

Espectros de estrellas

Alberto Bañón Serrano, febrero de 2022

1. Resumen.

Este documento recoge el trabajo realizado para la clasificación automática de estrellas por su espectro. Se hace un resumen de la evolución histórica de los conceptos relevantes de un espectro así como de los elementos esenciales del algoritmo de clasificación automática K-means, para finalmente presentar los resultados de su aplicación a un conjunto de 8.442 espectros de estrellas, con valores de intensidad de radiación para 4.563 longitudes de onda. Estos resultados se comparan con la clasificación por temperaturas de Harvard y se propone cambiar los rangos de temperatura de esta clasificación.

Los centroides de los grupos en que resultan clasificadas las estrellas constituyen un modelo de clasificación ya que dado un espectro cualquiera basta calcular las distancias a cada uno de ellos para determinar el grupo en que debe encuadrarse la estrella.

Las aplicaciones informáticas realizadas para este trabajo están a disposición de todo el mundo escribiendo al correo alberto@interajedrez.com.

2. Introducción y desarrollo teórico.

Las estrellas se encuentran tan lejos de nosotros que en 1835 el filósofo Auguste Comte se atrevió a decir que nunca seríamos capaces de saber su composición química, su densidad o temperatura porque no podríamos llegar hasta ellas para tomar muestras. La verdad es que el Sol no está tan lejos y lo que en 1835 era una distancia insalvable hoy no lo es. Pero no hace falta viajar a las estrellas para saber eso que a Comte le parecía imposible, las estrellas vienen a nosotros.

Primero vamos a repasar algunos conceptos esenciales, empezando por lo que se entiende como espectro.

Desde muy antiguo se sabía que la luz producía una serie de colores al atravesar un vaso de agua y en el siglo XIII Roger Bacon indicó que el arcoíris sería algo parecido. Newton observó que cuando la luz atraviesa un prisma de vidrio aparecen una serie de franjas de colores (figura 1) y dedujo que la luz blanca era en realidad la suma de varias luces de colores (corpúsculos) que el prisma conseguía separar porque tenían velocidades distintas (Opticks 1671). Newton llamó espectro a esta descomposición de la luz en sus elementos y así se sigue llamando desde entonces.

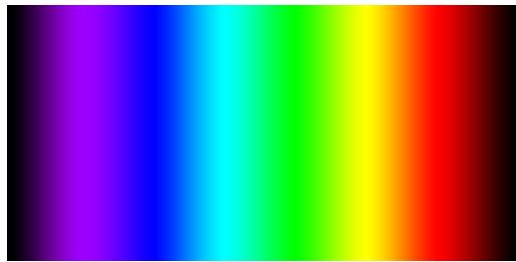


Figura 1. Espectro de la luz del sol.

Christian Huygens, coetáneo de Newton, discutió con este sobre la naturaleza de la luz defendiendo que era una onda y no una partícula y que el color se correspondía con la longitud de la onda y no con la velocidad de la partícula. Si colocamos la longitud de onda en el eje X del espectro (figura 2), el espectro visible por el ojo humano está comprendido entre las longitudes de onda de 400 nm y 750 nm.



Figura 2. Espectro visible por el ojo humano.

Aunque la mecánica cuántica sostiene que la luz es ambas cosas: onda y corpúsculo no veremos ningún espectro que en accisas tenga algo distinto a la longitud de onda o la frecuencia de esta.

En 1802 Wollaston, observando el espectro de la luz solar en detalle, encontró cuatro líneas oscuras que interpretó como puntos de separación de los colores hasta que en 1859 Kirchhoff y Bunsen demostraron que cada tipo de átomo emitía (al calentarlo) una serie de líneas características, o que absorbía la luz de esas mismas longitudes de onda, cuando pasaba a su través, siendo esta la verdadera razón de las líneas oscuras observadas por Wollaston. La explicación de la física teórica llegó en 1913 cuando Niels Bohr explico el espectro del átomo de hidrógeno, pero no entraremos en ella porque a los efectos de esta nota no se necesita.

2.1. La radiación del cuerpo negro.

Por otra parte, ya se sabía que todos los cuerpos, en función de su temperatura, emiten y absorben radiación electromagnética.

En el caso de los materiales sólidos, el espectro de la radiación se extiende a todas las longitudes de onda (emiten radiación de todas las longitudes de onda), aunque el máximo de emisión se produce a una determinada longitud de onda que depende de la temperatura, a bajas temperaturas el ojo humano no percibe radiación pero a temperatura suficientemente alta vemos al sólido brillar con luz propia.

Para el tratamiento teórico de este fenómeno Gustav Kirchhoff (1862) definió un sólido ideal que denominó cuerpo negro porque tenía la propiedad de absorber el cien por cien de la radiación que recibía así como de emitir el cien por cien de la energía que

recibía (por calentamiento) de aquí que estos estudios se denominen como la “radiación del cuerpo negro”.

Un aspecto que intrigaba a los físicos del momento es que el espectro de radiación era el mismo para todos los materiales a la misma temperatura, daba igual su composición.

Joseph Stefan encontró (1879) que la energía de la radiación total por unidad de área (un cuerpo negro con el doble de superficie que otro, radia el doble) era proporcional a la temperatura elevada a la cuarta potencia, cosa que Ludwig Boltzmann demostró teóricamente en 1884 a partir de consideraciones termodinámicas y electromagnéticas.

En 1894, Wihelm Wien dedujo teóricamente que el máximo de emisión se produce para una longitud de onda cuyo valor depende de la temperatura de forma inversamente proporcional, a más temperatura menor longitud de onda (fórmula 1).

$$\text{Longitud de onda (angstrom)} = 0.002898E+10 / \text{Temperatura (Kelvin)}$$

Fórmula 1. Ecuación que relaciona la temperatura con la longitud de onda para la que se presenta el máximo de la curva de radiación del cuerpo negro.

Aunque para los objetivos de esta nota podemos quedarnos aquí, vale la pena recordar el resto de la historia.

Los resultados teóricos de la ley de Wien difieren notablemente de los obtenidos experimentalmente. La discrepancia es mayor cuanto menor es la longitud de onda hasta el punto de que la teoría predecía que para longitudes de onda pequeñas (ultravioleta) la emisión de radiación tendía a infinito cuando los experimentos decían que tendía a cero. (figura 3). A este se le llamó la catástrofe ultravioleta.

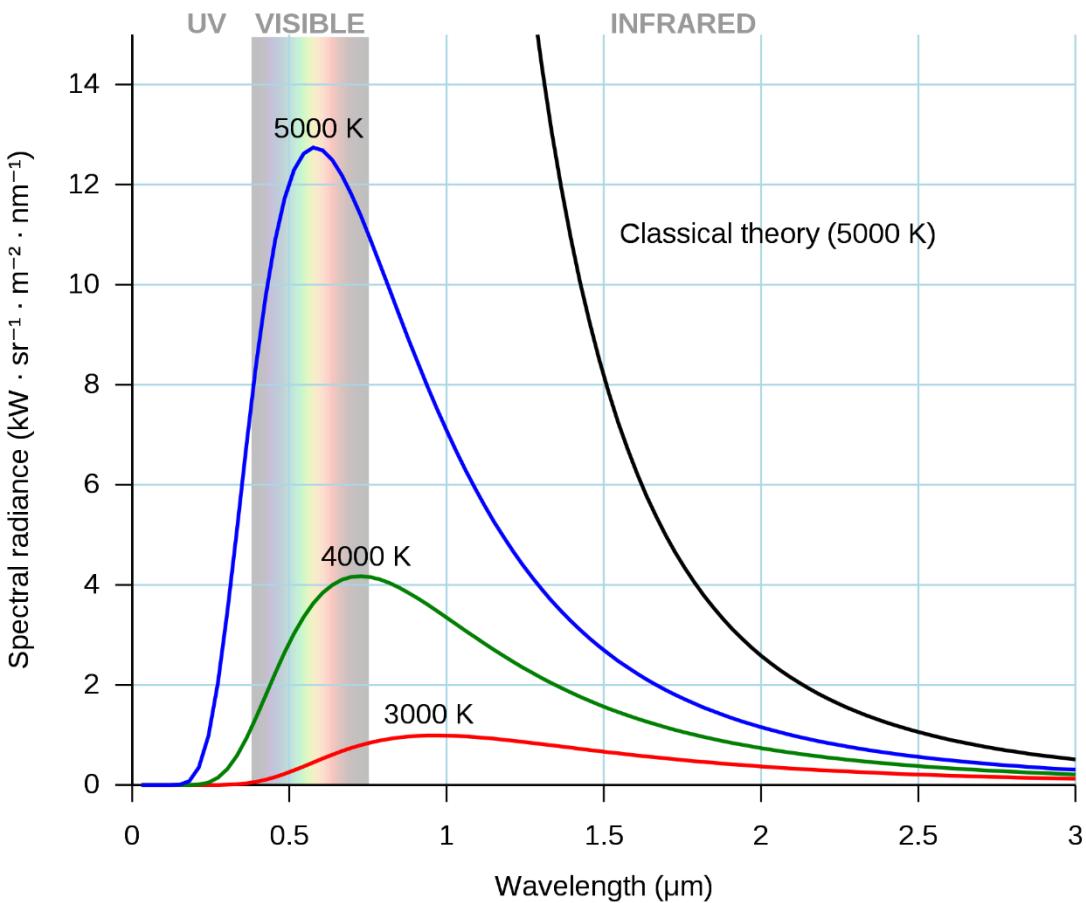


Figura 3. Radiación del cuerpo negro, curvas experimentales frente a los deducidos teóricamente (Classical theory).

La solución fue encontrada en 1900 por Max Planck al introducir el concepto de cuantización energética dando lugar al nacimiento de la mecánica cuántica. Básicamente, el mecanismo por el que la energía térmica se transforma en radiación electromagnética es la vibración de los átomos (ionizados) o moléculas del material que se calienta, este movimiento genera el campo electromagnético que origina la onda, a mayor temperatura la vibración es más intensa, es decir, mayor frecuencia de vibración y por tanto menor longitud de onda. La hipótesis de Planck es que la intensidad de vibración no cambia de forma continua, aunque la temperatura si lo haga, los osciladores cambian su intensidad de vibración a saltos. Con esta hipótesis y las funciones de probabilidad clásicas para la distribución de la energía total entre los osciladores, se reproducen las curvas experimentales con extraordinaria precisión.

2.2. El espectro de una estrella.

Aquí dejamos los aspectos históricos y teóricos para ir a la práctica.

Cuando miramos una estrella lo que vemos es radiación electromagnética que gracias a la Física nos aporta la información que Auguste Comte creyó que nunca tendríamos.

La luz que llega a nuestros telescopios se puede descomponer y obtener el correspondiente espectro de radiación con la intensidad de luz recibida para cada

longitud de onda, el máximo del espectro nos dirá la temperatura de la superficie de la estrella (que es lo que realmente vemos) y la presencia de líneas nos dirá que elementos químicos están presentes a su alrededor. Incluso podremos obtener información de la composición química de cualquier objeto estelar que este entre nosotros y la estrella porque el material del que está compuesto absorberá las radiaciones de determinadas longitudes de onda que faltaran en el espectro. Al igual que los átomos, las moléculas también absorben y emiten radiación de longitudes de onda que las identifican, lo que ha permitido detectar la presencia de muchas de ellas en el espacio interestelar.

La descomposición de la luz ya no se hace con prismas de vidrio, se hace con espectrómetros que usan redes de difracción que ya eran conocidas en tiempos de Newton (James Gregory) y fabricadas desde 1785 (David Rittenhouse), como curiosidad en el Anexo I se muestra como fabricar un espectrómetro en casa sin necesidad de ser muy habilidoso y en unos pocos minutos.

A continuación (figura 4) se muestra el espectro de una estrella visible en la dirección: 288,12 grados de ascensión recta y 51,448 grados de declinación. No creo que tenga asignado un nombre propio, cosa normal hoy en día que se han catalogado cientos de miles de estrellas. El rango de longitudes de onda del espectro va de 3.600 a 10.000 angstrom (360 a 1000 nm), todo el espectro visible y una parte del infrarrojo y ultravioleta. El máximo de radiación está sobre los 5.300 angstrom = 0,53 μ m que si miramos en la figura 3 veremos que corresponde a una temperatura alrededor de 5.000 grados Kelvin.

Entrando en el detalle del espectro veremos multitud de líneas, prácticamente todas de absorción (la intensidad de la radiación cae bruscamente para determinadas longitudes de onda) que muestran la existencia de ciertos elementos químicos (a 5.000 K no hay moléculas). En las figuras 5, 6 y 7 se han marcado las líneas de absorción del Hidrógeno, Helio y otros elementos químicos. La coincidencia de una línea espectral con una línea del espectro identifica a un elemento (a veces es difícil distinguir entre varios) y como puede verse se detectan una gran variedad de elementos, incluso se puede determinar su abundancia relativa a partir de la intensidad de radiación absorbida.

Las imágenes han sido obtenidas mediante la aplicación informática de la casa: ExploraFits.

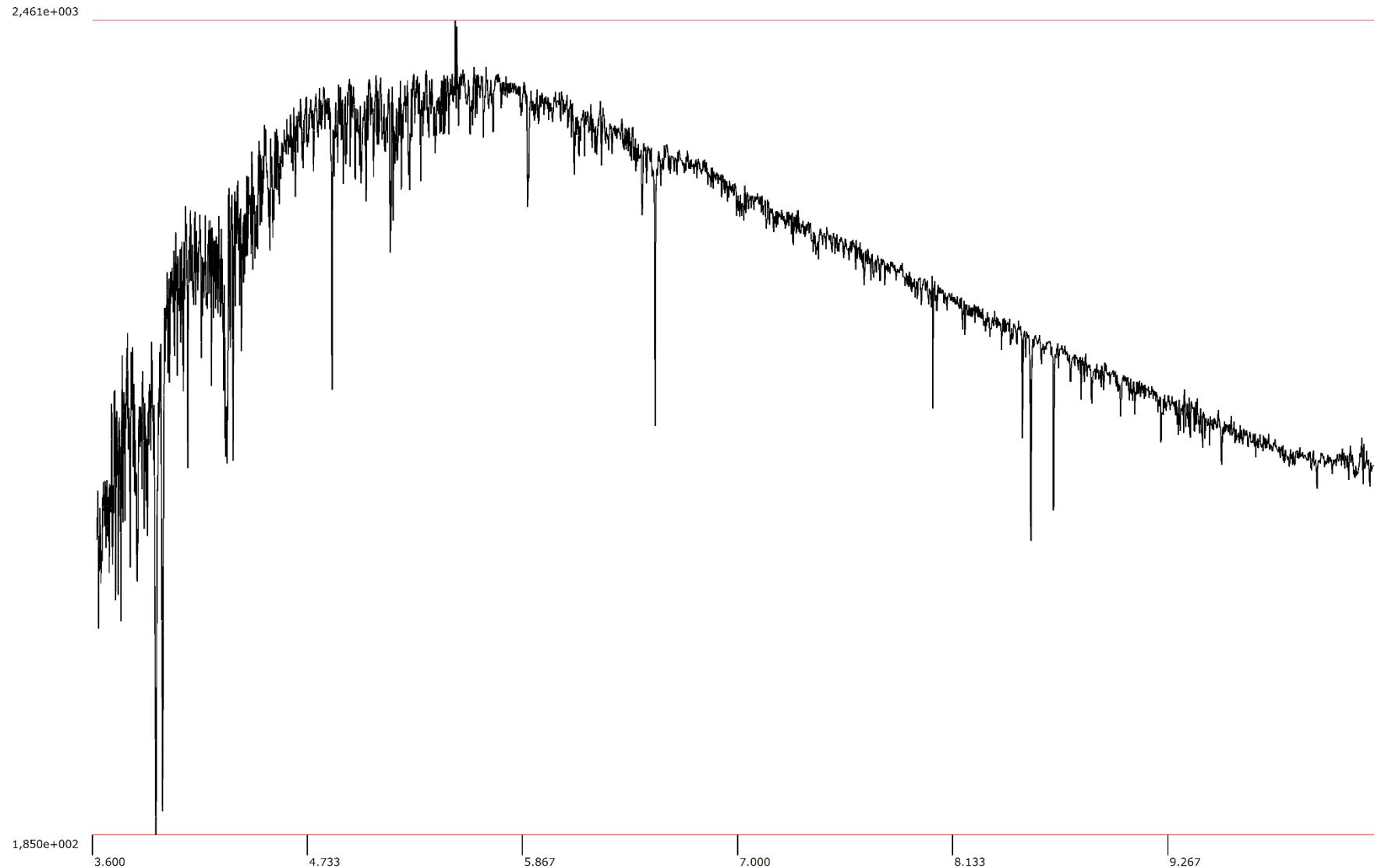


Figura 4. Espectro de una estrella. En *accisas* la longitud de onda en angstrom.

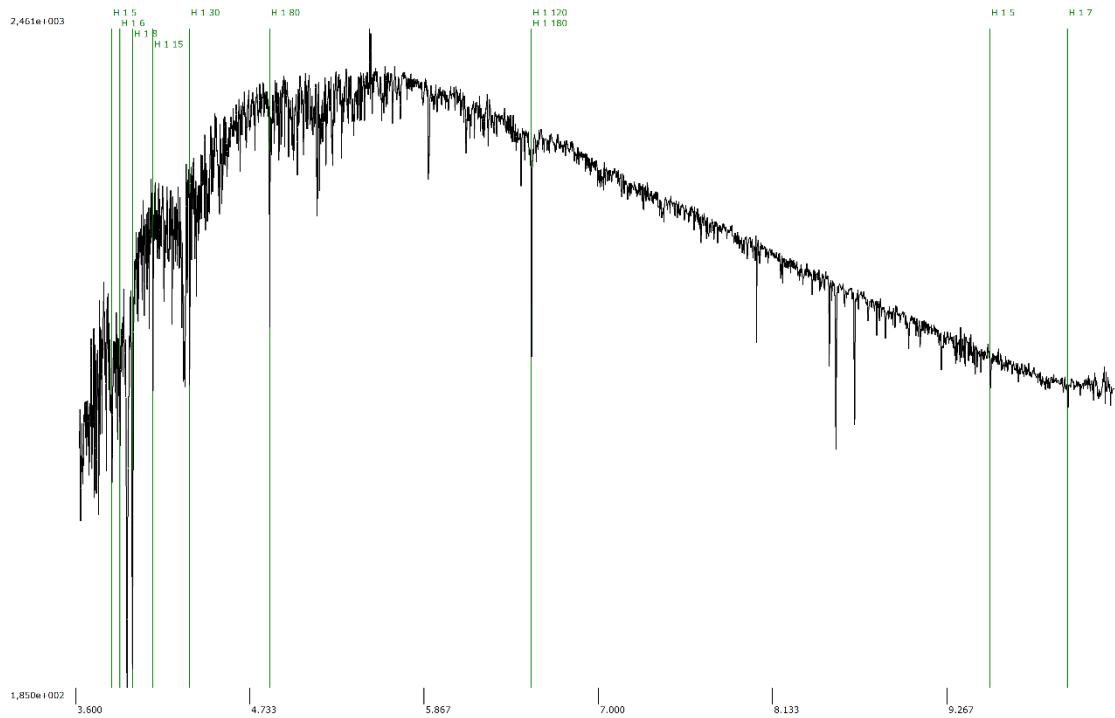


Figura 5. Líneas espectrales del Hidrógeno.

