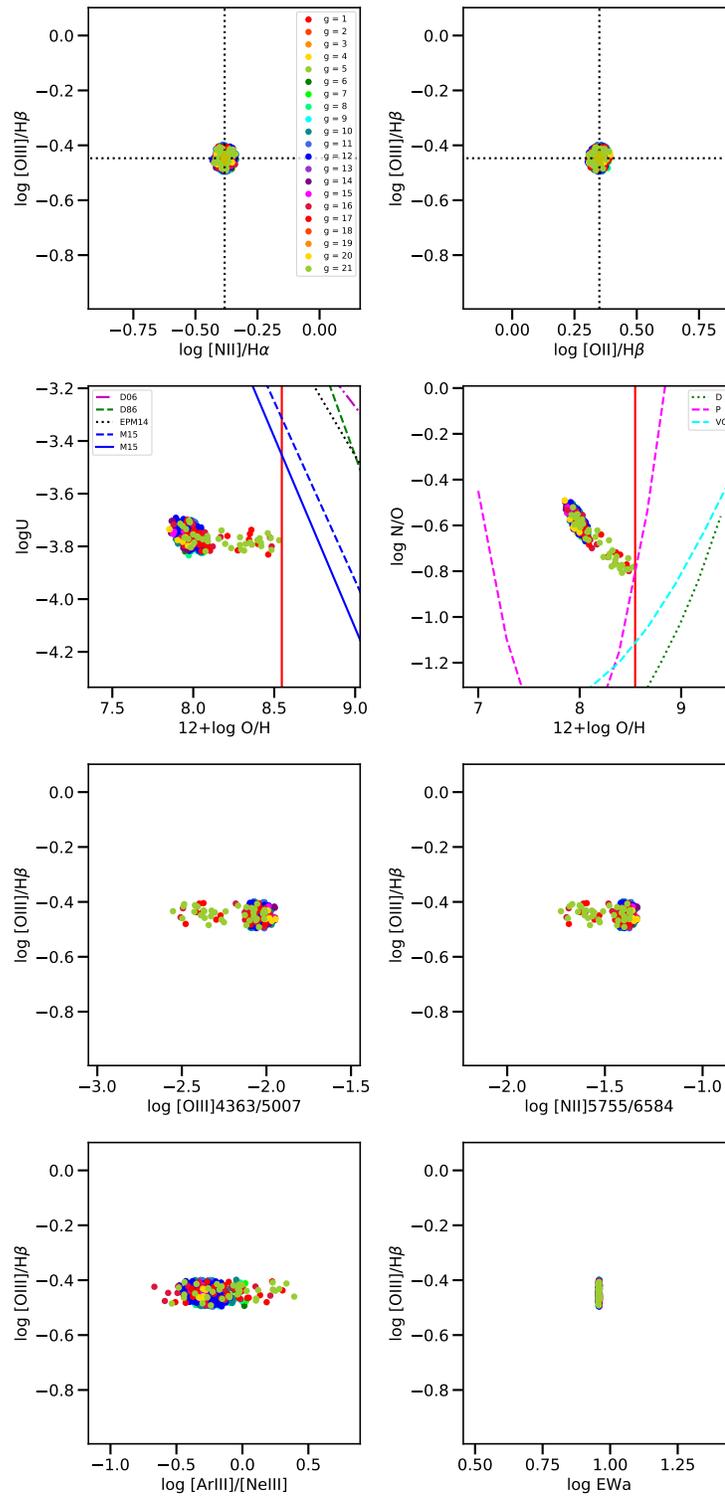


Figura A.15: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 21

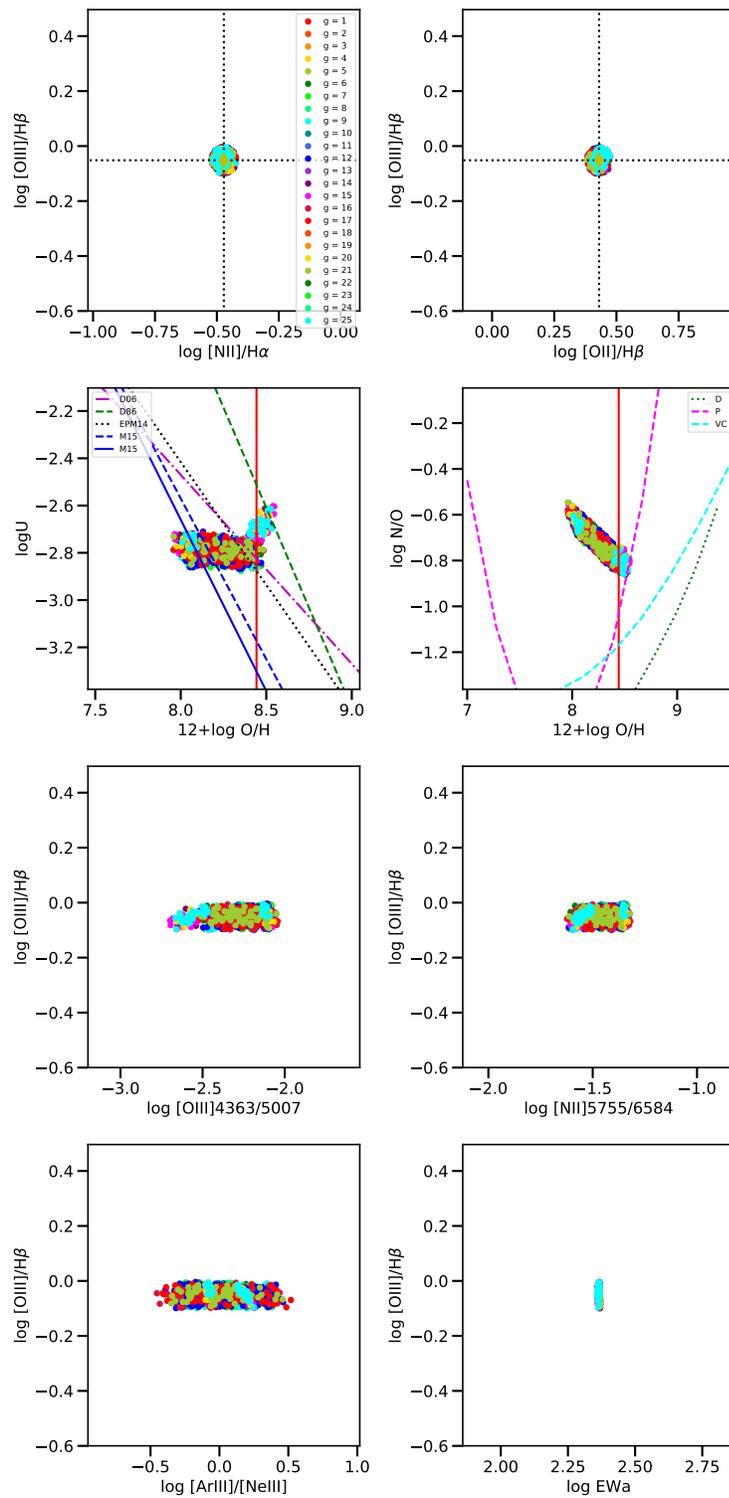


Figura A.16: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 24

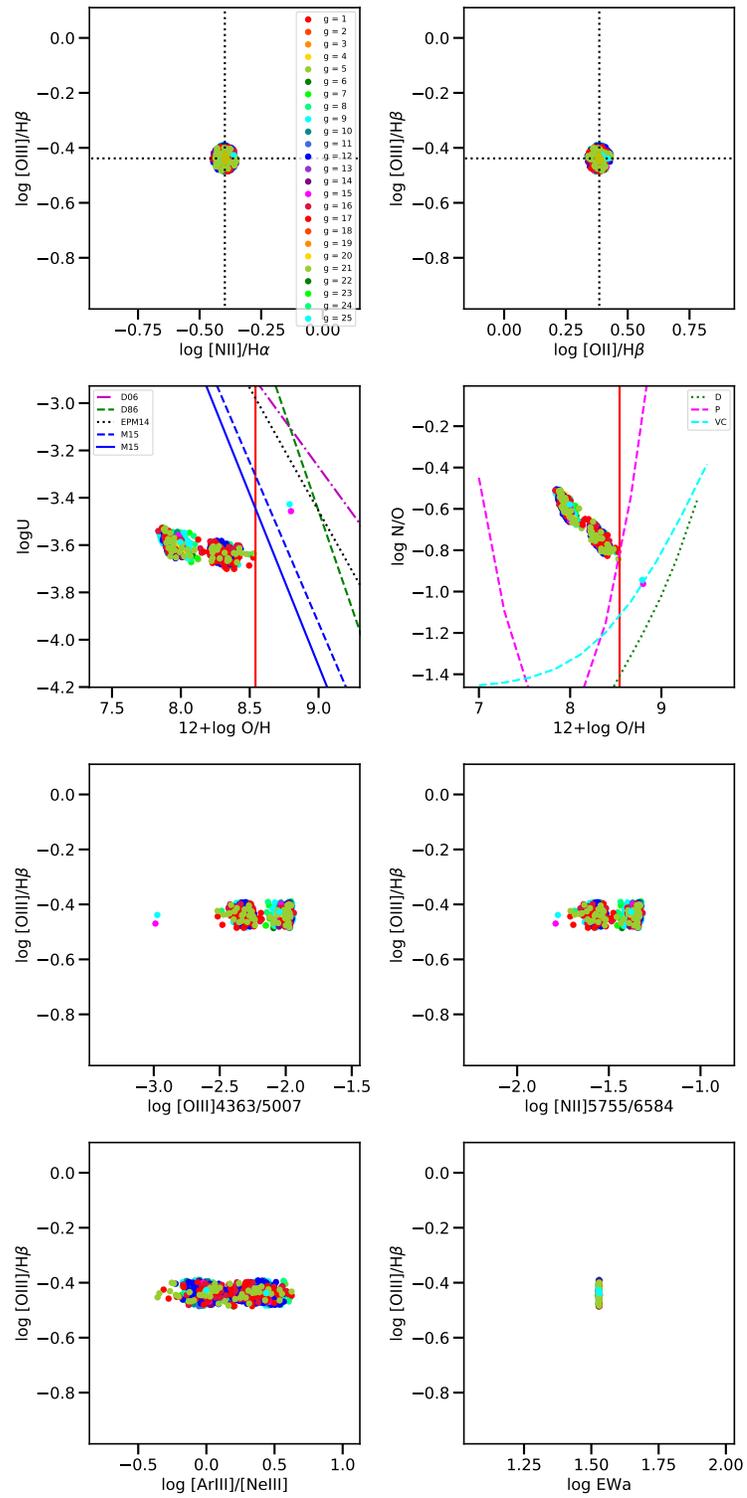


Figura A.17: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 25

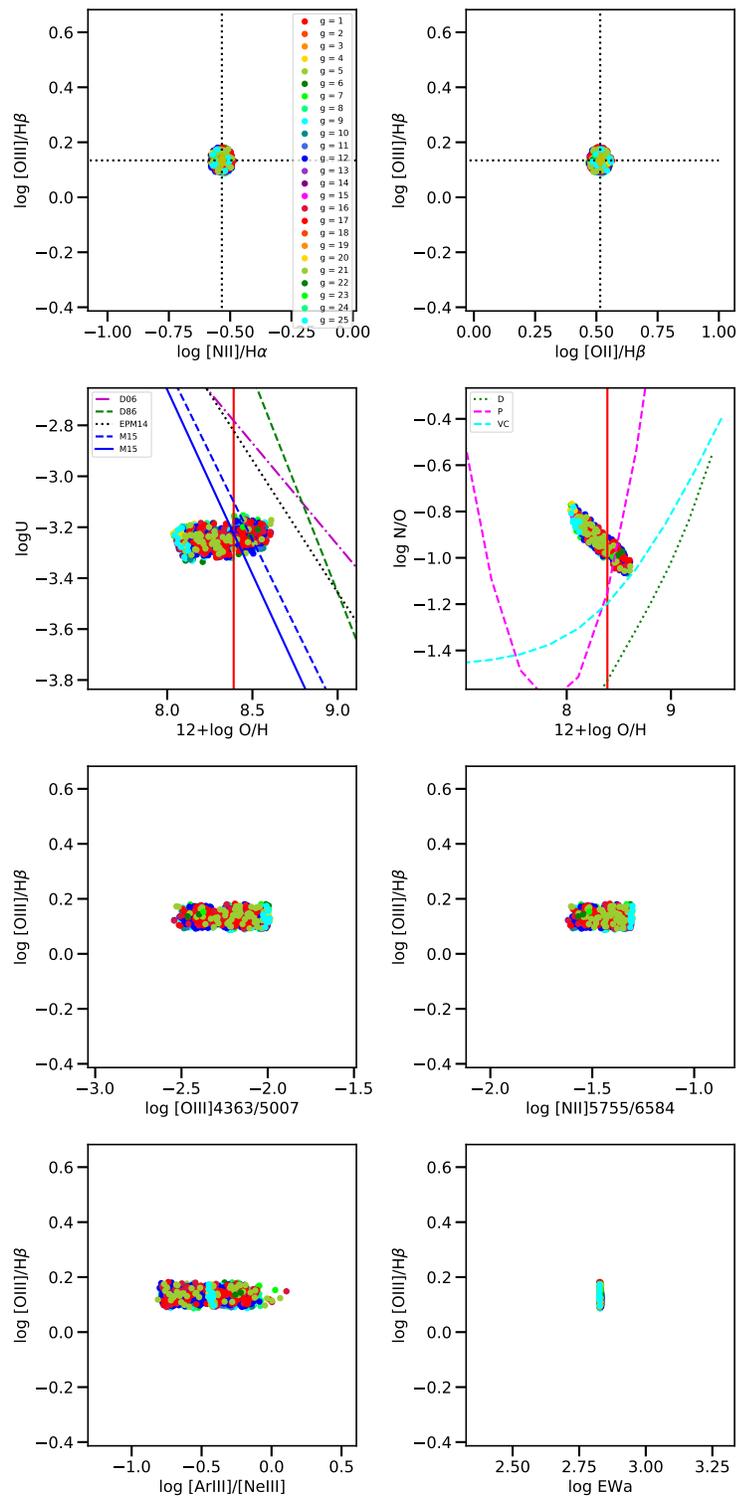


Figura A.18: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 26

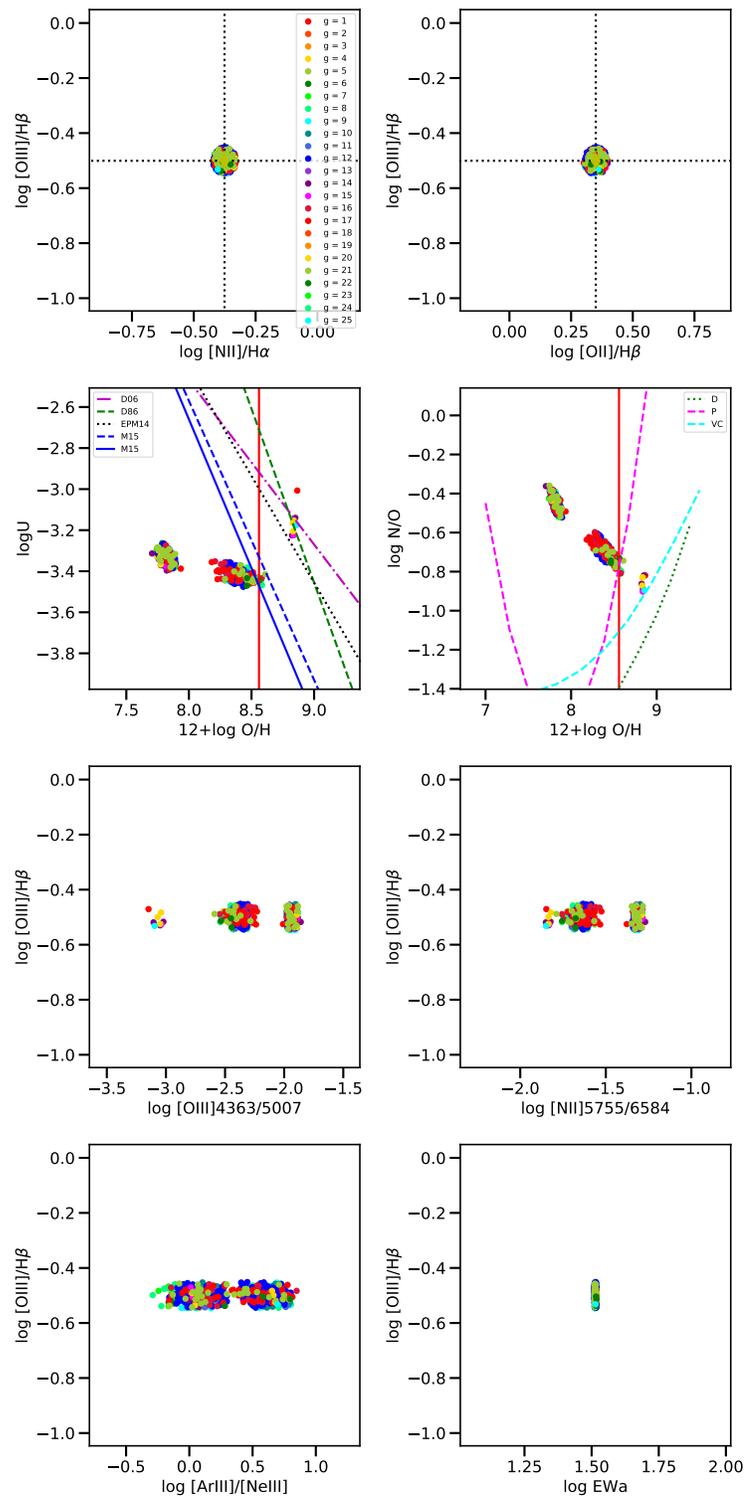


Figura A.19: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 27

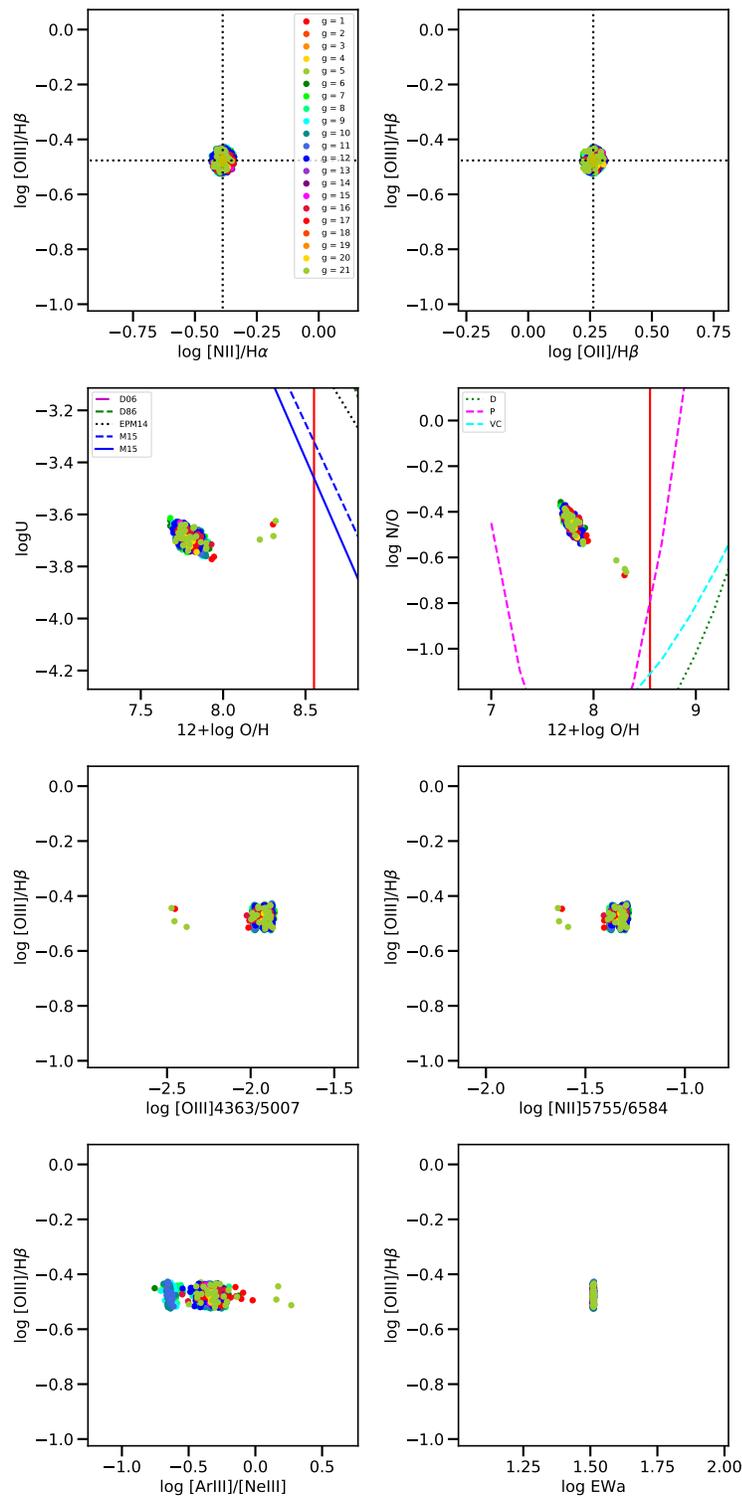


Figura A.20: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 28

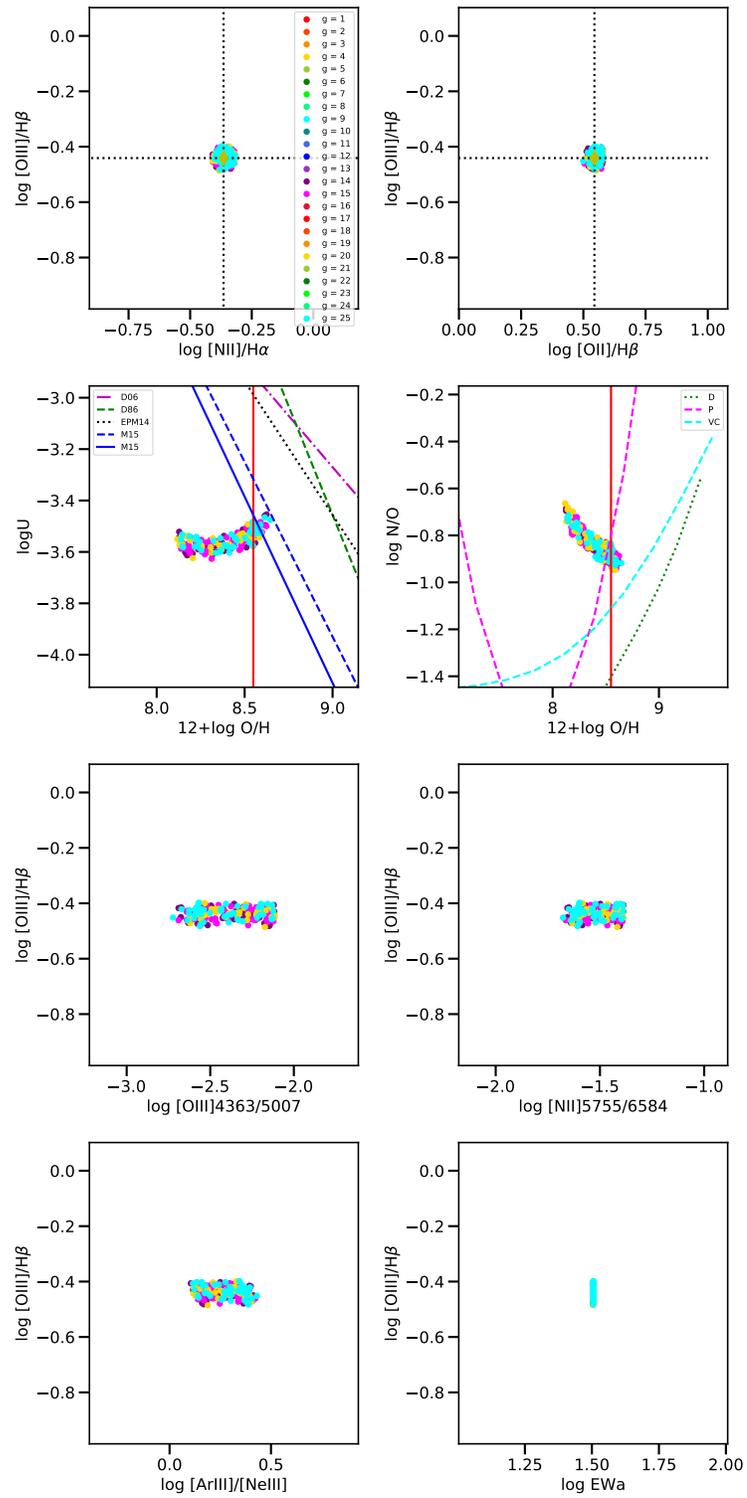