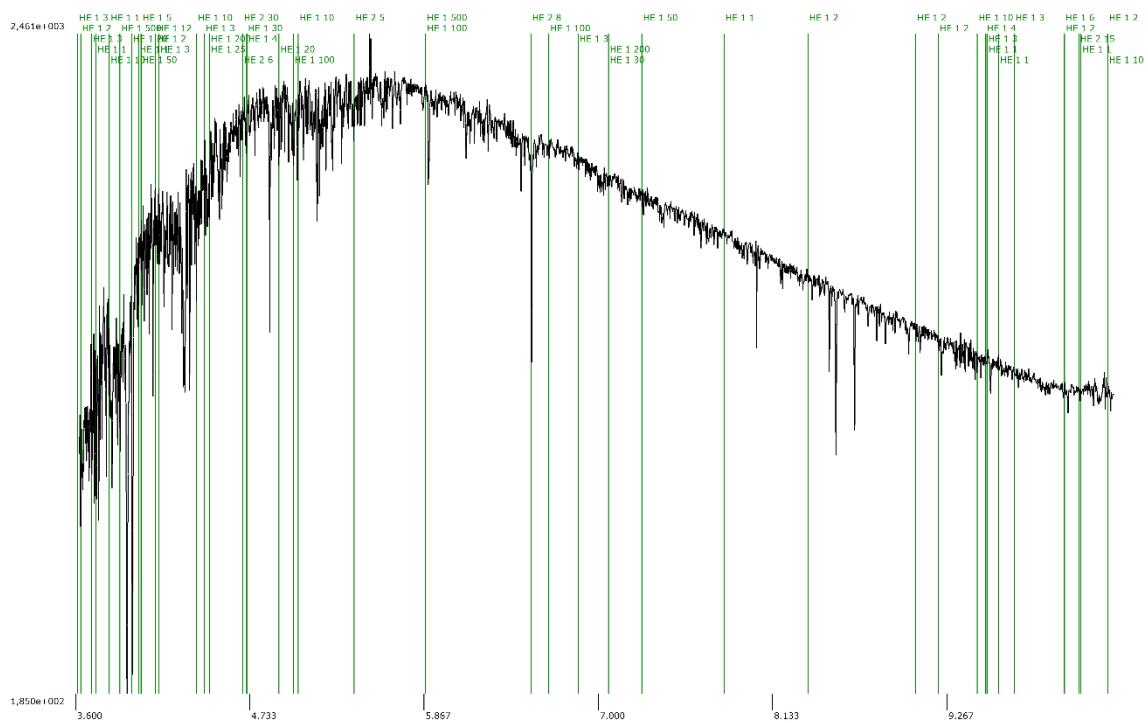


Figura 6. Líneas espectrales del Helio.

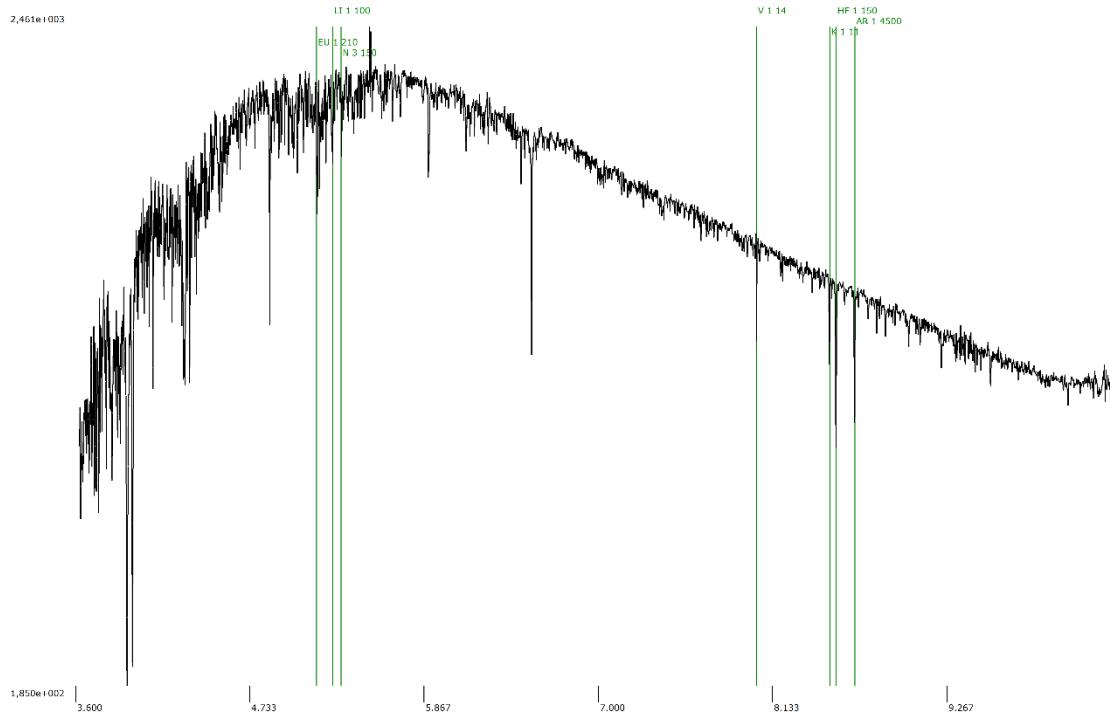


Figura 7. Líneas espectrales de algunos elementos químicos.

2.3. Clasificando estrellas.

El hombre siempre trata de poner orden en lo que observa y ya en la antigüedad inventó las constelaciones para agrupar las estrellas pero es a partir de 1860 cuando Angelo Secchi tras analizar miles de espectros de estrellas empezó a clasificarlas con criterios objetivos, inicialmente con el más evidente: el color.

1. estrellas blancas del tipo de Sirio y Vega,
2. estrellas amarillas de tipo solar,
3. estrellas anaranjadas y variables como Betelgeuse y Antares,
4. estrellas muy rojas como Mira.

Henry Draper amplió la clasificación de Secchi hasta 16 tipos identificados por las letras A, B, C y siguientes. Edward C. Pickering continuó el trabajo de Draper y catalogó 225.000 estrellas, su colaboradora Annie Jump Cannon redujo a 10 el número de tipos, ordenados por temperaturas decrecientes, con las denominaciones O B A F G K M N R S que actualmente se utiliza, si bien cada tipo se ha dividido en 10 subtipos.

A continuación me voy a extender un poco más de lo normal para esta nota, porque muestra como un simple diagrama puede conducir a algo totalmente inesperado.

Otra característica de las estrellas que resultaba evidente, además del color, era su brillo, al tratar la radiación del cuerpo negro vimos que hay una relación entre temperatura y luminosidad (energía) que se había concretado tanto experimentalmente como teóricamente (Stefan- Boltzmann) en que la energía (luminosidad) radiada por el cuerpo negro es proporcional a temperatura elevada a la cuarta potencia.

Pero el brillo de una estrella es en sí mismo bastante ambiguo porque se debe a su luminosidad real y a la distancia a la que se encuentra de nosotros, ninguna de ambas características se conoce de forma independiente salvo para muy pocas estrellas. Pero en casos concretos como las estrellas de un cúmulo se puede suponer una distancia similar para todas y entonces la diferencia en brillo lo es en luminosidad, con esta hipótesis, Ejnar Hertzsprung (1913) representó en un diagrama X-Y, siendo X la temperatura e Y la luminosidad relativa, las estrellas de las Pléyades y Hyades, a su vez y de forma independiente Russel realizo la misma representación con las estrellas para las que si se conocía su distancia a la tierra y por tanto su luminosidad, el diagrama (figura 8) que hoy se conoce como diagrama H-R por las iniciales de sus creadores, resultó ser extraordinariamente revelador. Pero antes que nada hay que tomar nota de que mientras la luminosidad se representa en el eje Y como es habitual con valores creciente hacia arriba, la temperatura lo hace en el eje X lo contrario a lo normal, decreciendo hacia la derecha (por aquello de seguir el orden de longitudes de onda del espectro, con el ultravioleta a la izquierda y el infrarrojo a la derecha).

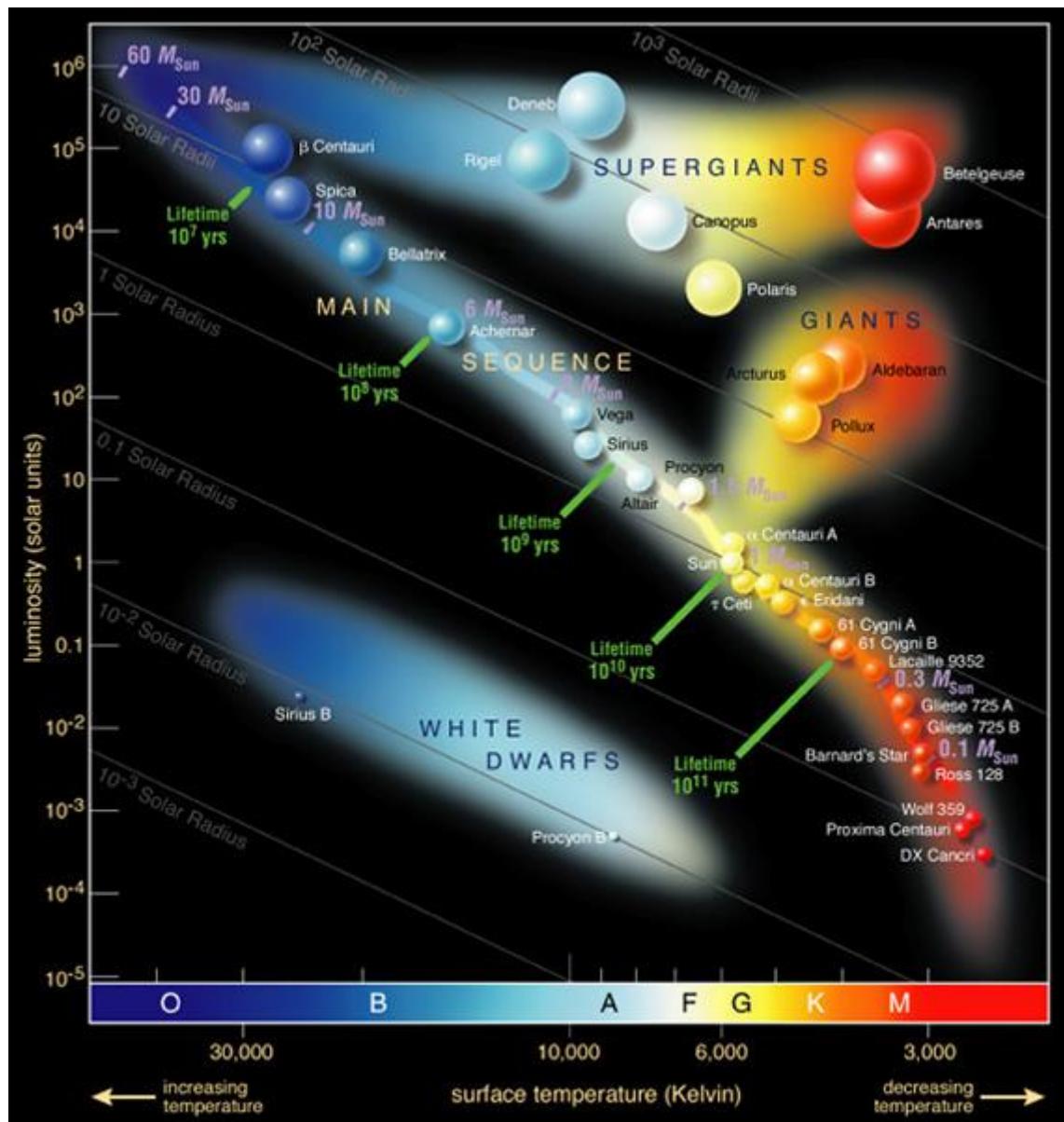


Figura8. Diagrama Hertzsprung-Russel (HR). Fuente:
https://upload.wikimedia.org/wikipedia/commons/1/17/Hertzsprung-Russel_StarData.png

Lo primero que cabe preguntarse es porque puede haber estrellas sobre una misma vertical, es decir con diferentes luminosidades a la misma temperatura. La respuesta lógica es porque tienen diferentes tamaños, a la misma temperatura más luminosidad implica mayor tamaño (radio si las consideramos esferas). Lo mismo se puede decir en horizontal, si dos estrellas muestran la misma luminosidad pero tiene temperaturas distintas, la más fría es la más grande (radia menos -menos temperatura- por unidad de superficie pero tiene más superficie).

Esto implica que en la parte superior derecha del diagrama encontraremos a las gigantes rojas (roja=baja temperatura y alta luminosidad=grande), por el contrario, en la parte inferior izquierda estarán las enanas blancas (blancas=alta temperatura y baja luminosidad =pequeña)

Hasta aquí nada en especial, lo sorprende del diagrama es que las estrellas no se dispersan por el mismo, se concentran en varias “familias”; una familia sobre la diagonal, otra en la parte derecha superior y otra en la parte izquierda inferior.

La diagonal que se conoce como secuencia principal sugiere que las estrellas evolucionan en el tiempo siguiendo una secuencia predeterminada, un ciclo de vida. Las estrellas nacen frías y poco luminosas (vértice inferior-derecho) y evolucionan a calientes y luminosas (vértice superior-izquierdo).

En el año 1924 Arthur Eddington descubrió la relación entre la luminosidad y la masa y luego sugirió que la energía de las estrellas provenía de la reacción nuclear de fusión de hidrógeno en helio, cosa que confirmó Hans Bethe en 1939. Hoy se conocen con detalle las distintas reacciones nucleares que rigen la evolución de las estrellas pero el diagrama H-R fue el que por primera vez sugirió la evolución estelar.

3. Clasificación automática

En los últimos tiempos el desarrollo de los métodos de clasificación se ha visto impulsado de forma exponencial por la capacidad de cálculo de los ordenadores y puesto de moda gracias a la inteligencia artificial.

Los trabajos en astronomía a este respecto se centran más en automatizar el reconocimiento de estrellas o galaxias, como se hace con el de objetos en imágenes, que en el análisis de los patrones que los métodos de clasificar establecen de una forma o de otra para automatizar el proceso de clasificación.

El análisis de los patrones de clasificación automática aunque no responden a nada que tenga que ver con el razonamiento humano (como a veces se trata de interpretar) pueden dar pistas sobre regularidades no observables a simple vista.

La clasificación de estrellas basada en la temperatura sólo hace uso de un punto del espectro: el correspondiente a la longitud de onda donde se da el máximo de radiación (fórmula 1), en este documento se utiliza el espectro en su conjunto mediante un análisis de agrupación (clustering) por el método K-means.

3.1. El modelo. K-means (K-medias).

Ahora se trata de agrupar las estrellas atendiendo a su espectro. El método de agrupación más antiguo y de los pocos no supervisados, es decir, que no requieren saber la solución de antemano, es el K-means.

Se establece un criterio de similitud que sea cuantificable para hacer grupos con los espectros similares (adelanto que no es con los más similares entre sí, sino con los más similares a una referencia o patrón). El objetivo más allá de saber quiénes se parecen a ese patrón es llegar a saber la razón del parecido.

El método se suele explicar con un ejemplo geométrico, donde se trata de agrupar una serie de puntos diseminados en un plano. De partida:

1. Se hacen N grupos de forma aleatoria.
2. A continuación se calculan los N puntos que son el “centro de masas” de cada uno de los N grupos. Esto se puede hacer de varias formas pero lo normal es calcular el punto medio de todos los puntos del grupo -la simple media (means) de cada una de las dos coordenadas X/Y-). A este punto se le llama centroide.
3. Una vez tenemos los N centroides nos olvidamos a que grupo pertenecían inicialmente los puntos y llevamos cada uno al grupo cuyo centroide sea el más cercano al punto (basta calcular la distancia del punto a los N puntos centroides).
4. Volvemos al paso 2 y repetimos el proceso hasta que ningún punto cambia de grupo.

Al final tendremos un centroide en cada grupo junto a los puntos más próximos a él, los que se parecen a él.

Los puntos son similares en el sentido de que son los más próximos al centroide, que en el caso de las distancias euclidianas significa que las coordenadas de los puntos se parecen más a las del centroide de su grupo que a las de cualquier otro centroide, pero es perfectamente posible (y es normal que ocurra) que dos puntos muy próximos entre si pertenezcan a dos grupos distintos.

El centroide es el patrón de similitud. En este ejemplo de los puntos, es posible que no valga la pena analizar el patrón para ver qué tiene de especial, pero en el caso de los espectros puede ser interesante.

Si consideramos un espectro como un vector donde cada coordenada es una longitud de onda y procedemos a clasificar un gran número de ellos, el centroide resultante para cada grupo será, por sí mismo, un espectro con un valor de la “radiación”, para cada longitud de onda, igual a la media de los valores de todos los espectros en el grupo para esa longitud de onda. Se trata de ver si podemos extraer alguna conclusión del análisis del centroide y de la comparación de este con los otros centroides. Hay que insistir en que mediante K-means no se agrupan las estrellas más parecidas entre si (que sería un mero ejercicio taxonómico), sino las más parecidas a un espectro patrón. La similitud entre espectros es indirecta.

Son varias las formas de tratar los espectros antes de agruparlos y las comentaremos más adelante pero primero hagamos algunas consideraciones sobre el método K-means.

Lo primero es mencionar que el número de grupos hay que establecerlo a priori, es un problema menor que se salva mediante prueba y error y se han propuesto algunos indicadores para seleccionar el número óptimo de conjuntos. Es evidente que cuanto mayor sea el número de grupos mejor será la clasificación resultante, pero si realmente hay objetos similares, lo que se espera es que a partir de un número de grupos determinado la calidad de la clasificación no mejore significativamente.