Figura A.21: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 31

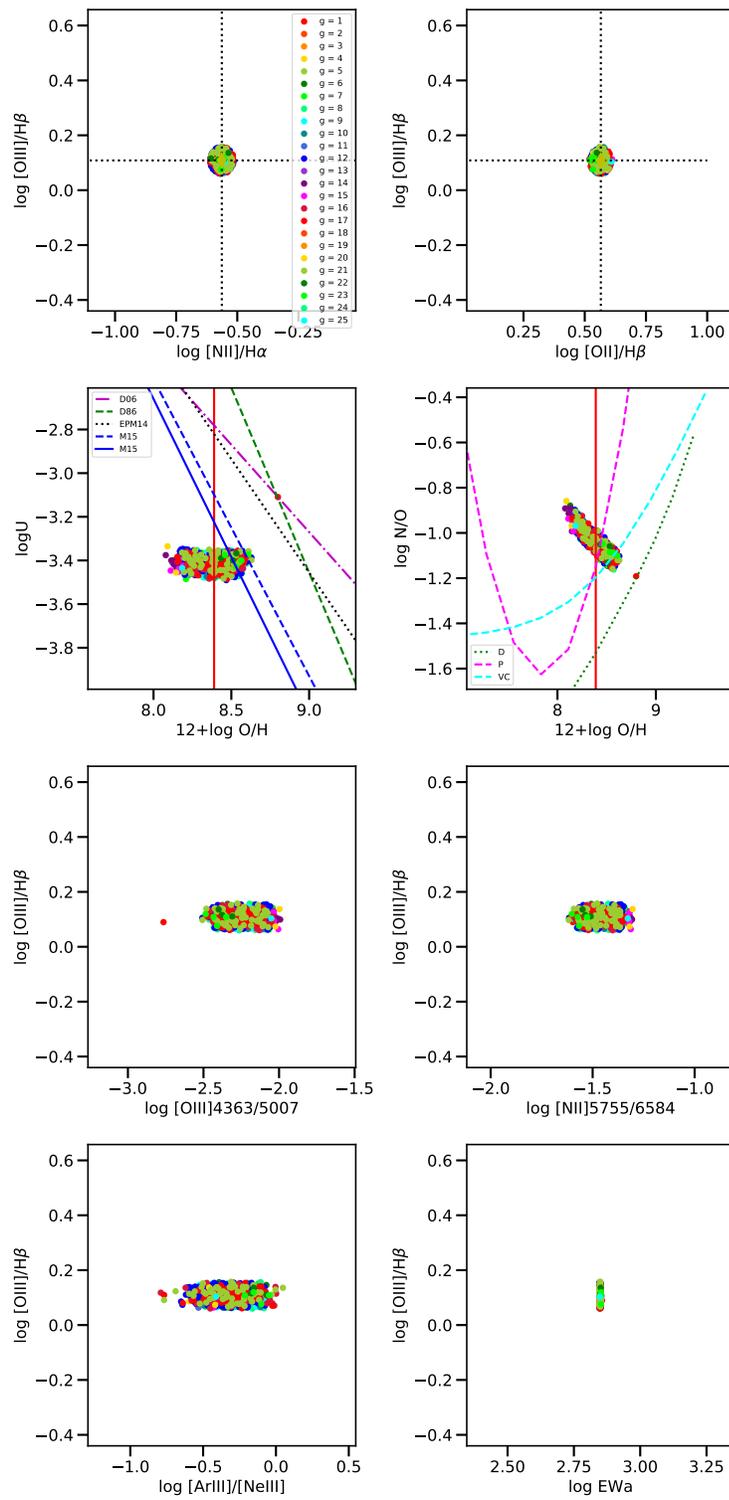


Figura A.22: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 33

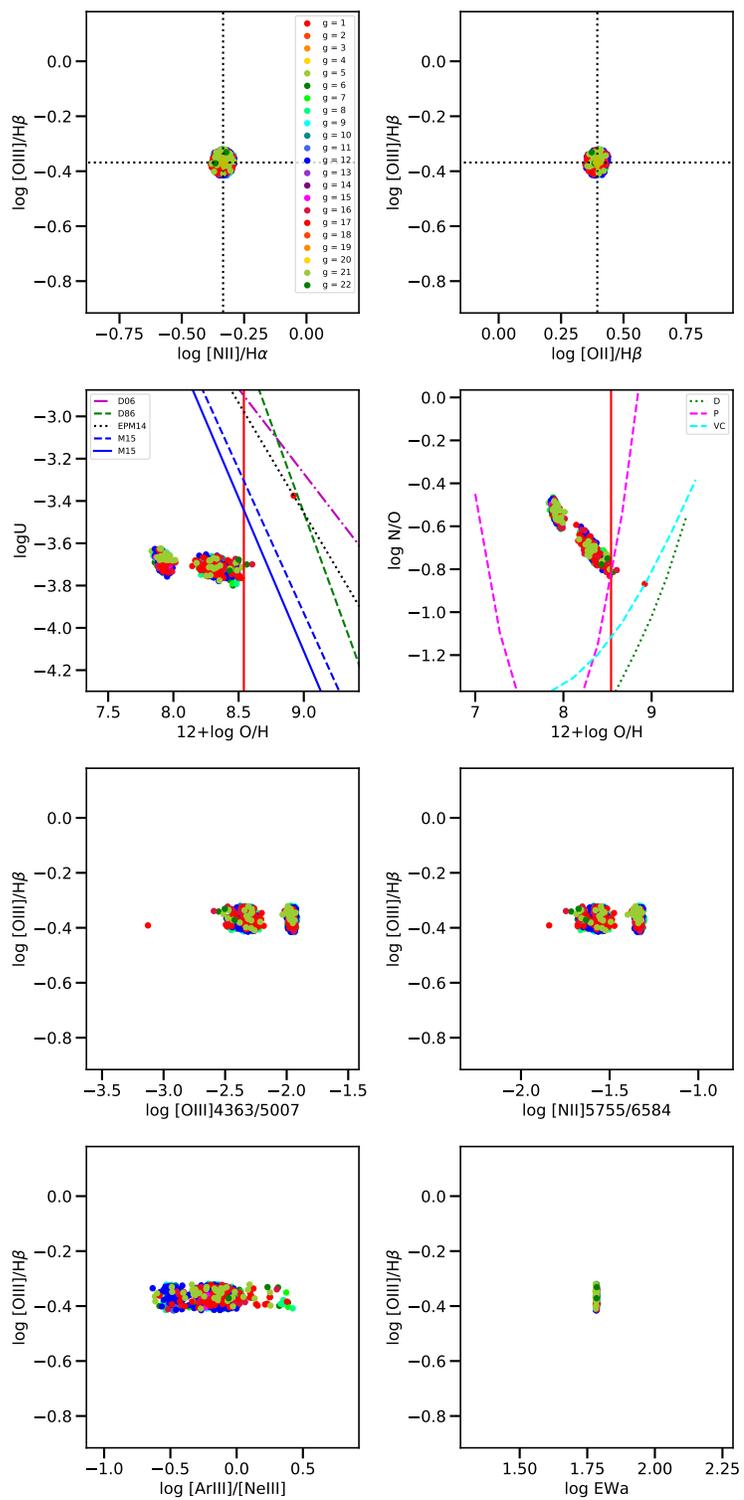


Figura A.23: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 35

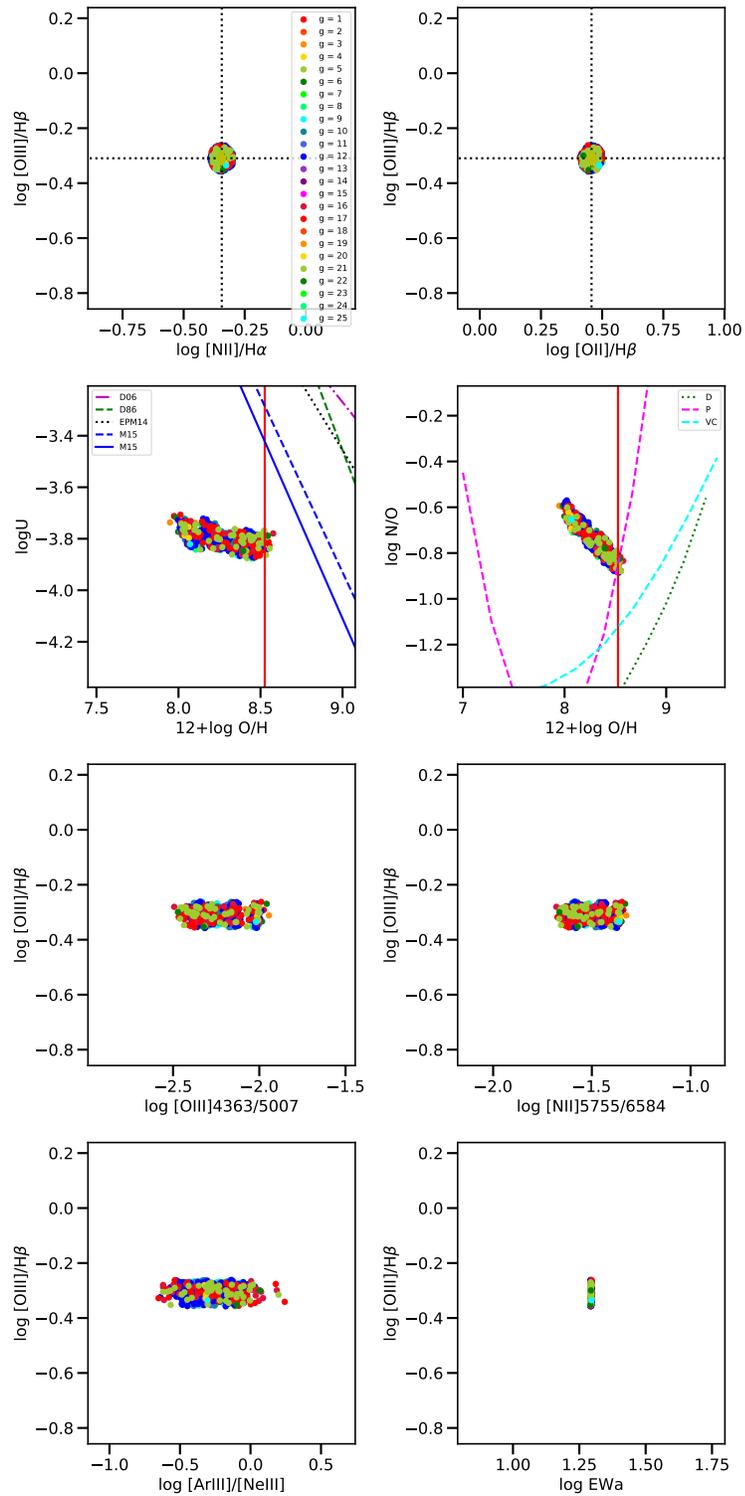


Figura A.24: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 37

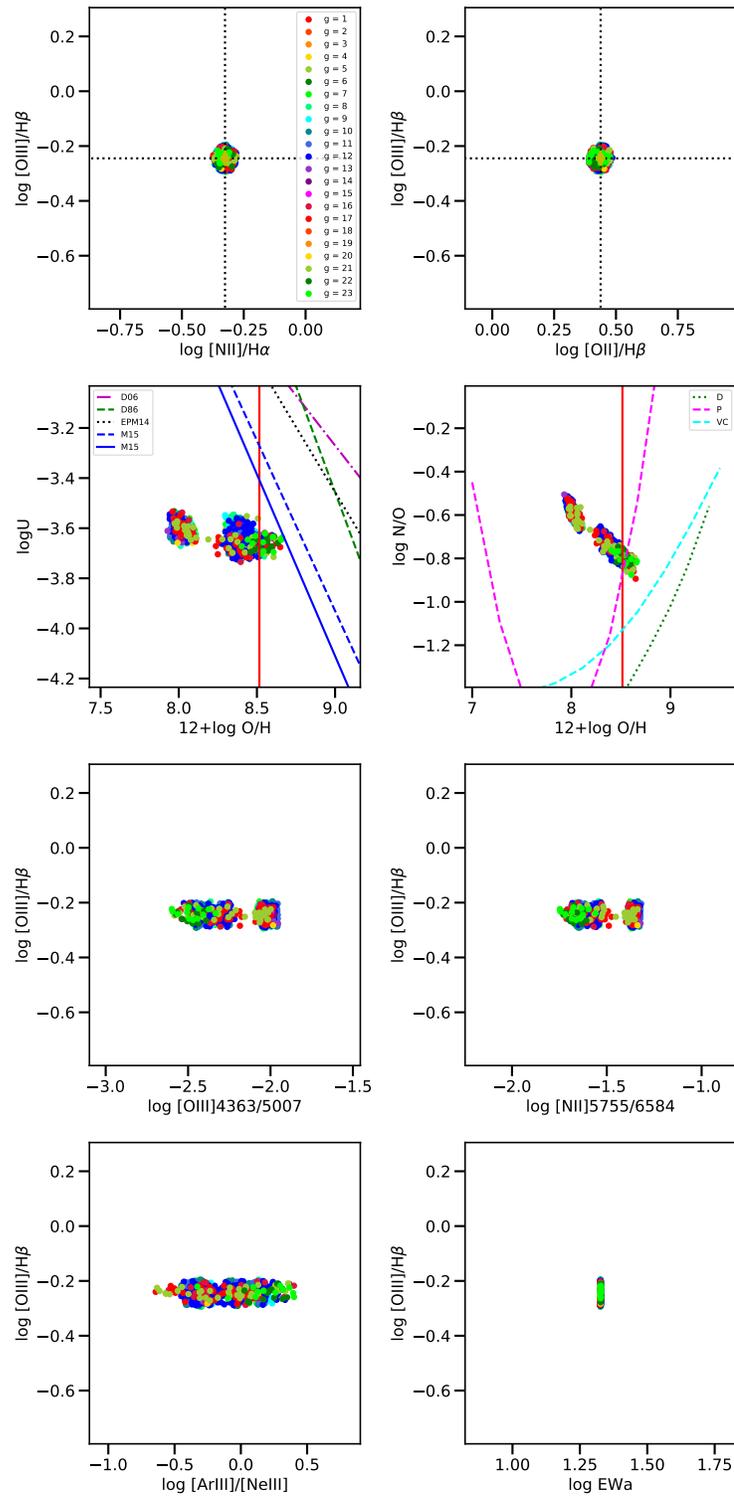


Figura A.25: Modelos finales que difieren de la observación un máximo de 0.05 de la solución de la región 38

Bibliografía

- ASPLUND, M., GREVESSE, N., SAUVAL, A. J. & SCOTT, P., 2009. The Chemical Composition of the Sun. *ARA&A*, **47**, 481–522.
- BASU, B., CHATTOPADHYAY, T. & BISWAS, S., 2010. *An Introduction To Astrophysics*. PHI Learning. ISBN 9788120340718.