Más importante que lo anterior es la dependencia de los resultados finales a la selección inicial de los grupos (primer paso del método), es el problema de los óptimos locales de los que no se consigue salir una vez se cae en ellos. A menudo se busca que los grupos iniciales tengan centroides lo más alejados posibles entre ellos (hay un par de métodos sencillos y generalmente aceptados para hacerlo, como el denominado Kmeans++). Esto suele dar buenos resultados pero no es una garantía, sólo una heurística, la mejor solución es repetir el proceso de clasificación tantas veces como se pueda partiendo de grupos iniciales distintos y quedarse con la mejor y es aquí donde surge el siguiente problema: ¿Cuál es la mejor?

Después de cada clasificación (agrupamiento) podemos empezar por calcular la distancia media de los espectros respecto al centroide de su grupo, así como la desviación estándar y luego la media y desviación estándar extendida a todos los grupos, pero lo normal es que la clasificación con la mejor media global no sea la de mejor desviación estándar, ¿A que nos atenemos?, yo me decanto por la media ya que al fin y al cabo las iteraciones que realiza el método buscan minimizar ese valor medio global llevando al espectro al grupo con el centroide más cercano (con independencia de la desviación respecto de la media). Una vez alcanzada la solución, si se pasa un espectro de un conjunto a otro se empeora necesariamente la media global, aunque la desviación estándar puede mejorar.

Si queremos que no sea la media la que conduce el proceso de optimización, podemos modificar el segundo paso del método, en el que se calculan los centroides, por ejemplo, no considerando en el cálculo de los centroides a los peores espectros de cada grupo, excluimos a los más alejados del centroide. Esto conduce a soluciones con una mejor desviación estándar porque en ella interviene el cuadrado de las distancias, pero no sabemos de antemano si la solución será una mejor clasificación de los espectros. Más adelante propongo un método para conducir las optimizaciones de otras formas que es más versátil.

El problema de valorar las clasificaciones obtenidas, aunque sea de forma relativa de una clasificación a otra, es una constante en cualquier método de clasificación y la idea que subyace en todos los casos es combinar lo cerca que están los objetos de su centroide (cohesión) con lo lejos que están de los otros (separación). El más utilizado es el índice Davies Bouldin (DB) que mide la cohesión por la distancia media de los espectros con su centroide y la separación por la distancia entre centroides, uno más sofisticado es el índice silhouette que mide la separación con más detalle calculando la distancia entre cada espectro del grupo y los espectros de los otros grupos, pero es prohibitivo por motivos de tiempo de cálculo, si tratamos de clasificar 10.000 espectros hay que calcular:

$$10.000 \times 10.000 = 10^8 \text{ distancias}$$

Si tratamos de conducir la optimización mediante estos indicadores, por ejemplo el índice DB, se llega a clasificaciones absurdas, al menos en el caso de los espectros es lo que he observado.

El absurdo consiste en que prácticamente todos los espectros se ubican en un mismo grupo y sólo unos pocos, a veces uno sólo, en los otros grupos. Situación que ocurre si una minoría es muy distinta del resto, al final es preferible llevar los “diferentes” a grupos independientes de forma que los centroides de estos grupos quedan muy alejados del centroide del grupo muy poblado, maximizándose la componente de “separación” del índice y por tanto el índice.

Para evitar la dependencia de los resultados de la selección inicial de los grupos he mencionado la posibilidad de repetir muchas veces la clasificación desde diferentes puntos de inicio, es decir, un método Monte Carlo de fuerza bruta, pero además, lo que he hecho es combinar K-means con un algoritmo genético.

Cada posible clasificación es un cromosoma y los genes son los centroides, mediante el Monte Carlo de fuerza bruta generamos un conjunto inicial de cromosomas y luego aplicamos el algoritmo genético para recombinar los genes (centroides) en nuevos cromosomas, es decir, en nuevas soluciones.

Sólo hay un problema, mientras que estamos acostumbrados a que los genes tengan nombre propio y como tales podemos identificar al mismo gen en los cromosomas del padre y de la madre, nuestros centroides son indistinguibles y no demos decir que el hijo toma el gen A del padre y el gen B de la madre porque no hay centroide A ni B. Si elegimos al azar la mitad de los centroides de cada uno de los dos progenitores la mayoría de las veces obtendremos cromosomas sin sentido.

La solución que he adoptado es elegir al azar la mitad de los centroides del padre y descartar los centroides de la madre que más cerca estén de los elegidos del padre, pasando al hijo los restantes.

La ventaja de introducir el algoritmo genético es que podemos elegir el criterio de supervivencia que es lo que en el fondo determina la evolución. Si establecemos que los cromosomas (clasificaciones) mejor adaptados son los de menor desviación estándar, la evolución conducirá a cromosomas con menor desviación estándar, lo mismo ocurrirá si valoramos como mejor adaptados a los de mejor índice DB. En definitiva, podemos conducir la evolución de la forma que queramos, cosa que con K-means no podemos.

Este método mixto K-means-genético ha funcionado bastante bien. Si partimos de un conjunto inicial de cromosomas muy grande, por ejemplo 10.000, a la parte genética le cuesta mucho mejorar la media global de la clasificación porque es probable que alguno de esos 10.000 cromosomas iniciales ya esté próximo al óptimo global porque los hemos generado mediante K-means que busca el valor mínimo de la media global. Por el contrario, mejora fácilmente las clasificaciones si lo que se quiere es minimizar la desviación estándar o el índice DB.

3.2 Los Datos.

Una vez hecho el planeamiento del modelo hay que pasar al tratamiento de los datos. Lo primero es conseguir un buen número de espectros y esto lo hemos conseguido gracias al proyecto Sloan Digital Sky Survey (SDSS):

<https://www.sdss.org/>

Este proyecto ha obtenido el espectro de decenas de miles de estrellas en condiciones homogéneas y los ha puesto a disposición publica, en particular ha seleccionado, de entre todas sus observaciones, 8.646 espectros especialmente buenos todos con valores para las mismas 4.563 longitudes de onda, de 3.600 a 10.350 Angstrom.

<https://www.sdss.org/dr15/mastar/mastar-data-access/>

El rango de temperaturas que corresponde a este rango de longitudes de onda es de 8.000 a 2.800 grados Kelvin, de acuerdo con la fórmula de Wein (fórmula 1). En la figura 9 se muestra el histograma con la distribución de los espectros según su temperatura, calculada a partir de la longitud de onda del valor máximo de radiación, para determinar este máximo se ha ajustado un polinomio a cada espectro y para todos, salvo para 204, el polinomio presenta un máximo dentro del rango de longitudes de onda. Estos 204 espectros se han eliminado del conjunto de datos que se utiliza en este estudio.

La figura 10 es la distribución de los espectros de acuerdo con la clasificación de Harvard (tabla 1)

Clase	Temperatura
O	$\geq 33\,000\text{ K}$
B	10 000–33 000 K
A	7500–10 000 K
F	6000–7500 K
G	5200–6000 K
K	3700–5200 K
M	$\leq 3700\text{ K}$

Tabla 1. Clasificación de estrellas de Harvard

La figura 4 es uno de los espectros del proyecto SDSS y la figura 11 el mapa de los 8.442 espectros de acuerdo con la ascensión recta y declinación de cada espectro (en grados).

452

G 1. 8.442. 8.442

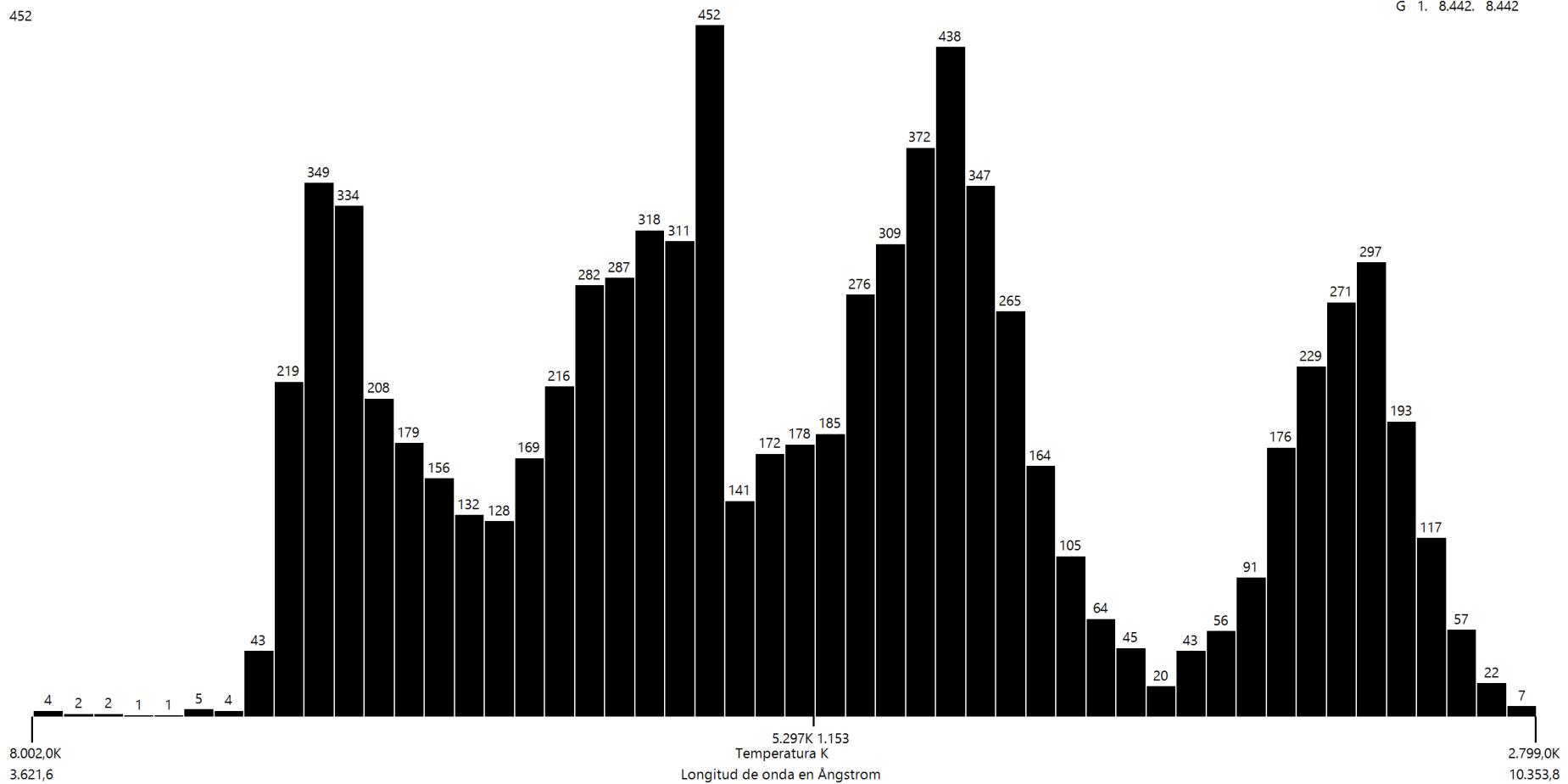


Figura 9. Histograma con la distribución de los 8.442 espectros por su temperatura. Los histogramas tienen un ancho de 102 K.

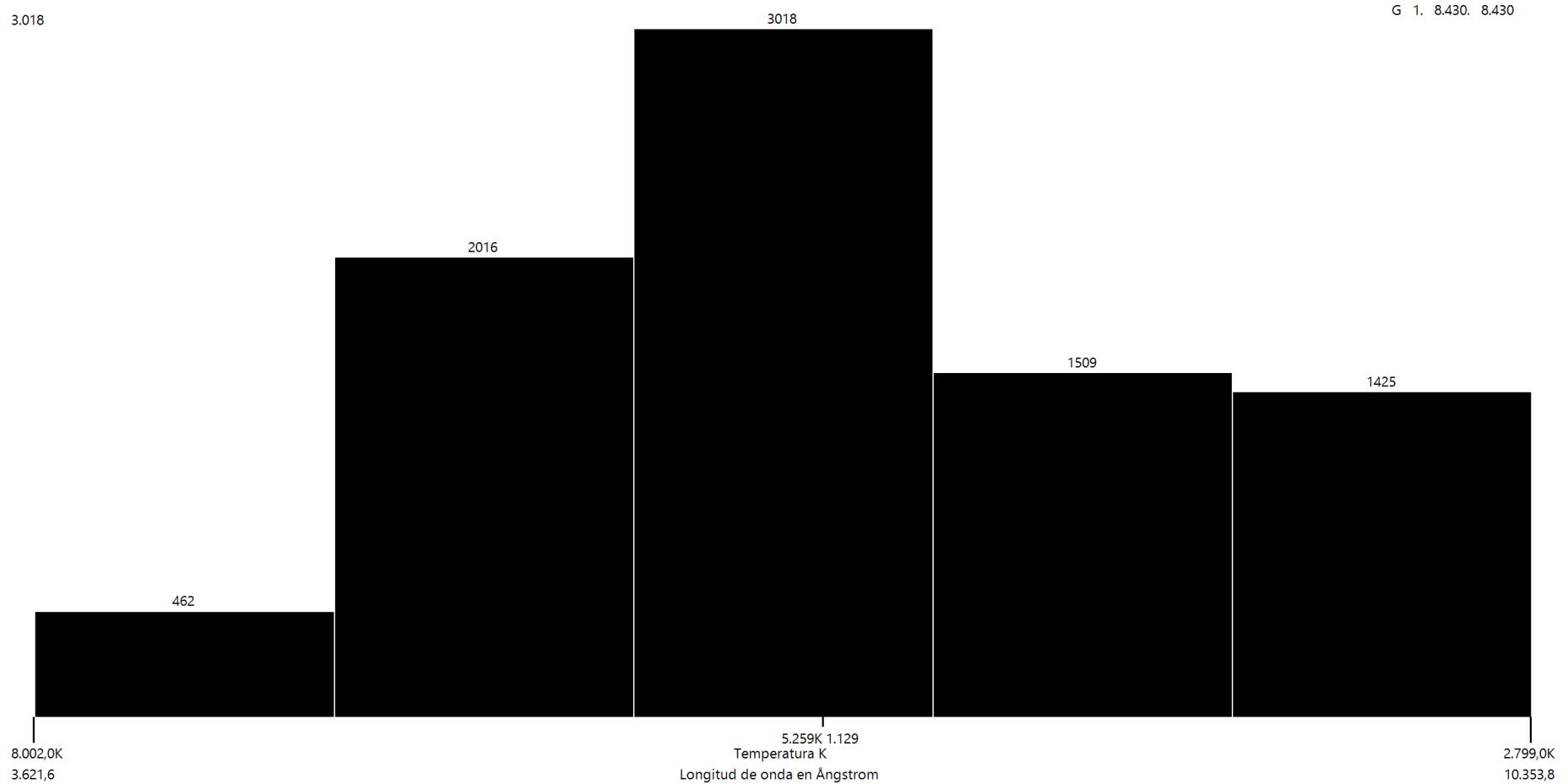


Figura 10. Histograma con la distribución de los 8.442 espectros conforme a la clasificación de Harvard: A, F, G, K, M de izquierda a derecha.

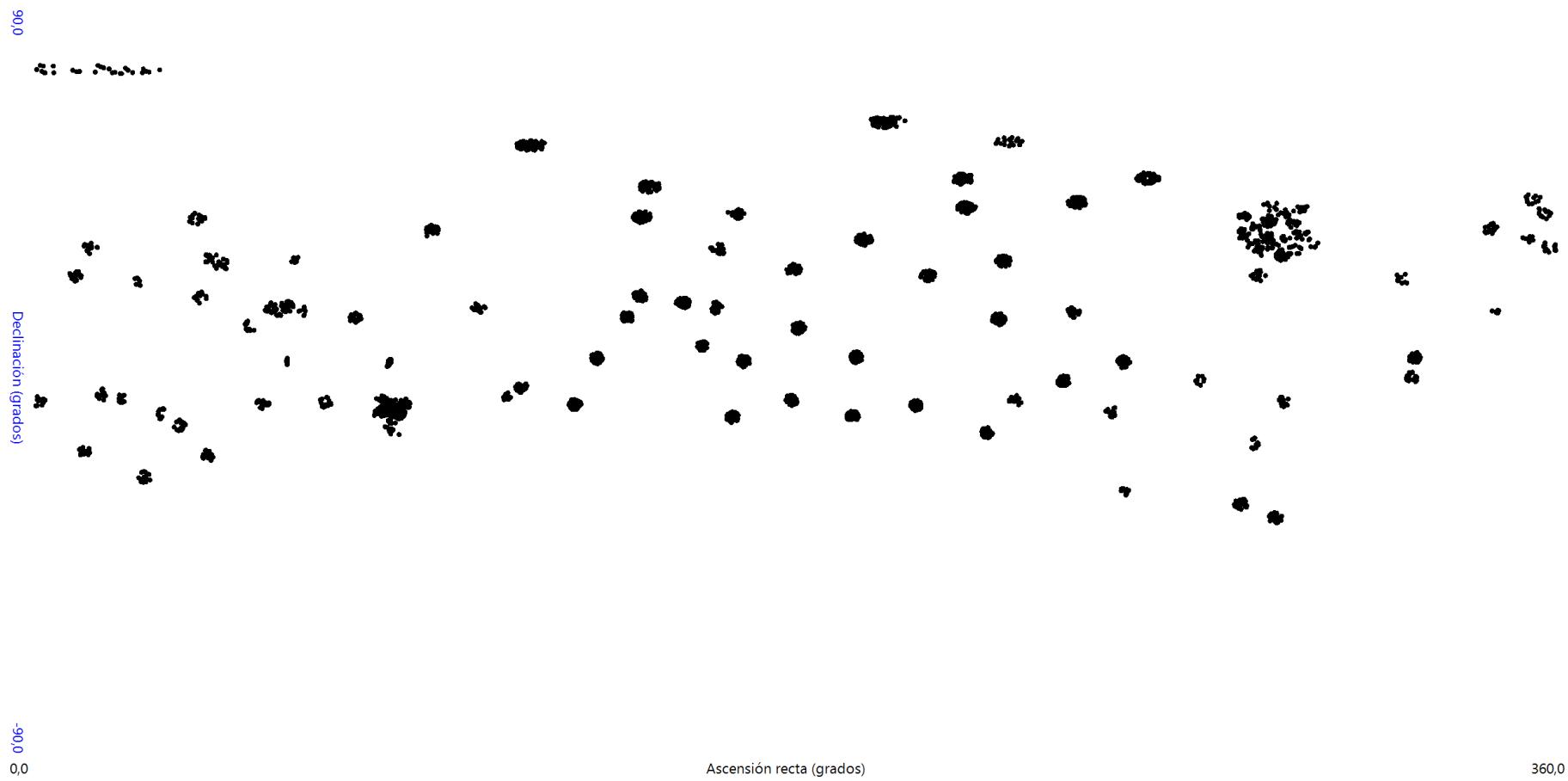


Figura 11. Mapa de las 8.442 estrellas cuyos espectros se van a clasificar en este estudio. La proyección en el gráfico no considera curvatura en los ejes.