- BISBAS, T. G., 2016. *The Interstellar Medium, Expanding Nebulae and Triggered Star Formation: Theory and Simulations*. SpringerBriefs in Astronomy. Springer International Publishing, 1st ed. ISBN 9783319261409,9783319261423.
- BOENDER, C. G. E. & ROMELJN, H. E., 1995. *Stochastic Methods*, 829–869. Springer US, Boston, MA. ISBN 9781461520252.
- BOHIGAS, J., 2008. Photoionization Models Applied to Planetary Nebulae. *ApJ*, **674**, 954–975.
- CALIŃSKI, T. & HARABASZ, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics*, **3**(1), 1–27.
- CANTÓ, J., CURIEL, S. & MARTÍNEZ-GÓMEZ, E., 2009. A simple algorithm for optimization and model fitting: AGA (asexual genetic algorithm). *A&A*, **501**, 1259–1268.
- CHARBONNEAU, P., 1995. Genetic Algorithms in Astronomy and Astrophysics. *ApJS*, **101**, 309.
- CID FERNANDES, R., PÉREZ, E., GARCÍA BENITO, R., GONZÁLEZ DELGADO, R. M., DE AMORIM, A. L., SÁNCHEZ, S. F., HUSEMANN, B., FALCÓN BARROSO, J., SÁNCHEZ-BLÁZQUEZ, P., WALCHER, C. J. & MAST, D., 2013. Resolving galaxies in time and space. I. Applying STARLIGHT to CALIFA datacubes. *A&A*, **557**, A86.

- COMANICIU, D.; MEER, P., 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**.
- DE AMORIM, R. C. & HENNIG, C., 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, **324**, 126 – 145. ISSN 0020-0255.
- DOPITA, M. A. & EVANS, I. N., 1986. Theoretical models for H II regions. II - The extragalactic H II region abundance sequence. *ApJ*, **307**, 431–440.
- DOPITA, M. A., FISCHERA, J., SUTHERLAND, R. S., KEWLEY, L. J., TUFFS, R. J., POPESCU, C. C., VAN BREUGEL, W., GROVES, B. A. & LEITHERER, C., 2006. Modeling the Pan-Spectral Energy Distribution of Starburst Galaxies. II. Control of the H II Region Parameters. *ApJ*, **647**, 244–255.
- DOPITA, M. A. & SUTHERLAND, R. S., 2003. *Astrophysics of the diffuse universe*.
- DOPITA, M. A., SUTHERLAND, R. S., NICHOLLS, D. C., KEWLEY, L. J. & VOGT, F. P. A., 2013. New Strong-line Abundance Diagnostics for H II Regions: Effects of κ -distributed Electron Energies and New Atomic Data. *ApJS*, **208**, 10.
- DORS, JR., O. L., ARELLANO-CÓRDOVA, K. Z., CARDACI, M. V. & HÄGELE, G. F., 2017. New quantitative nitrogen abundance estimations in a sample of Seyfert 2 active galactic nuclei. *MNRAS*, **468**, L113–L117.
- DRABAS, T., 2016. *Practical Data Analysis Cookbook*. Packt Publishing. ISBN 9781783558513.
- DRAINE, B., 2010. *Physics of the Interstellar and Intergalactic Medium*. Princeton Series in Astrophysics. Princeton University Press. ISBN 9781400839087.
- DRAINE, B. T., 2011. On Radiation Pressure in Static, Dusty H II Regions. *ApJ*, **732**, 100.
- DYSON, J. & WILLIAMS, D., 1997. *The Physics of the Interstellar Medium, Second Edition*. Series in Astronomy and Astrophysics. CRC Press. ISBN 9780585368115.
- EPANECHNIKOV, V. A., 1969. Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability & Its Applications*, **14**.