Para efectuar la clasificación de las estrellas se han ensayado varias posibilidades en cuanto a los datos y a la forma de calcular la distancia entre espectros:

1. Utilizar los espectros tal cual: vectores con 4.563 valores para las correspondientes 4.563 longitudes de onda.

a. Distancia euclíadiana

b. Covarianza. En realidad se usa el valor $(1 - \text{Covarianza})$ para que tenga el mismo sentido que la distancia euclíadiana, dos espectros están más cerca cuanto menor es la distancia, para dos espectros idénticos la covarianza es 1.

2. Utilizar los espectros normalizados por la radiación máxima. Cada dato se divide por el valor máximo del espectro.

a. Distancia euclíadiana

b. $1 - \text{Covarianza}$.

3. Utilizar la diferencia entre el dato del espectro y el valor del polinomio ajustado. De esta forma se pretende eliminar la radiación térmica para destacar las líneas de absorción y emisión del espectro. Para ser preciso, no se utilizan las diferencias estrictamente algebraicas, teniendo en cuenta que los datos del espectro están sometidos a cierta incertidumbre (ruido) lo que he hecho ha sido lo siguiente:

Si la diferencia entre espectro y polinomio, en valor absoluto, es menor que un determinado valor, se registra un 0, en caso contrario se registra la diferencia. Como valor de corte se ha utilizado un valor que es distinto para cada longitud de onda ya que es un porcentaje del valor del espectro. La figura 12 muestra esto para el espectro que nos está sirviendo de ejemplo. Todos los puntos dentro de la banda delimitada por las líneas naranja se consideran cero y las únicas longitudes de onda para las que el valor no es cero son las correspondientes a los puntos azules y verdes. Como valor de corte (diferente para cada longitud de onda) se ha utilizado la desviación estándar del error que resulta del ajuste al polinomio, dividido por la media del valor del espectro y el ajuste para cada longitud de onda y multiplicada por 1,4, este coeficiente es arbitrario y se ha elegido para obtener un cierto grado de simplificación sin perder los picos significativos. Las pruebas realizadas con coeficientes 1 y 2 no han ofrecido resultados muy distintos. Finalmente la diferencia se expresa en términos relativos dividiéndola por la media entre el dato del espectro y el polinomio ajustado, para cada longitud de onda.

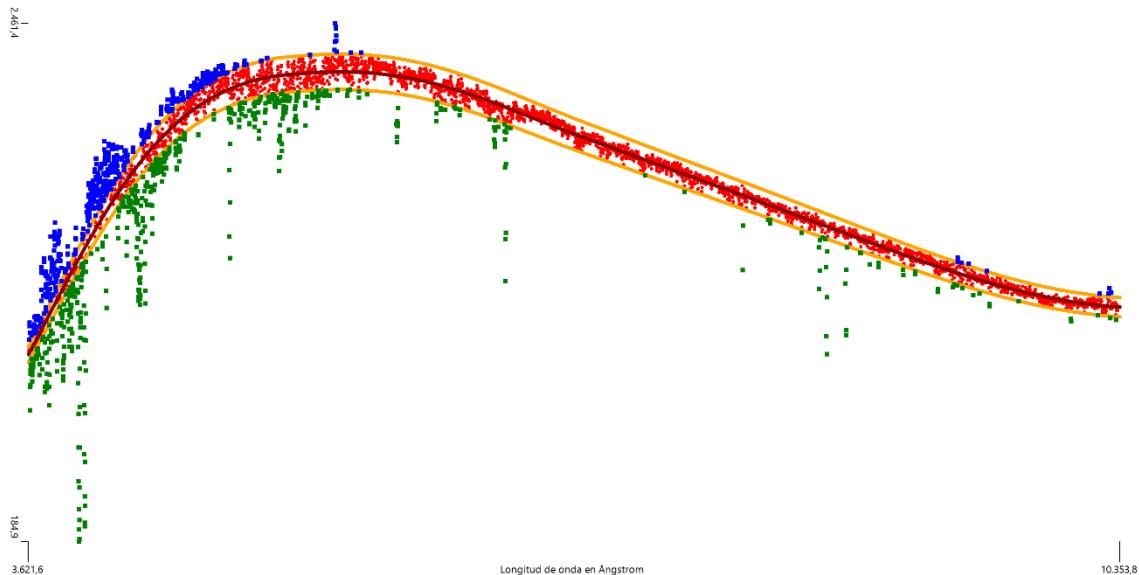


Figura 12. Preparación de los datos mediante ajuste a un polinomio.

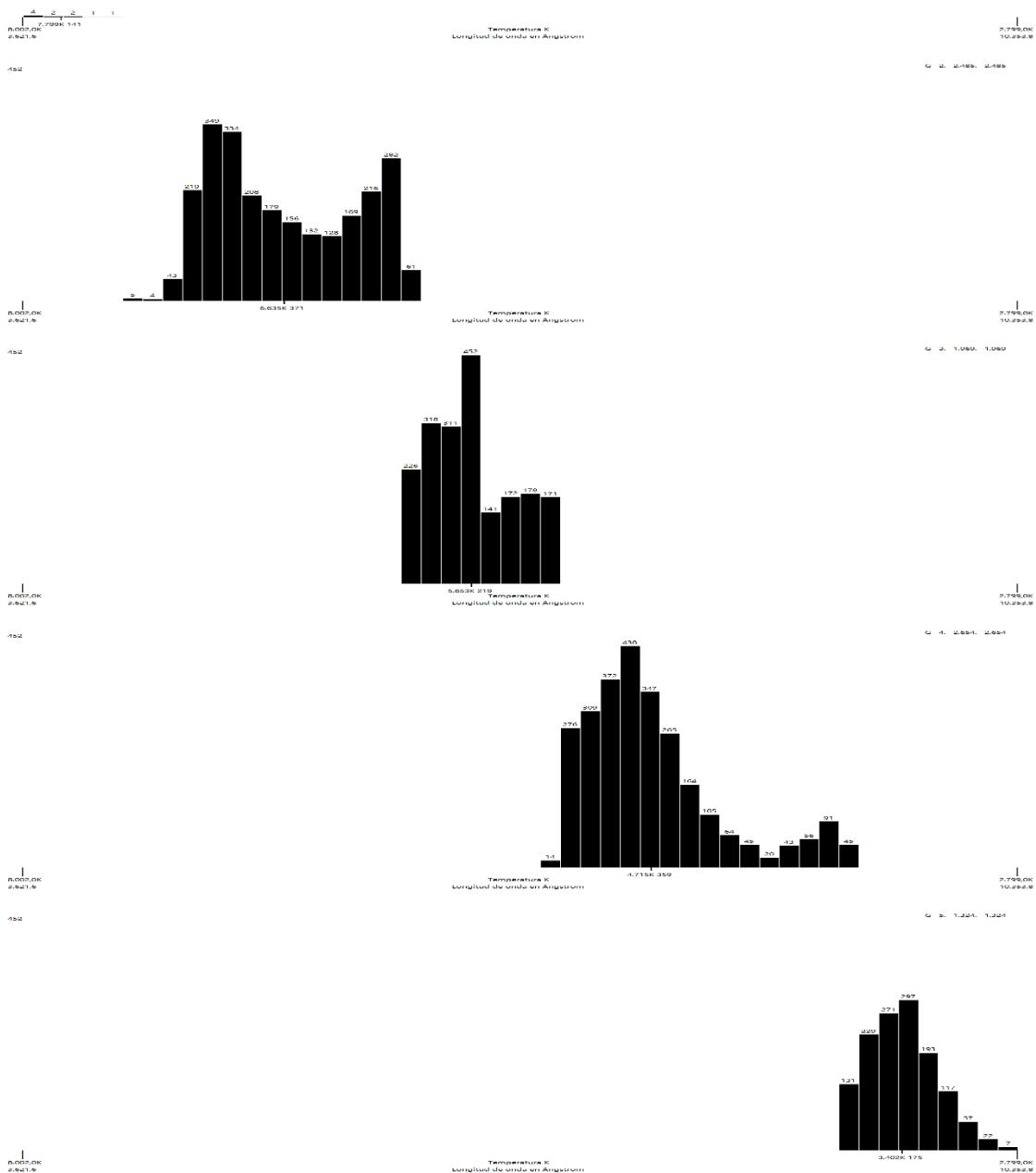
4. El instrumento.

Para realizar las clasificaciones y el análisis de los resultados he escrito una aplicación informática (Kspectro), que juega el papel de instrumento de experimentación, se ha ido dotando de todas las opciones de cálculo y tratamiento de los datos que se querían probar y se describe en el Anexo II.

La aplicación incluye su propia programación del método K-means, pero también puede realizar una clasificación K-means usando la librería de Microsoft, con el propósito de validar lo programado. Era imprescindible programar el algoritmo K-means para poder adaptarlo a todos los escenarios que se han probado, cosa imposible con las librerías estándar.

5. Resultados.

El escenario de referencia es la clasificación de Harvard, no es una clasificación K-means ya que los espectros se clasifican simplemente por su temperatura dando lugar a cinco grupos como se ve en la figura 10, que corresponden a las clases: A, F, G, K y M de Harvard. La distribución de espectros en cada grupo se ve en la figura 13.



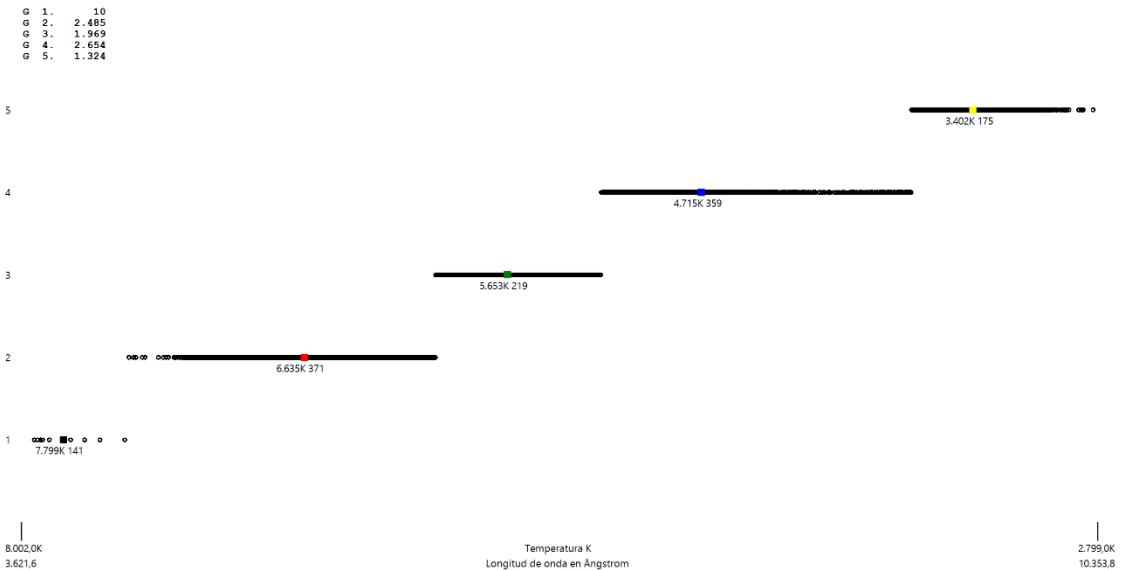


Figura 13. Distribución de los espectros con la clasificación Harvard, por definición los grupos son disjuntos respecto a la temperatura.

A esta clasificación le hemos aplicado las mismas métricas que usaremos en las clasificaciones K-means, incluido el cálculo de los centroides de cada grupo (media aritmética de los espectros incluidos en el grupo). En las tablas 2, 3 y 4 se muestran los resultados en diversos escenarios de datos y distancias.

- Tabla 2. Datos originales y distancia euclídea.
- Tabla 3. Datos originales normalizados y distancia euclídea.
- Tabla 4. Datos originales y distancia 1-Covarianza.

grupo	número	distancia					desviación	grupo			Desviación por dato				
		espectros	d_min	d_max	media	estándar		más próximo	índice DB	L.o.Max Å	temperatura K media	d.estándar T	%media		
1	10	1.481	45.563	11.589	12.146		2	2,083	3.717	7.799	141		1,459		
2	2.485	1.182	340.242	30.818	32.896		3	3,889	4.382	6.635	371		1,480		
3	1.969	1.182	224.149	36.441	32.253		2	4,420	5.134	5.653	219		1,197		
4	2.654	1.646	316.815	47.910	42.929		3	4,420	6.187	4.715	359		1,210		
5	1.324	1.503	448.378	32.255	42.477		3	2,702	8.541	3.402	175		1,437		
	8.442			37.706	37.732			3,503		5.297	312		1,291		

Tabla 2. Datos originales y distancia euclidiana.

grupo	número	distancia					desviación	grupo			Desviación por dato				
		espectros	d_min	d_max	media	estándar		más próximo	índice DB	L.o.Max Å	temperatura K media	d.estándar T	%media		
1	10	1,49	8,93	2,85	2,11		2	0,496	3.717	7.799	141		0,174		
2	2.485	1,57	63,21	8,28	4,40		3	0,842	4.382	6.635	371		0,286		
3	1.969	0,88	14,00	3,98	1,48		4	0,842	5.134	5.653	219		0,115		
4	2.654	1,41	42,04	6,20	2,84		3	0,814	6.187	4.715	359		0,181		
5	1.324	1,22	26,51	7,01	3,91		4	0,633	8.541	3.402	175		0,263		
	8.442			6,42	3,34			0,726		5.297	312		0,203		

Tabla 3. Datos originales normalizados y distancia euclidiana.

grupo	número espectros	distancia			desviación estándar	grupo más próximo	índice DB	L.o.Max Å	temperatura K		Desviación por dato %media
		d_min	d_max	media					media	d.estándar T	
1	10	0,00	0,07	0,01	0,02	2	0,291	3.717	7.799	141	0,000
2	2.485	0,00	0,70	0,05	0,04	1	0,291	4.382	6.635	371	0,000
3	1.969	0,00	0,35	0,04	0,03	4	0,306	5.134	5.653	219	0,000
4	2.654	0,00	0,51	0,07	0,06	3	0,306	6.187	4.715	359	0,000
5	1.324	0,00	0,46	0,03	0,03	4	0,271	8.541	3.402	175	0,000
		8.442		0,05	0,04		0,293		5.297	312	0,000

Tabla 4. Datos originales y distancia 1-Covarianza.

Estas tablas tienen las siguientes columnas:

grupo.

número de espectros clasificados en el grupo.

d.min. Distancia del espectro más próximo al centroide.

d.max. Distancia del espectro más alejado del centroide.

media. Distancia media de los espectros al centroide.

desviación estándar de la distancia de los espectros al centroide.

grupo más próximo. Determinado mediante la distancia entre centroides.

índice DB. Índice Davies Bouldin.

Lo.Max. Longitud de onda del máximo de radiación del espectro (del polinomio ajustado).

temperatura media de los espectros del grupo.

desviación estándar de la temperatura de los espectros incluidos en el grupo.

Desviación por dato (lo llamaremos DPD). Para obtener una idea de la similitud de los espectros con el centroide de su grupo se calcula la desviación media, en tanto por ciento, de cada uno de los datos de los espectros respecto al valor correspondiente del centroide.

Las distancias no son comparables entre tablas por corresponder a formas de cálculo distintas, pero si es comparable la desviación porcentual por dato (DPD). Como cabía esperar, normalizar los datos mejora sustancialmente el DPD ya que sólo aparecerán las diferencias en la forma del espectro no es un su valor absoluto y usar la covarianza lo mejora aún más porque las diferencias se deben fundamentalmente a la posición del máximo que es quien determina la temperatura que es el criterio de clasificación Harvard.

Las figuras 14 muestra esto mismo a través de las distancias entre centroides (separación) y la distancia media de los espectros de cada grupo respecto a su centroide (cohesión). Cuanto menor es la relación entre la cohesión y la distancia al centroide más próximo, mejor es la clasificación, aunque en este caso la clasificación es siempre la misma, lo que estamos viendo es como se valora la clasificación dependiendo del criterio de similitud que se adopte. Es evidente que la mejor valoración de una clasificación por temperatura es la similitud por temperatura (1-covarianza).

Estas comparaciones serán útiles cuando clasifiquemos mediante K-means.

En las figuras 15 a 20 se muestran los centroides que surgen de la clasificación Harvard.