- ERCOLANO, B., BARLOW, M. J., STOREY, P. J. & LIU, X.-W., 2003. MOCASSIN: a fully three-dimensional Monte Carlo photoionization code. *MNRAS*, **340**, 1136–1152.
- ESTER, M., KRIEGEL, H.-P., SANDER, J. & XU, X., 1996. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231. AAAI Press.
- FERLAND, G. J., HENNEY, W. J., O'DELL, C. R. & PEIMBERT, M., 2016. The abundance discrepancy factor and t^2 in nebulae: are non-thermal electrons the culprits? *Rev. Mexicana Astron. Astrofis.*, **52**, 261–269.
- FERLAND, G. J., PORTER, R. L., VAN HOOF, P. A. M., WILLIAMS, R. J. R., ABEL, N. P., LYKINS, M. L., SHAW, G., HENNEY, W. J. & STANCIL, P. C., 2013. The 2013 Release of Cloudy. *Rev. Mexicana Astron. Astrofis.*, **49**, 137–163.
- FLORES-FAJARDO, N., MORISSET, C., STASIŃSKA, G. & BINETTE, L., 2011. Ionization of the diffuse gas in galaxies: hot low-mass evolved stars at work. *MNRAS*, **415**, 2182–2192.
- FUKUNAGA, K.; HOSTETLER, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, **21**.
- GARCÍA-BENITO, R., PÉREZ, E., DÍAZ, Á. I., MAÍZ APELLÁNIZ, J. & CERVIÑO, M., 2011. The 100 Myr Star Formation History of NGC 5471 from Cluster and Resolved Stellar Photometry. *AJ*, **141**, 126.
- GESICKI, K., ZIJLSTRA, A. A. & MORISSET, C., 2016. 3D pyCloudy modelling of bipolar planetary nebulae: Evidence for fast fading of the lobes. *A&A*, **585**, A69.
- GONZÁLEZ DELGADO, R. M. & PÉREZ, E., 1997. H II Region Population in a Sample of Nearby Galaxies with Nuclear Activity. II. Luminosity Function, Size, and Radial Distributions. *ApJS*, **108**, 199–228.
- HARTIGAN, J., 1975. *Clustering algorithms*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. ISBN 9780471356455.
- HORST, R. & TUY, H., 1996. *Global Optimization: Deterministic Approaches*. Springer-Verlag Berlin Heidelberg, 3rd ed. ISBN 978-3-642-08247-4, 978-3-662-03199-5.

- HUMMER, D. G. & SEATON, M. J., 1963. The ionization structure of planetary nebulae, I. Pure hydrogen nebulae. *MNRAS*, **125**, 437.
- HYNDMAN, R. J. & FAN, Y., 1996. Sample Quantiles in Statistical Packages. *The American Statistician*, **50**(4), 361–365. ISSN 00031305.
- KALLMAN, T. R., 2010. Modeling of Photoionized Plasmas. *Space Sci. Rev.*, **157**, 177–191.
- KAUFMAN, L. & ROUSSEEUW, P., 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780470317488.
- KELZ, A., VERHEIJEN, M. A. W., ROTH, M. M., BAUER, S. M., BECKER, T., PASCHKE, J., POPOW, E., SÁNCHEZ, S. F. & LAUX, U., 2006. PMAS: The Potsdam Multi-Aperture Spectrophotometer. II. The Wide Integral Field Unit PPak. *PASP*, **118**, 129–145.
- KWOK, S., 2007. *Physics and Chemistry of the Interstellar Medium*. University Science Books. ISBN 9781891389467.
- LEIGHLY, K. M. & CASEBEER, D., 2007. Photoionization Models of the Broad-line Region. In L. C. Ho & J.-W. Wang, eds., *The Central Engine of Active Galactic Nuclei*, vol. 373 of *Astronomical Society of the Pacific Conference Series*, 365.
- LETOKHOV, V. & JOHANSSON, S., 2008. *Astrophysical Lasers*. OUP Oxford. ISBN 9780191563355.
- LÓPEZ, L. A., KRUMHOLZ, M. R., BOLATTO, A. D., PROCHASKA, J. X. & RAMIREZ-RUIZ, E., 2011. What Drives the Expansion of Giant H II Regions?: A Study of Stellar Feedback in 30 Doradus. *ApJ*, **731**, 91.
- MACIEL, W. J., 2013. *Astrophysics of the Interstellar Medium*. Springer-Verlag New York, 1st ed. ISBN 978-1-4614-3766-6, 978-1-4614-3767-3.
- MARINO, R. A., ROSALES-ORTEGA, F. F., SÁNCHEZ, S. F., GIL DE PAZ, A., VÍLCHEZ, J. ET AL., 2013. The O3N2 and N2 abundance indicators revisited: improved calibrations based on CALIFA and T_e -based literature data. *A&A*, **559**, A114.
- MCGAUGH, S. S., 1991. H II region abundances - Model oxygen line ratios. *ApJ*, **380**, 140–150.

- MOLLÁ, M., GARCÍA-VARGAS, M. L. & BRESSAN, A., 2009. PopStar I: evolutionary synthesis model description. *MNRAS*, **398**, 451–470.
- MORISSET, C., 2013. pyCloudy: Tools to manage astronomical Cloudy photoionization code. Astrophysics Source Code Library.
- MORISSET, C., 2014. PyCloudy: a new tool to model 3D nebulae. In *Asymmetrical Planetary Nebulae VI Conference*, 62.
- MORISSET, C., DELGADO-INGLADA, G. & FLORES-FAJARDO, N., 2015. A virtual observatory for photoionized nebulae: the Mexican Million Models database (3MdB). *Rev. Mexicana Astron. Astrofis.*, **51**, 103–120.
- MORISSET, C., DELGADO-INGLADA, G., SÁNCHEZ, S. F., GALBANY, L., GARCÍA-BENITO, R., HUSEMANN, B., MARINO, R. A., MAST, D. & ROTH, M. M., 2016. Photoionization models of the CALIFA H II regions. I. Hybrid models. *A&A*, **594**, A37.
- NAGAO, T., MAIOLINO, R. & MARCONI, A., 2006. Gas metallicity diagnostics in star-forming galaxies. *A&A*, **459**, 85–101.
- OSTERBROCK, D. & FERLAND, G., 2006. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*. University Science Books. ISBN 9781891389344.
- PAGEL, B. E. J., EDMUNDS, M. G., BLACKWELL, D. E., CHUN, M. S. & SMITH, G., 1979. On the composition of H II regions in southern galaxies. I - NGC 300 and 1365. *MNRAS*, **189**, 95–113.
- PATEL, P., SIGUT, T. A. A. & LANDSTREET, J. D., 2016. The Inner Gaseous Disk Around the Herbig Be Star BD+65 1637. In T. A. A. Sigut & C. E. Jones, eds., *Bright Emissaries: Be Stars as Messengers of Star-Disk Physics*, vol. 506 of *Astronomical Society of the Pacific Conference Series*, 291.
- PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