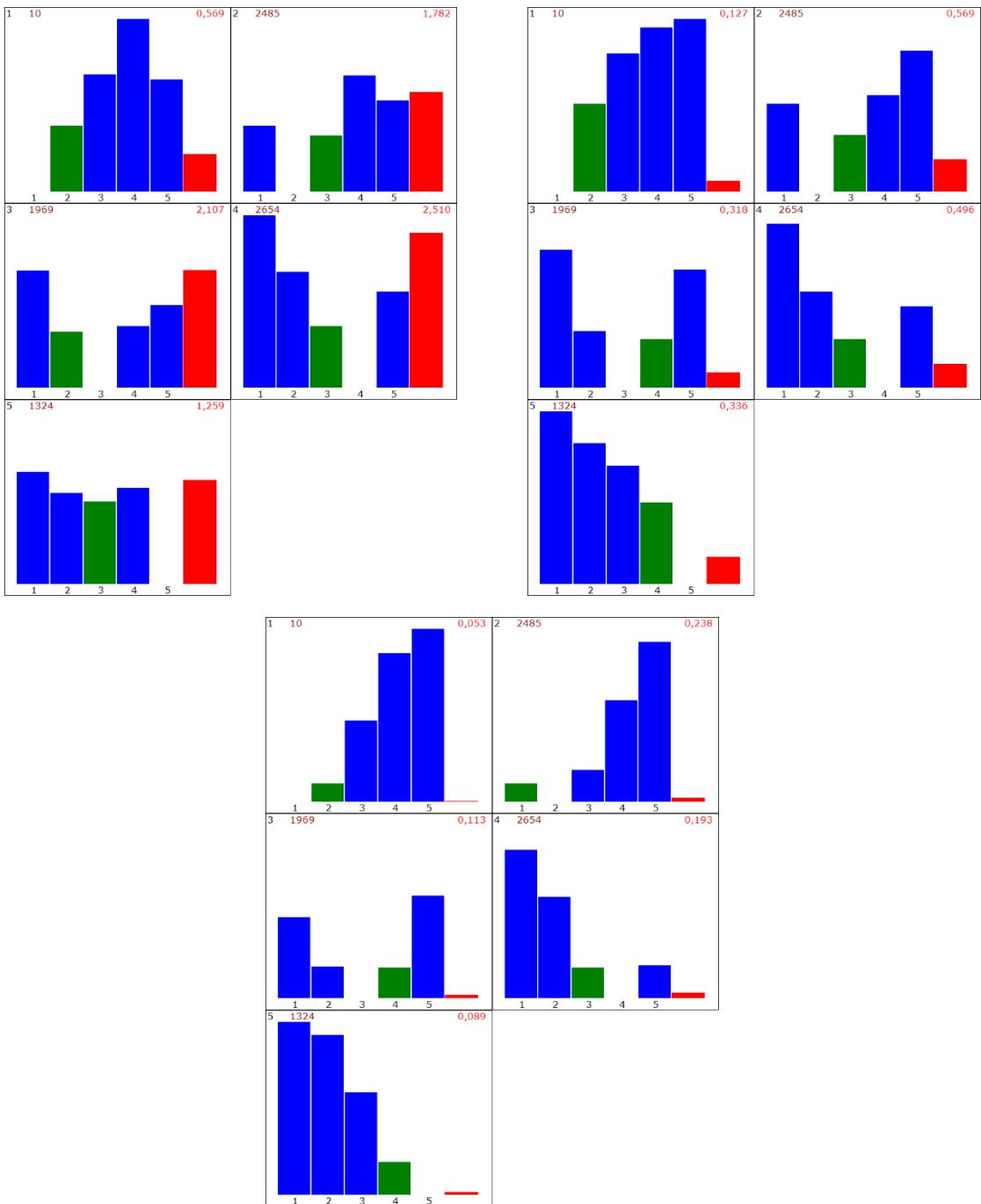


Figura 14. Para cada grupo (1 al 5): distancia del centroide a los demás centroides (azul y verde- el más cercano-) y distancia media de los espectros en su grupo respecto a él (rojo).. El número en rojo (arriba a la derecha) es la relación entre el histograma en rojo y el verde, cuanto menor más “separada” es la clasificación. Arriba a la izquierda para datos originales-euclidiana, a la derecha datos normalizados-euclidiana y abajo para datos originales-(1-covarianza).

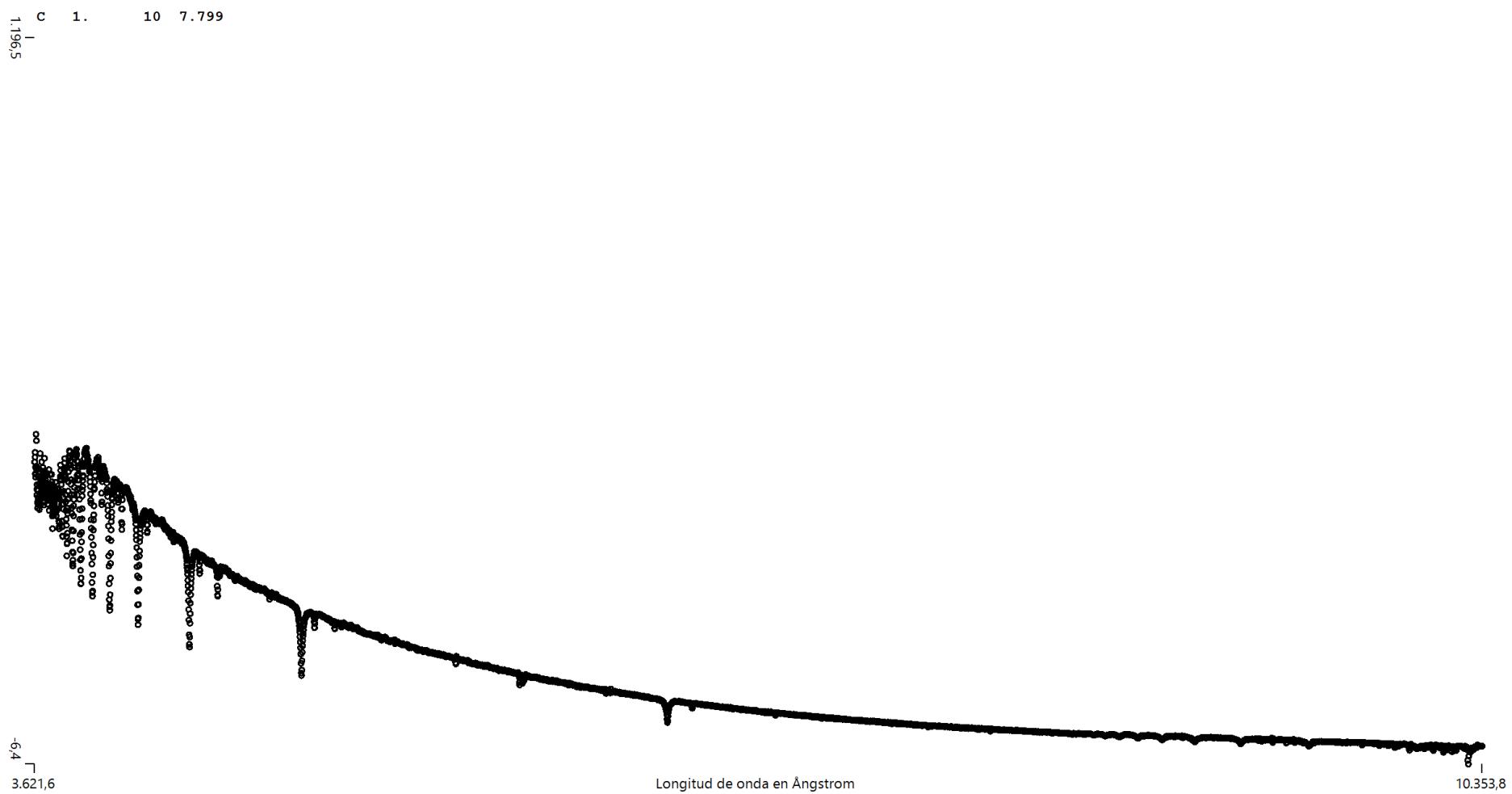


Figura 15. Centroide del grupo 1. A en la clasificación Harvard.

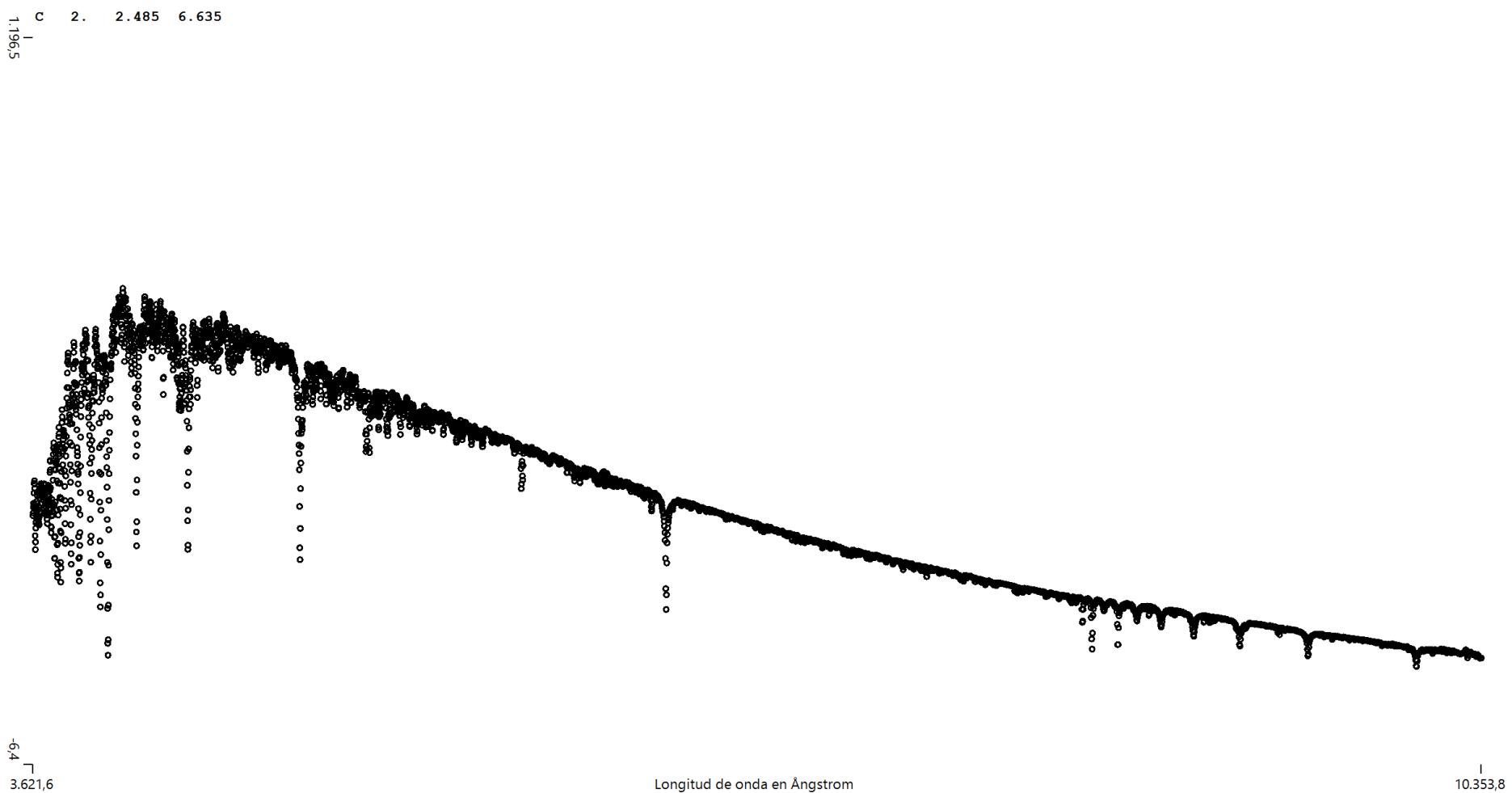


Figura 16. Centroide del grupo 2. F en la clasificación Harvard.

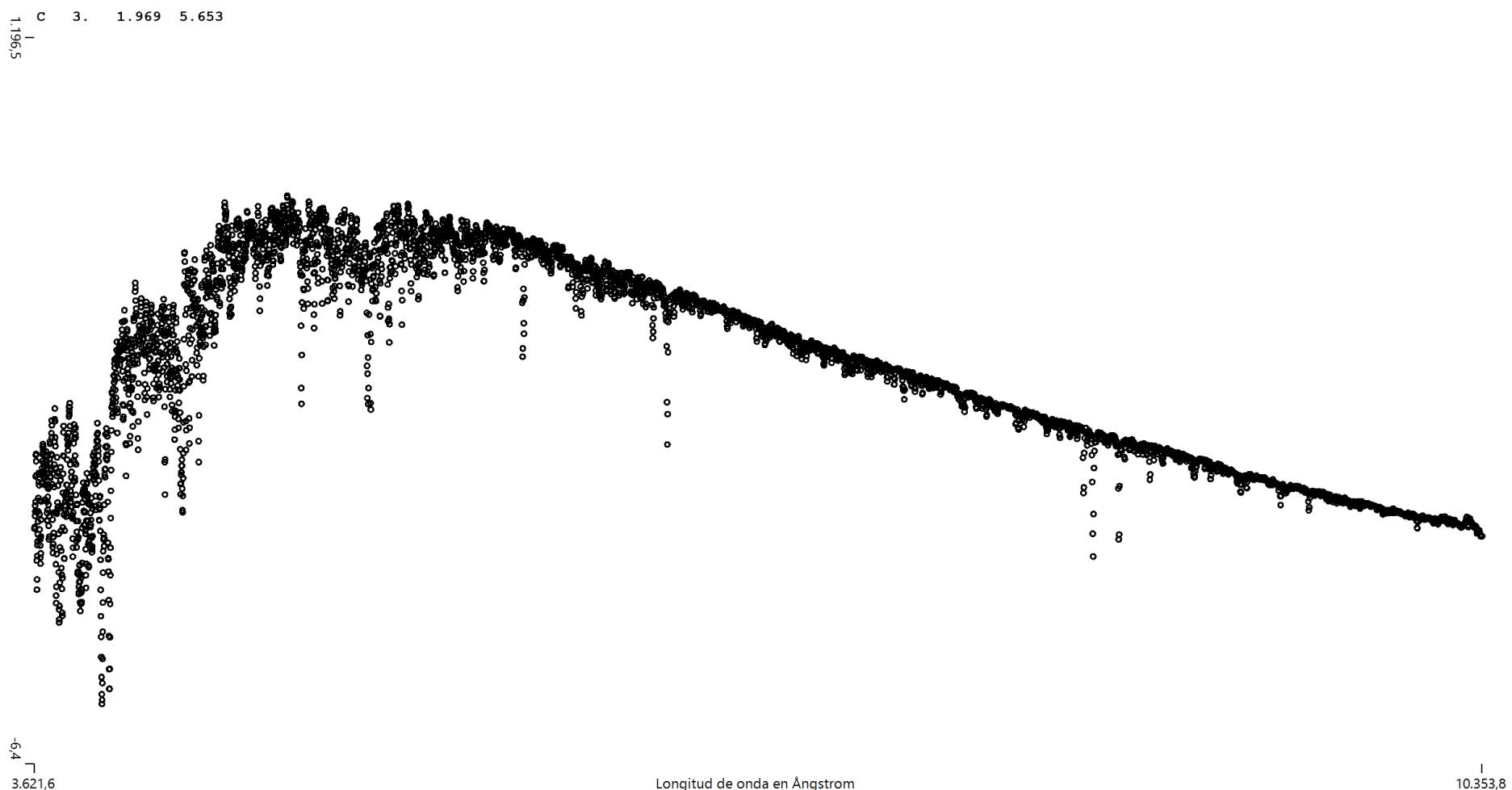


Figura 17. Centroide del grupo 2. G en la clasificación Harvard.

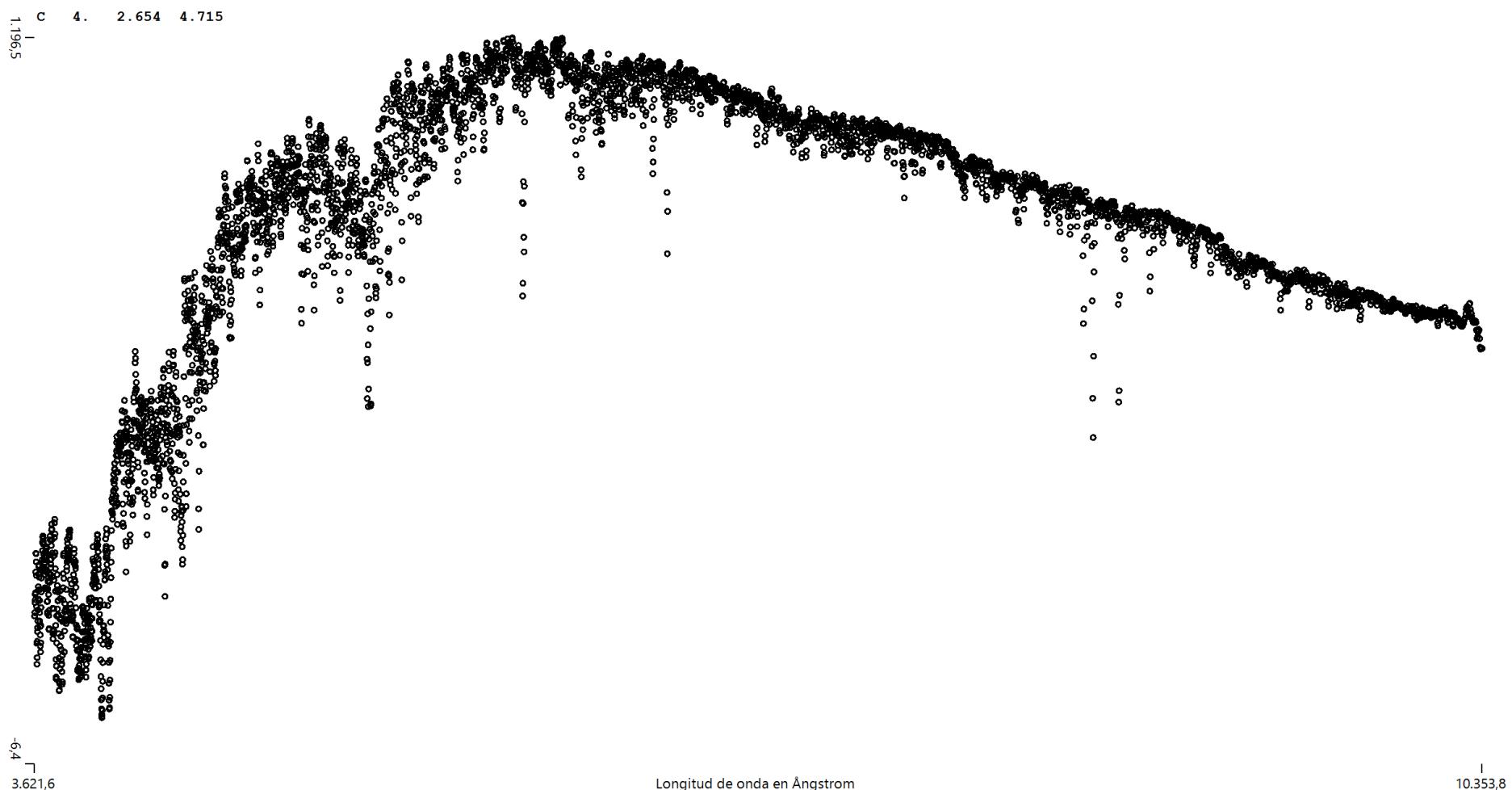


Figura 18. Centroide del grupo 2. K en la clasificación Harvard.

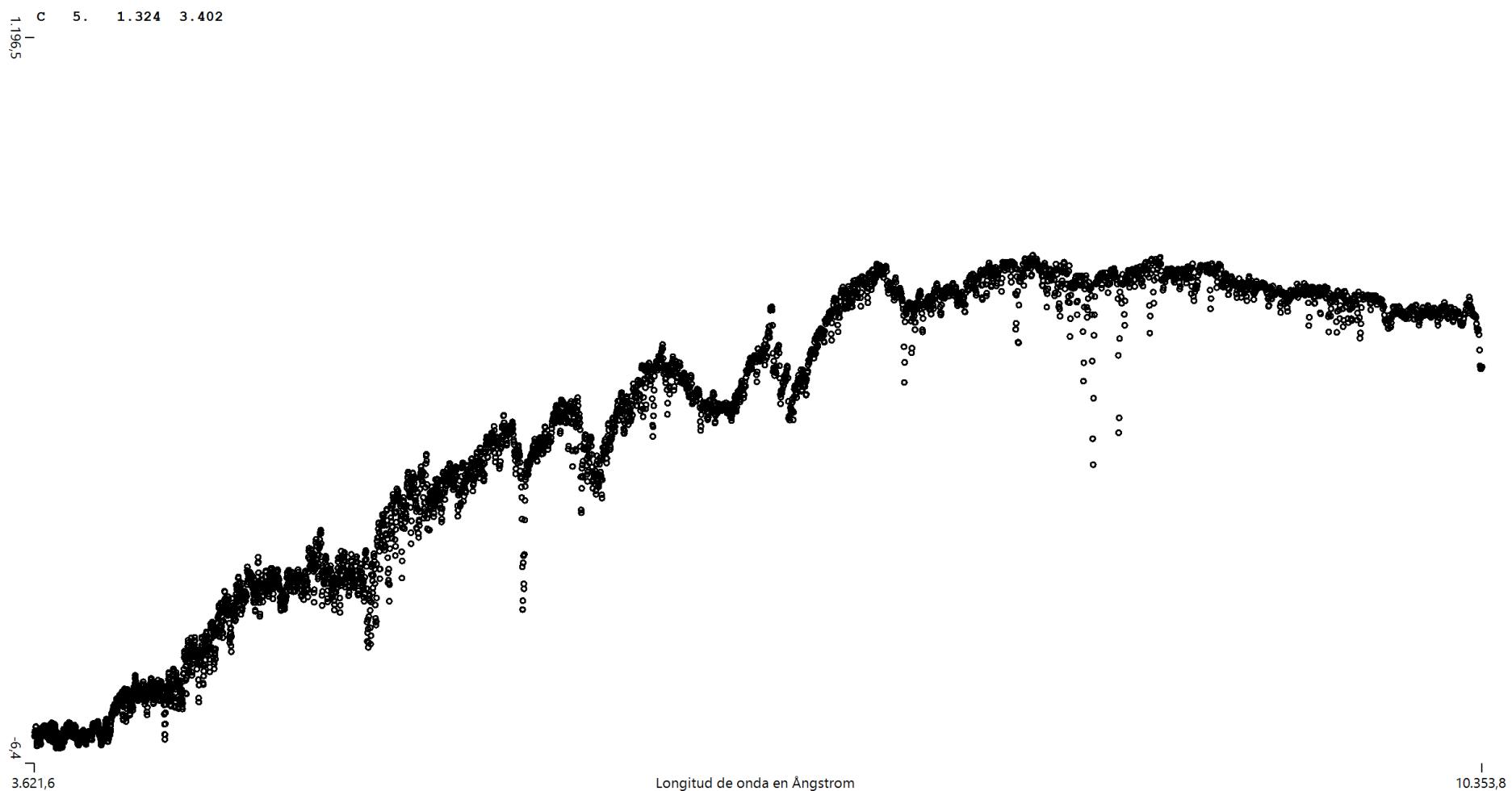


Figura 19. Centroide del grupo 2. M en la clasificación Harvard.

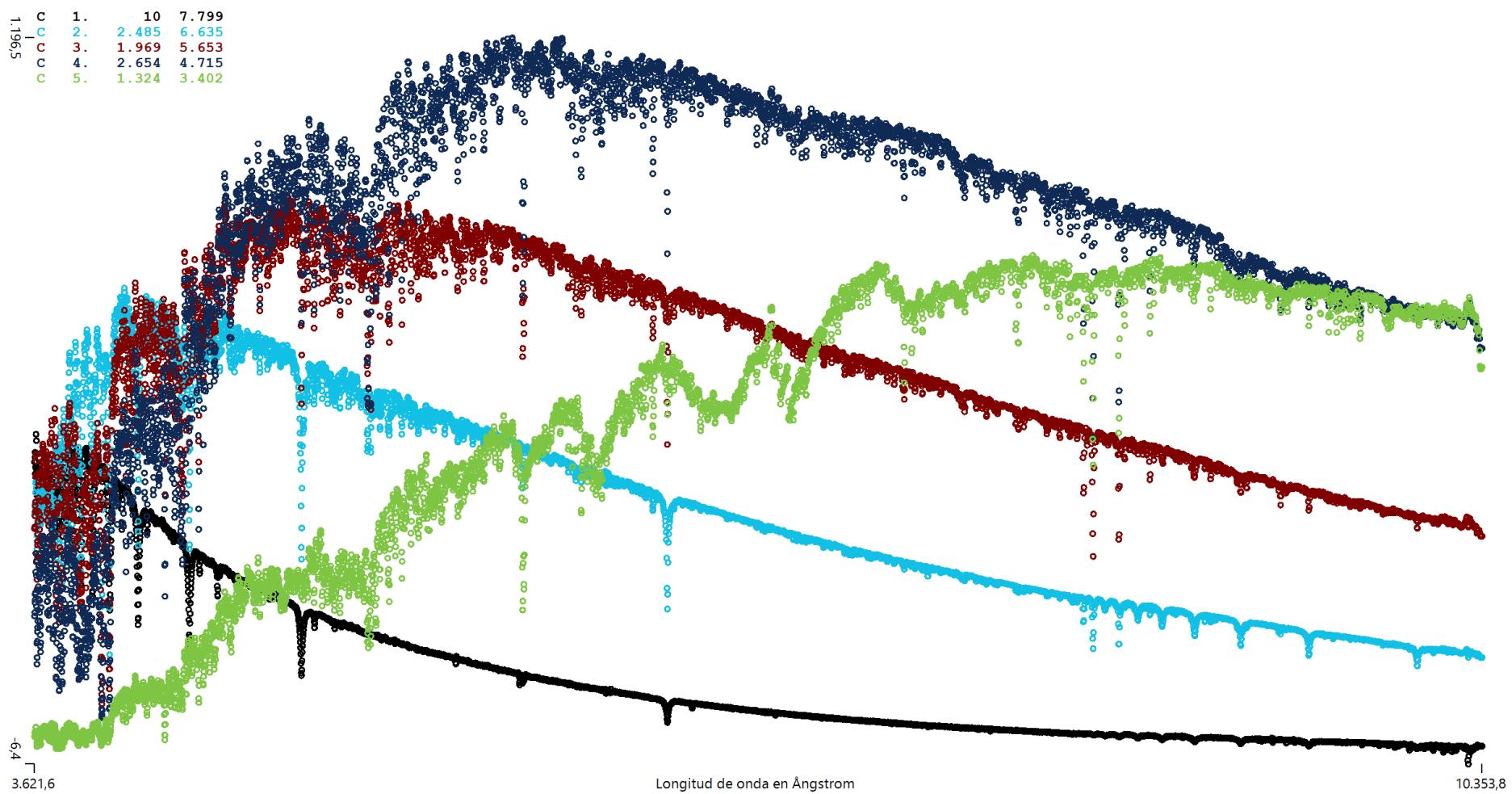


Figura 20. Centroides de la clasificación Harvard.

Una vez establecida la referencia y mostrados los elementos de evaluación pasamos a clasificar los 8.442 espectros mediante K-means. Los parámetros usados en todos los casos han sido los siguientes:

- Número de grupos (clústeres) = 5.
- Número de espectros = 8.442.
- Número de datos por espectro = 4.563.
- Selección inicial de grupos por el método Kmeans++.
- Número de simulaciones Monte Carlo = 1.000

Las tablas 5, 6 y 7 muestran las clasificaciones en los mismos escenarios de datos y distancias usados en la clasificación Harvard:

- Tabla 5. Datos originales y distancia euclidiana.
- Tabla 6. Datos originales normalizados y distancia euclidiana.
- Tabla 7. Datos originales y distancia 1-Covarianza.

grupo	número espectros	distancia			desviación estándar	grupo más próximo			índice L.o.Max			temperatura K		Desviación por dato %media
		d_min	d_max	media		DB	Å	media	d.estándar T					
1	552	5.373	127.847	39.057	20.721	3	0,896	5.535	5.372	834			0,382	
2	4.667	847	32.025	7.866	4.157	4	0,974	5.788	5.361	1.292			0,937	
3	1.032	3.899	121.871	24.765	14.563	4	0,753	5.666	5.287	910			0,432	
4	1.961	2.393	60.095	14.821	8.012	2	0,934	5.781	5.217	986			0,525	
5	230	10.816	281.712	58.764	38.090	1	0,974	6.568	4.535	756			0,371	
	8.442			14.974	10.863		0,906		5.297		1.145		0,513	

Tabla 5. Datos originales y distancia eucladiana.

grupo	número espectros	distancia			desviación estándar	grupo más próximo			índice L.o.Max			temperatura K		Desviación por dato %media
		d_min	d_max	media		DB	Å	media	d.estándar T					
1	833	1,08	16,29	5,50	2,55	2	0,568	4.130	7.019	143			0,245	
2	2.378	1,04	63,87	4,90	2,48	3	0,629	4.638	6.270	379			0,150	
3	2.754	1,07	16,55	4,67	1,90	2	0,729	5.581	5.214	335			0,134	
4	1.488	1,71	39,75	5,44	2,29	3	0,729	6.953	4.222	468			0,166	
5	989	0,83	26,14	5,72	3,34	4	0,689	8.673	3.349	159			0,228	
	8.442			5,07	2,41		0,669		5.297		347		0,161	

Tabla 6. Datos originales normalizados y distancia eucladiana.

grupo	número espectros	grupo						Desviación por dato		
		distancia			desviación estándar	más próximo	índice	L.o.Max	temperatura K	d.estándar T
		d_min	d_max	media			DB	Å	media	
1	1.835	0,00	0,70	0,04	0,03	2	0,270	4.256	6.819	260
2	2.022	0,00	0,24	0,03	0,03	1	0,321	4.935	5.881	229
3	1.952	0,00	0,37	0,03	0,03	4	0,321	5.685	5.109	248
4	1.264	0,00	0,48	0,03	0,03	3	0,317	6.510	4.477	318
5	1.369	0,00	0,47	0,03	0,03	4	0,270	8.510	3.416	187
	8.442			0,03	0,03		0,300		5.297	250
										0,000

Tabla 7. Datos originales y distancia 1-Covarianza.

Como no podía ser de otra forma, las clasificaciones K-means presentan mejor valoración que las correspondiente Harvard (tabla 8).

Datos	Distancia	DPD	
		Harvard	K-means
Originales	Euclíadiana	1,291	0,513
Normalizados	Euclíadiana	0,203	0,161
Originales	1-Covarianza	1,7E-06	1,1E-06

Tabla 8. Desviación porcentual por dato (DPD).

Esta mejor valoración significa que los espectros en cada grupo son más parecidos al centroide (e indirectamente entre ellos) en la clasificación K-means que en la de Harvard.

La mejora de las métricas por sí sola no es garantía de mejor clasificación, primero hay que comprobar que las diferencias entre grupos responden a diferencias reales y significativas entre las estrellas, a este respecto la clasificación Harvard es clara: distintos grupos significan distintas temperaturas, mientras que las clasificaciones obtenidas con K-means son muchos los espectros que con una misma temperatura se asignan a grupos distintos, aunque normalmente vecinos.

Empezando el análisis de los resultados, La tabla 5 pone de manifiesto que clasificar usando los datos originales y la distancia euclíadiana no es un buen criterio porque no discrimina adecuadamente entre espectros y clasifica a más de la mitad (4.667) en el mismo grupo, por lo que no profundizaremos más en este escenario.

Por su parte, el escenario de datos originales y distancia igual a (1-covarianza) debería ser el que más se aproxime a la clasificación Harvard por tener la localización del máximo de radiación (temperatura) como elemento esencial de clasificación. La primera diferencia que se observa es el rango de temperaturas en cada grupo, fijado de antemano en Harvard pero no en K-means. En la figura 21 se comparan estos rangos representados por el valor medio de la temperatura en cada grupo y la dispersión en el mismo calculada como la desviación estándar (radio de los círculos). En la tabla 7 ya se ve que la desviación estándar global de la temperatura es inferior en K-means (250) que en Harvard (312), lo cual es lógico porque K-means tiene la “libertad” de fijar el centro de los intervalos y su anchura y así reducir la desviación estándar. La desviación estándar de la temperatura por grupos no es comparable porque el rango de temperaturas de cada grupo en ambas clasificaciones es distinto y a más rango cabe esperar mayor desviación estándar.

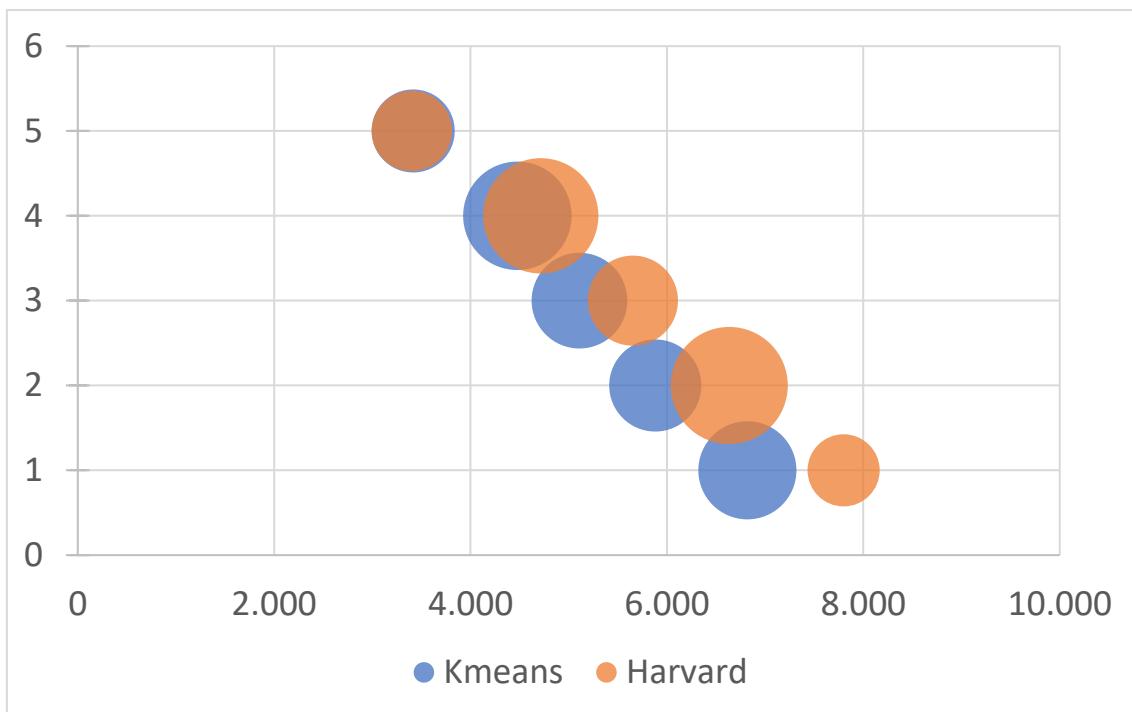


Figura 21. Temperatura media y dispersión de esta dentro de cada grupo, para las clasificaciones K-means y Harvard. El radio de los círculos es la desviación estándar de la temperatura

La clasificación K-means es más regular, tanto en temperatura media como en dispersión que la clasificación Harvard. Cosa que también se refleja en las figuras 23 y 24 que comparan la cohesión y dispersión de las clasificaciones Harvard y K-means, ambas son similares pero K-means es más homogénea con índices parecidos para los 5 grupos.

Pero a pesar de que la dispersión de temperatura (medida por la desviación estándar) es inferior en la clasificación K-means que en la clasificación Harvard, ocurre que K-means no conduce a grupos disjuntos de temperatura, como puede verse en la figura 22. Como ya se ha dicho, la clave para decidir entre ambas clasificaciones es saber si aquello que hace semejante a los espectros en una clasificación K-means es más importante que la semejanza en temperatura.

Las figuras 25 a 31 muestran los centroides de la clasificación con los datos originales mediante distancia igual a 1-covarianza. Vemos que es difícil encontrar líneas de absorción que no estén presentes en todos los centroides, en el anexo III se marcan las líneas del H, del He y de otros elementos presentes en casi todos los centroides. El grupo 5 es el más singular, su forma sugiere que acumula un gran número de estrellas variables.

G	1.	1.876
G	2.	2.068
G	3.	2.030
G	4.	1.283
G	5.	1.185

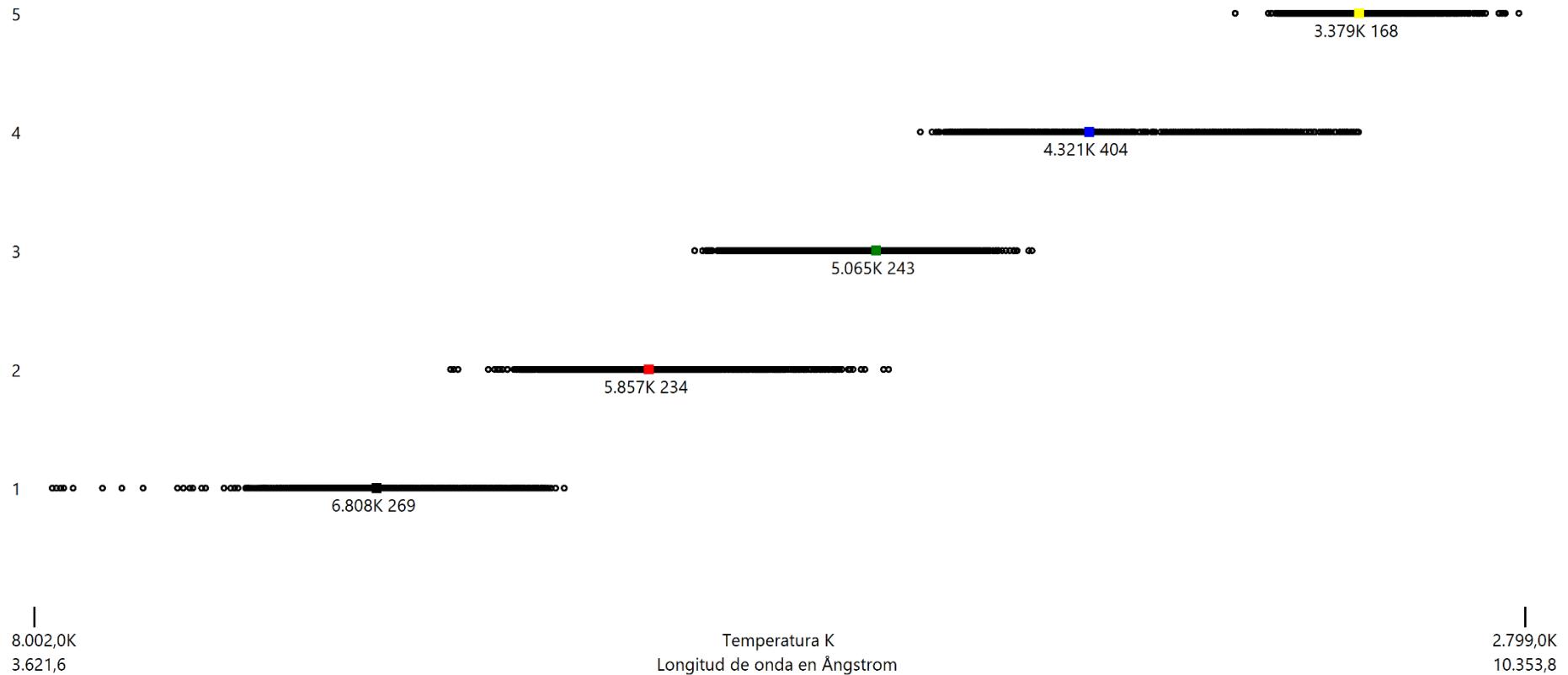


Figura 22. Distribución de los espectros resultante de clasificar los datos originales mediante distancia igual a 1-covarianza.