- PÉREZ-MONTERO, E., 2014. Deriving model-based T_e -consistent chemical abundances in ionized gaseous nebulae. *MNRAS*, **441**, 2663–2675.
- PÉREZ-MONTERO, E. & CONTINI, T., 2009. The impact of the nitrogen-to-oxygen ratio on ionized nebula diagnostics based on [NII] emission lines. *MNRAS*, **398**, 949–960.
- PETTINI, M. & PAGEL, B. E. J., 2004. [OIII]/[NII] as an abundance indicator at high redshift. *MNRAS*, **348**, L59–L63.
- PILYUGIN, L. S., 2000. On the oxygen abundance determination in H II regions. The problem of the line intensities - oxygen abundance calibration. *A&A*, **362**, 325–332.
- PILYUGIN, L. S., VÍLCHEZ, J. M., MATTSSON, L. & THUAN, T. X., 2012. Abundance determination from global emission-line SDSS spectra: exploring objects with high N/O ratios. *MNRAS*, **421**, 1624–1634.
- REYNOLDS, D., 2015. Gaussian mixture models. *Encyclopedia of biometrics*, 827–832.
- RICHARD O. DUDA, PETER E. HART, D. G. S., 2001. *Pattern classification*. Wiley, 2nd ed. ISBN 9780471056690,0471056693.
- RIX, H.-W. & WHITE, S. D. M., 1992. Optimal estimates of line-of-sight velocity distributions from absorption line spectra of galaxies - Nuclear discs in elliptical galaxies. *MNRAS*, **254**, 389–403.
- ROTH, M. M., KELZ, A., FECHNER, T., HAHN, T., BAUER, S.-M., BECKER, T., BÖHM, P., CHRISTENSEN, L., DIONIES, F., PASCHKE, J., POPOW, E., WOLTER, D., SCHMOLL, J., LAUX, U. & ALTMANN, W., 2005. PMAS: The Potsdam Multi-Aperture Spectrophotometer. I. Design, Manufacture, and Performance. *PASP*, **117**, 620–642.
- ROUSSEEUW, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53 – 65. ISSN 0377-0427.
- RYBICKI, G. & LIGHTMAN, A., 2008. *Radiative Processes in Astrophysics*. Physics textbook. Wiley. ISBN 9783527618187.
- SÁNCHEZ, S. F., GARCÍA-LORENZO, B., JAHNKE, K., MEDIAVILLA, E., GONZÁLEZ-SERRANO, J. I., CHRISTENSEN, L. & WISOTZKI, L., 2006. Decoupling the host and

- nuclear spectra of type I AGNs using integral field spectroscopy: A test on 3C 120. *New A Rev.*, **49**, 501–507.
- SÁNCHEZ, S. F., KENNICUTT, R. C., GIL DE PAZ, A., VAN DE VEN, G., VÍLCHEZ, J. M. ET AL., 2012a. CALIFA, the Calar Alto Legacy Integral Field Area survey. I. Survey presentation. *A&A*, **538**, A8.
- SÁNCHEZ, S. F., PÉREZ, E., SÁNCHEZ-BLÁZQUEZ, P., GONZÁLEZ, J. J., ROSÁLEZ-ORTEGA, F. F., CANO-DÍAZ, M., LÓPEZ-COBÁ, C., MARINO, R. A., GIL DE PAZ, A., MOLLÁ, M., LÓPEZ-SÁNCHEZ, A. R., ASCASIBAR, Y. & BARRERA-BALLESTEROS, J., 2016. Pipe3D, a pipeline to analyze Integral Field Spectroscopy Data: I. New fitting philosophy of FIT3D. *RMXAA*, **52**, 21–53.
- SÁNCHEZ, S. F., ROSALES-ORTEGA, F. F., KENNICUTT, R. C., JOHNSON, B. D., DIAZ, A. I., PASQUALI, A. & HAO, C. N., 2011. PPAK Wide-field Integral Field Spectroscopy of NGC 628 - I. The largest spectroscopic mosaic on a single galaxy. *MNRAS*, **410**, 313–340.
- SÁNCHEZ, S. F., ROSALES-ORTEGA, F. F., MARINO, R. A., IGLESIAS-PÁRAMO, J., VÍLCHEZ, J. M. ET AL., 2012b. Integral field spectroscopy of a sample of nearby galaxies. II. Properties of the H II regions. *A&A*, **546**, A2.
- SOBELMAN, I., 1972. *Introduction to the theory of atomic spectra*. International series of monographs in natural philosophy. Pergamon Press.
- STASIŃSKA, G., 1978. Ionization and temperature structure of H II regions - The influence of the star, the gas density and its chemical composition. *A&AS*, **32**, 429–438.
- STASIŃSKA, G., 2002. The Electron Temperature in Ionized Nebulae. In W. J. Henney, J. Franco & M. Martos, eds., *Revista Mexicana de Astronomía y Astrofísica Conference Series*, vol. 12 of *Revista Mexicana de Astronomía y Astrofísica Conference Series*, 62–69.
- STASIŃSKA, G., 2004. Abundance determinations in H II regions and planetary nebulae. In *Cosmochemistry. The melting pot of the elements*, 115–170.
- STASIŃSKA, G., 2009. *What can emission lines tell us?*, 1.

- TAN, P.-N., STEINBACH, M. & KUMAR, V., 2005. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. ISBN 0321321367.
- THOMAS, A. D., GROVES, B. A., SUTHERLAND, R. S., DOPITA, M. A., KEWLEY, L. J. & JIN, C., 2016. A Physically based Model of the Ionizing Radiation from Active Galaxies for Photoionization Modeling. *ApJ*, **833**, 266.
- TUFFS, R. J., POPESCU, C. C., VÖLK, H. J., KYLAFIS, N. D. & DOPITA, M. A., 2004. Modelling the spectral energy distribution of galaxies. III. Attenuation of stellar light in spiral galaxies. *A&A*, **419**, 821–835.
- VALE ASARI, N., STASIŃSKA, G., MORISSET, C. & CID FERNANDES, R., 2016. BOND: Bayesian Oxygen and Nitrogen abundance Determinations in giant H II regions using strong and semistrong lines. *MNRAS*, **460**, 1739–1757.
- VERNER, D. A., FERLAND, G. J., KORISTA, K. T. & YAKOVLEV, D. G., 1996. Atomic Data for Astrophysics. II. New Analytic FITS for Photoionization Cross Sections of Atoms and Ions. *ApJ*, **465**, 487.
- VILA COSTAS, M. B. & EDMUNDS, M. G., 1993. The Nitrogen-To Ratio in Galaxies and its Implications for the Origin of Nitrogen. *MNRAS*, **265**, 199.