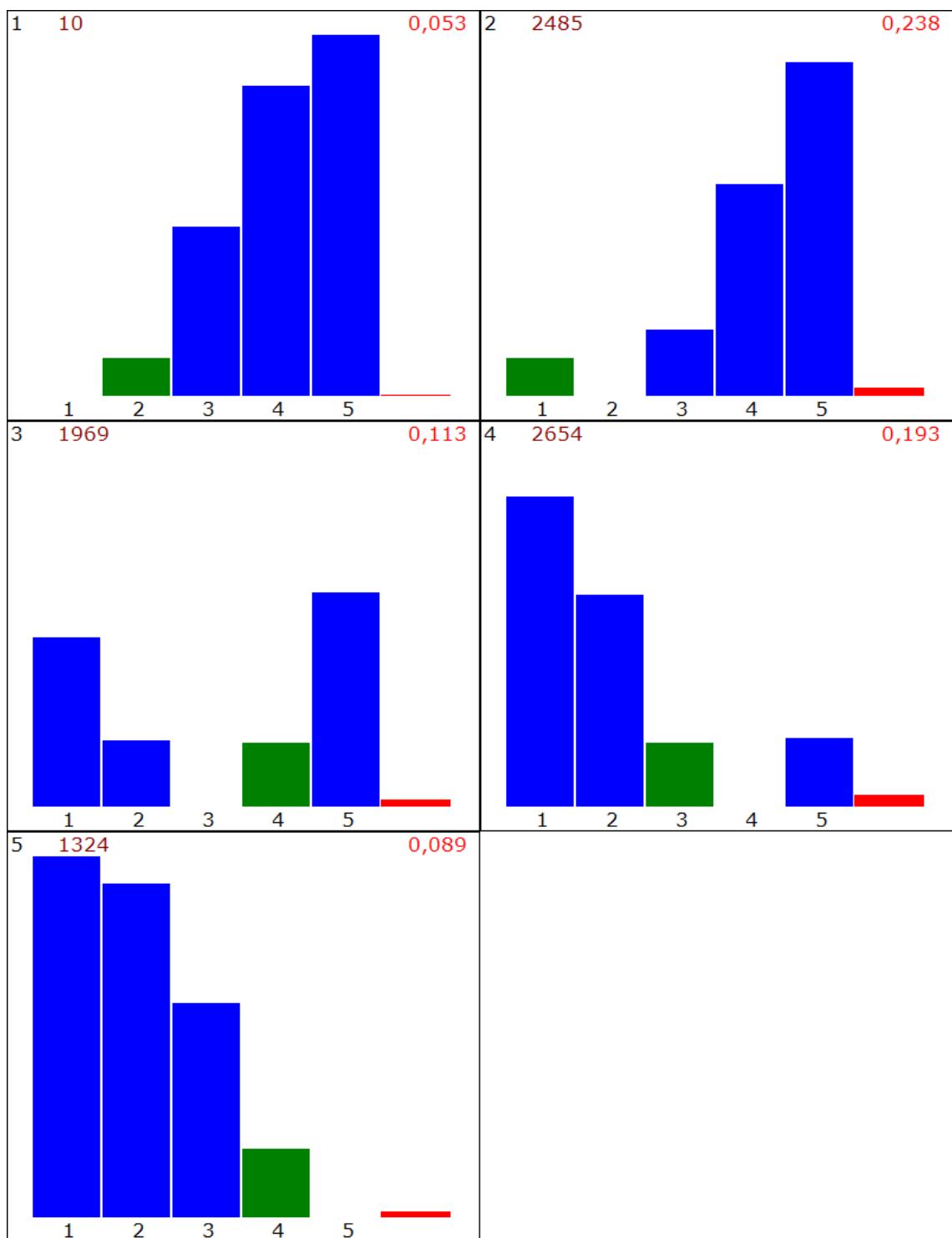


Figura 23. Coherencia y dispersión de la clasificación Harvard.

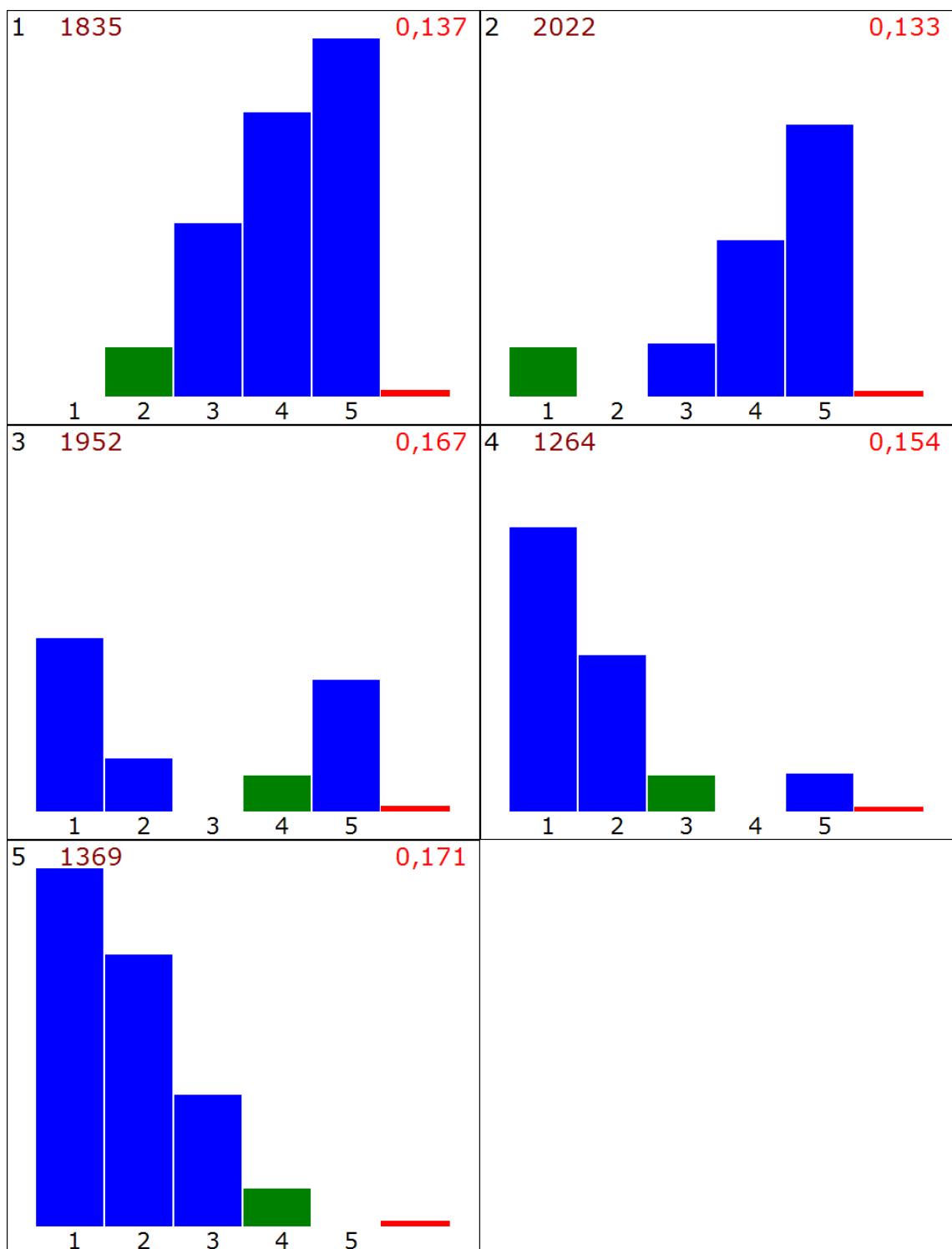


Figura 24. Coherencia y dispersión de la clasificación K-means con Datos originales y distancia = (1-covariancia).

C 1. 1.835 6.819
1.220,0

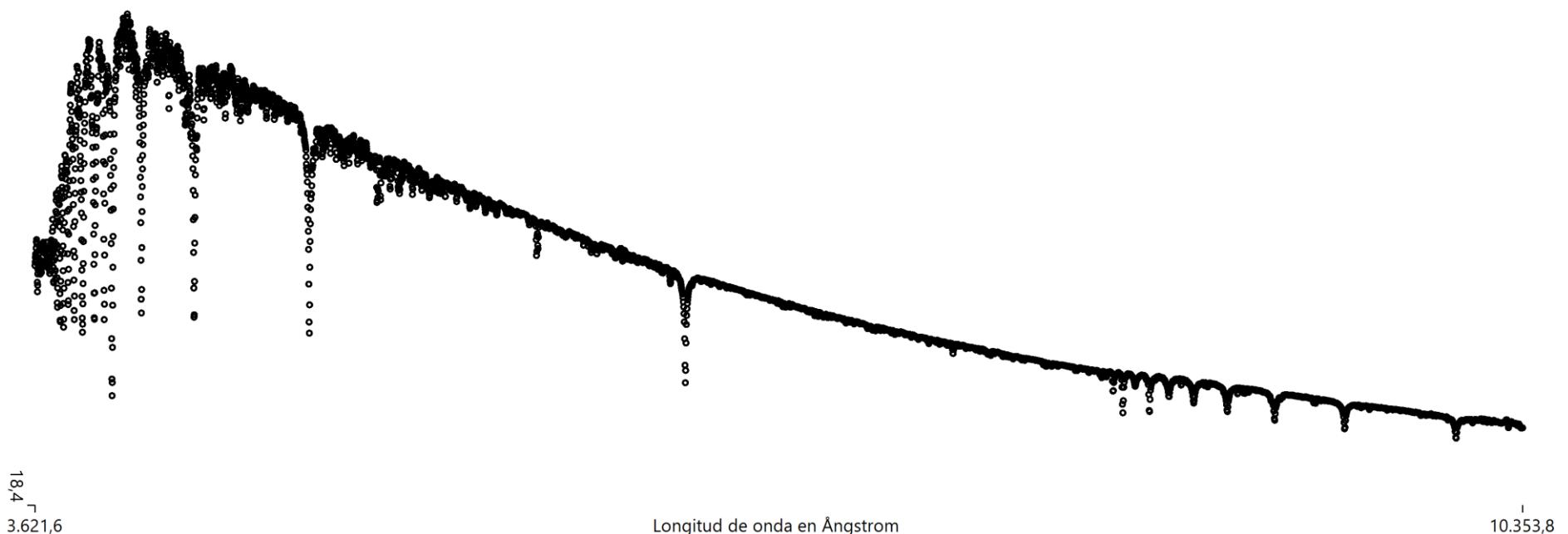


Figura 25. Centroide del grupo 1 de la clasificación K-means con datos originales y distancia = 1-covarianza.

C 2. 2.022 5.881
1.220,0

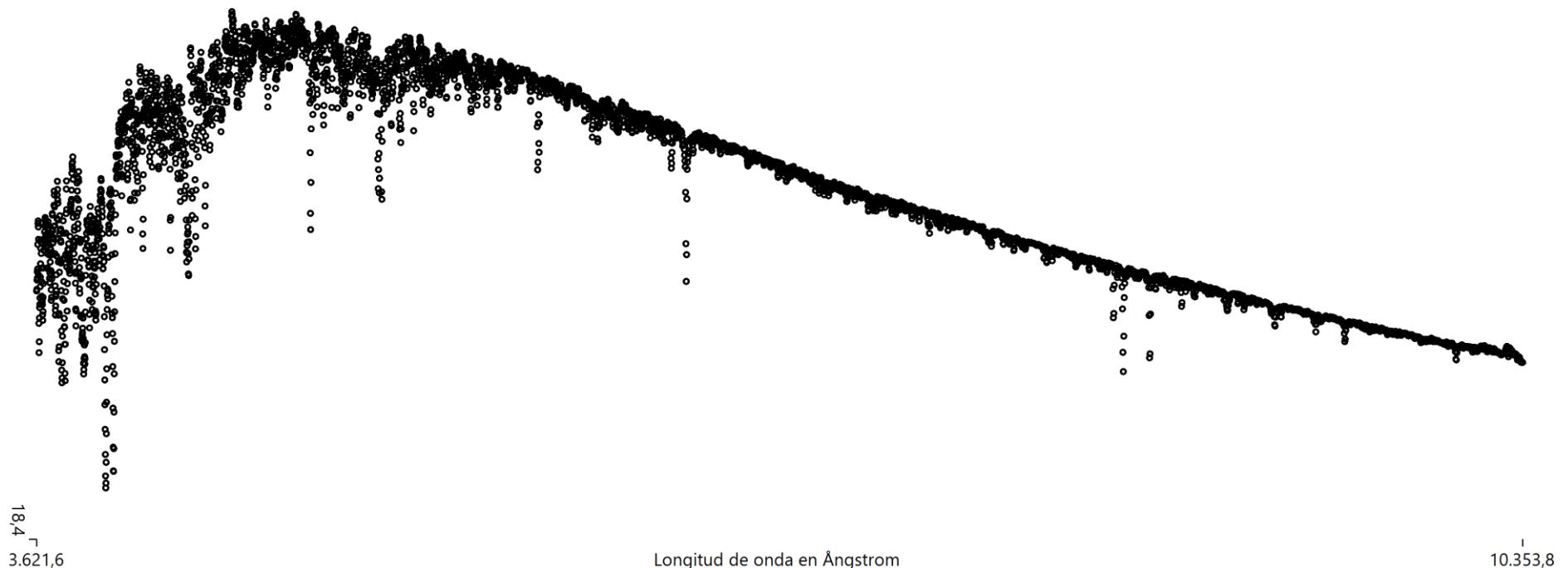


Figura 26. Centroide del grupo 2 de la clasificación K-means con datos originales y distancia = 1-covarianza.

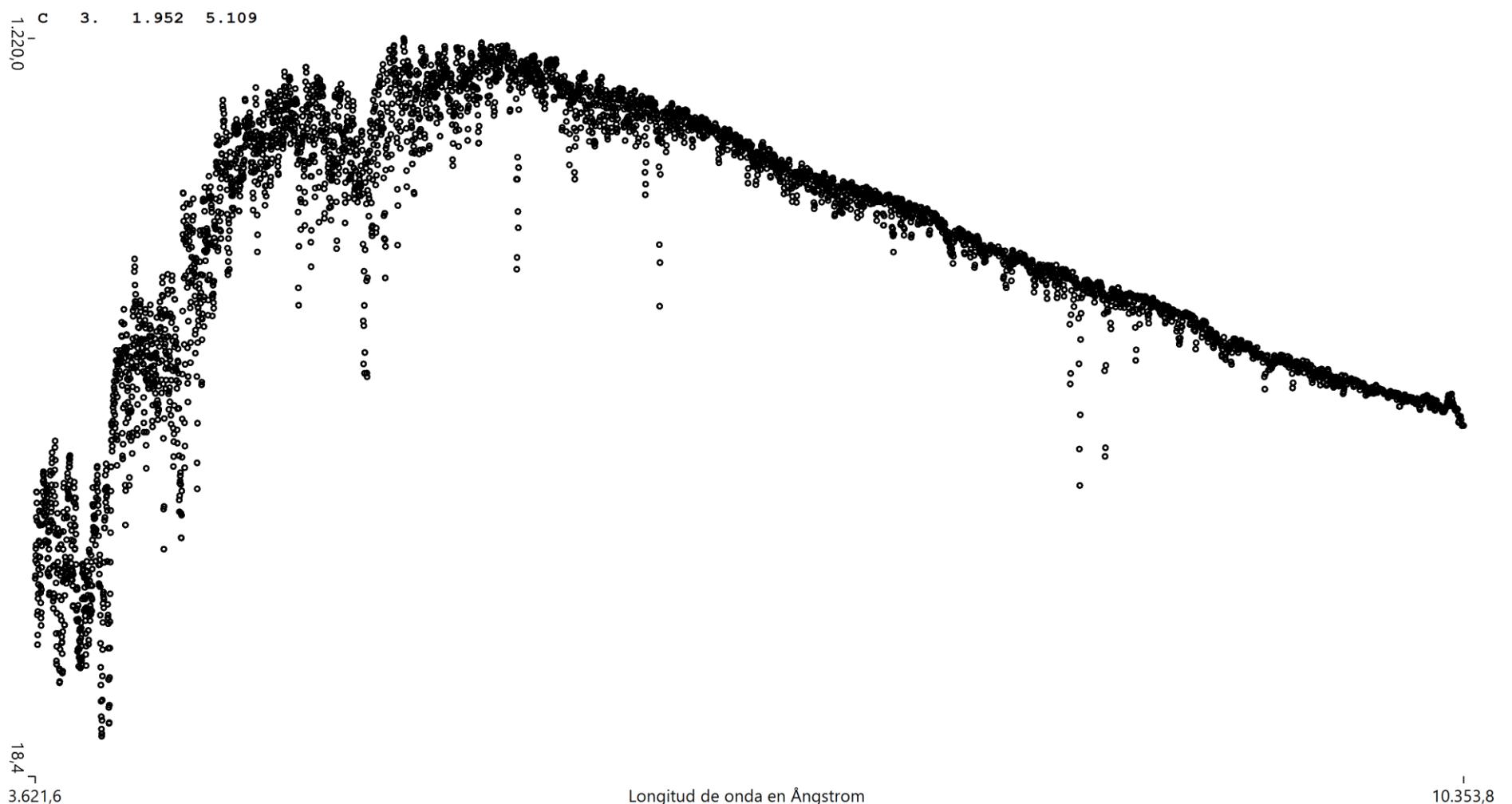


Figura 27. Centroide del grupo 3 de la clasificación K-means con datos originales y distancia = 1-covarianza.

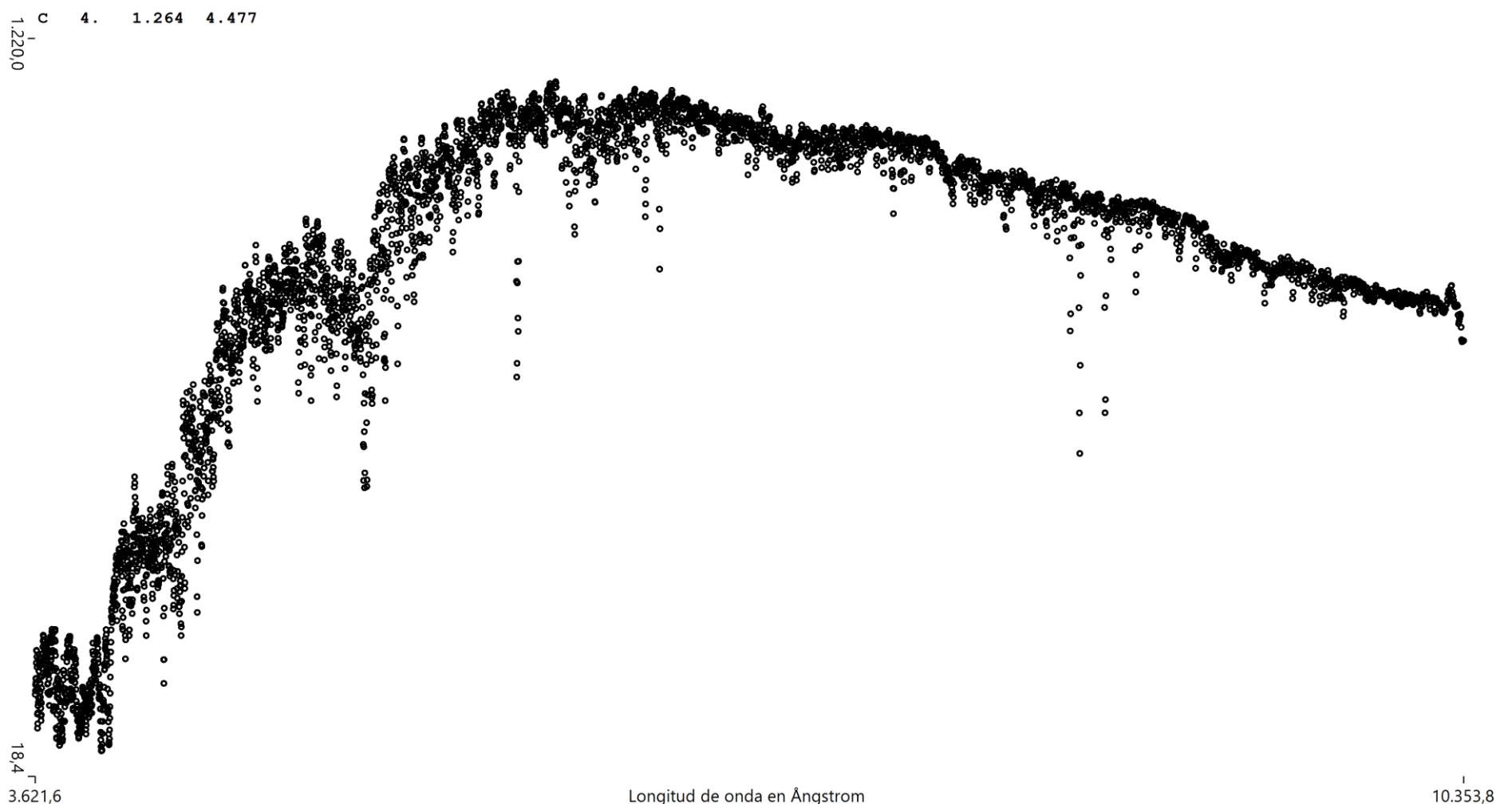


Figura 28. Centroide del grupo 4 de la clasificación K-means con datos originales y distancia = 1-covarianza.

C 5. 1.369 3.416
1.220,0

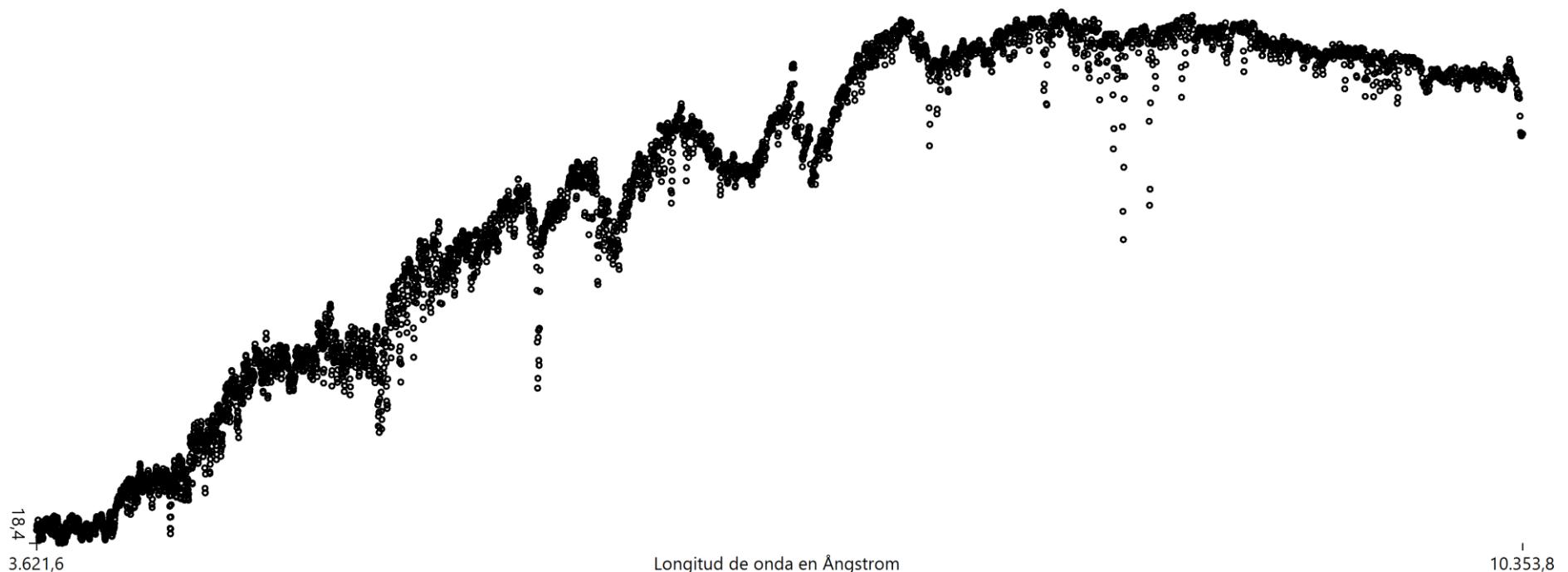


Figura 29. Centroide del grupo 5 de la clasificación K-means con datos originales y distancia = 1-covarianza.

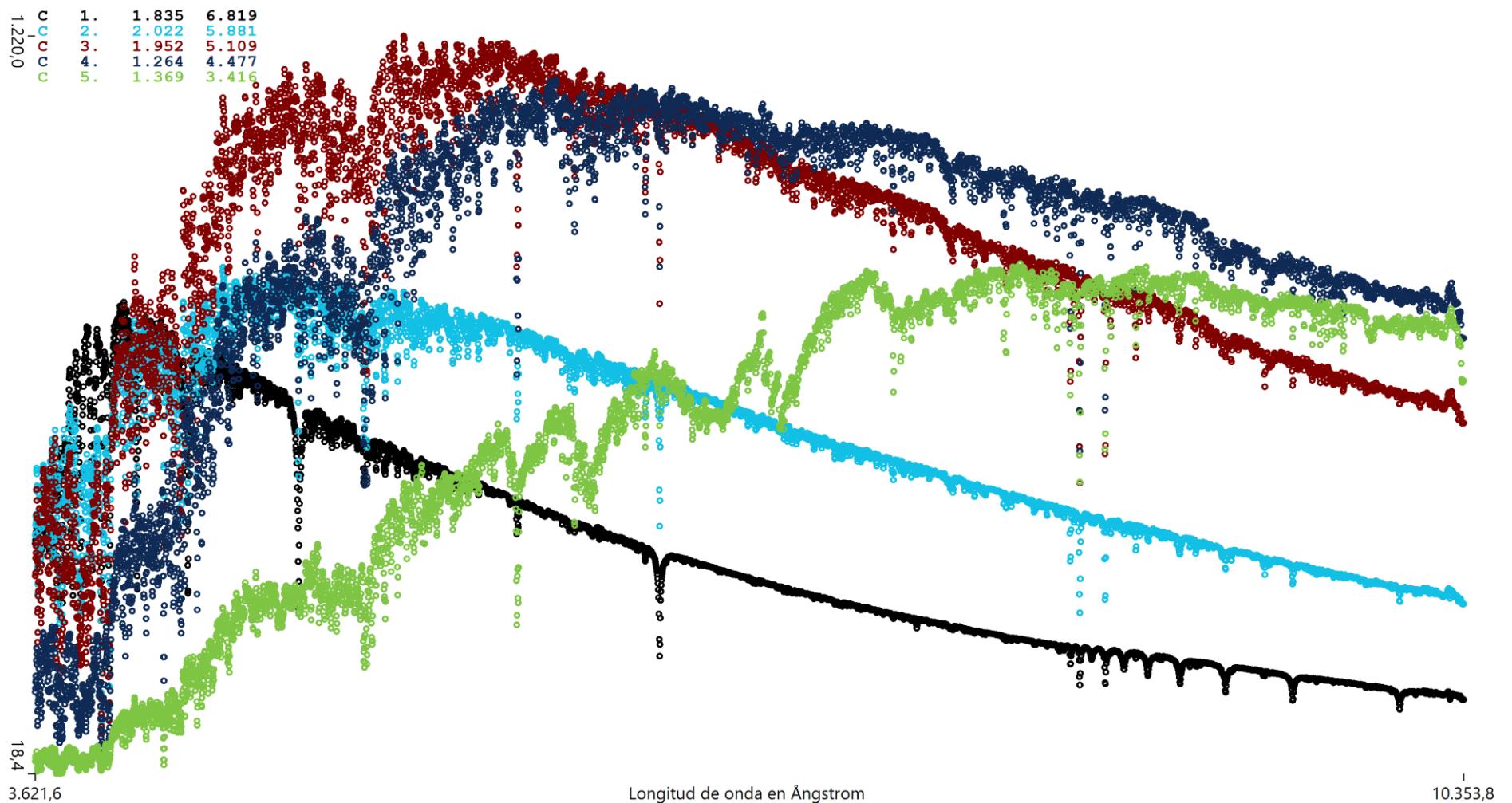


Figura 30. Centroides de la clasificación K-means con datos originales y distancia = 1-covarianza.

La figura 31 compara los centroides Harvard y K-means, visualmente son similares pero se aprecian diferencias, los centroides K-mean están más separados.

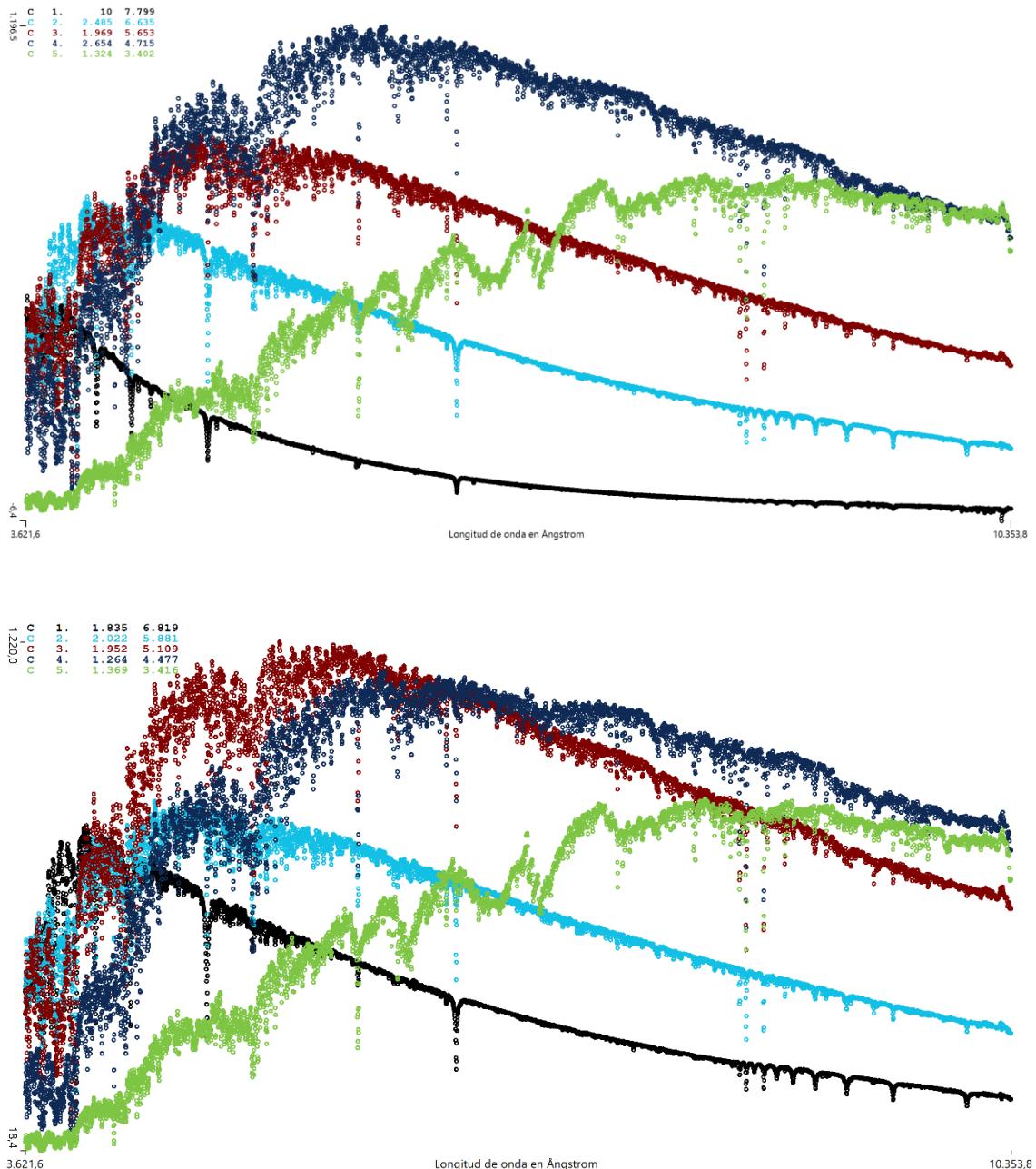


Figura 31. Centroides de la clasificación Harvard (arriba) frente a los centroides K-means con datos originales y distancia = 1-covarianza.

La clasificación K-means con datos normalizados y distancia euclíadiana presenta algunas diferencias importantes respecto de la clasificación con distancia igual a 1-covarianza (figuras 32 y 33) y las diferencias con la clasificación Harvard se acentúan al tomarse en consideración la forma total del espectro y no sólo el máximo de este.

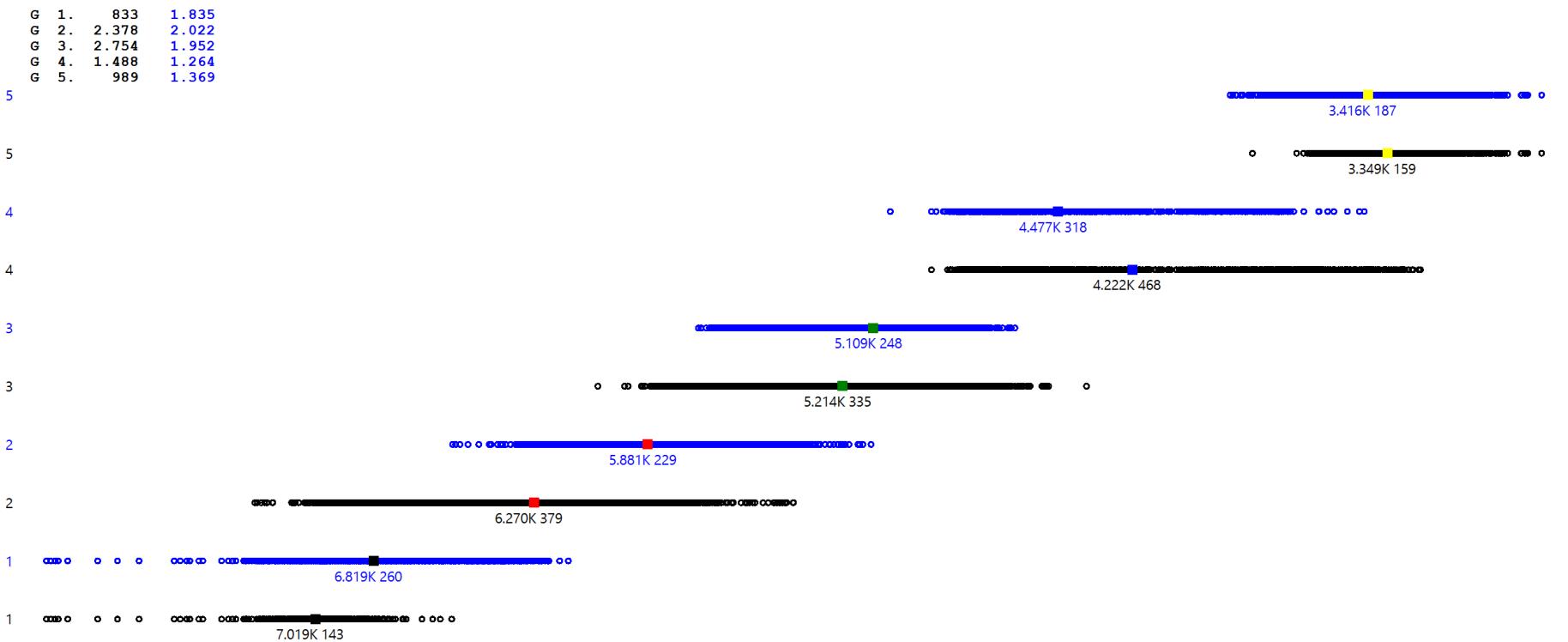


Figura 32. Comparación de las clasificaciones K-means con datos originales y distancia igual a (1- covarianza) (azul) con la clasificación con datos normalizados y distancia euclíadiana (negro).

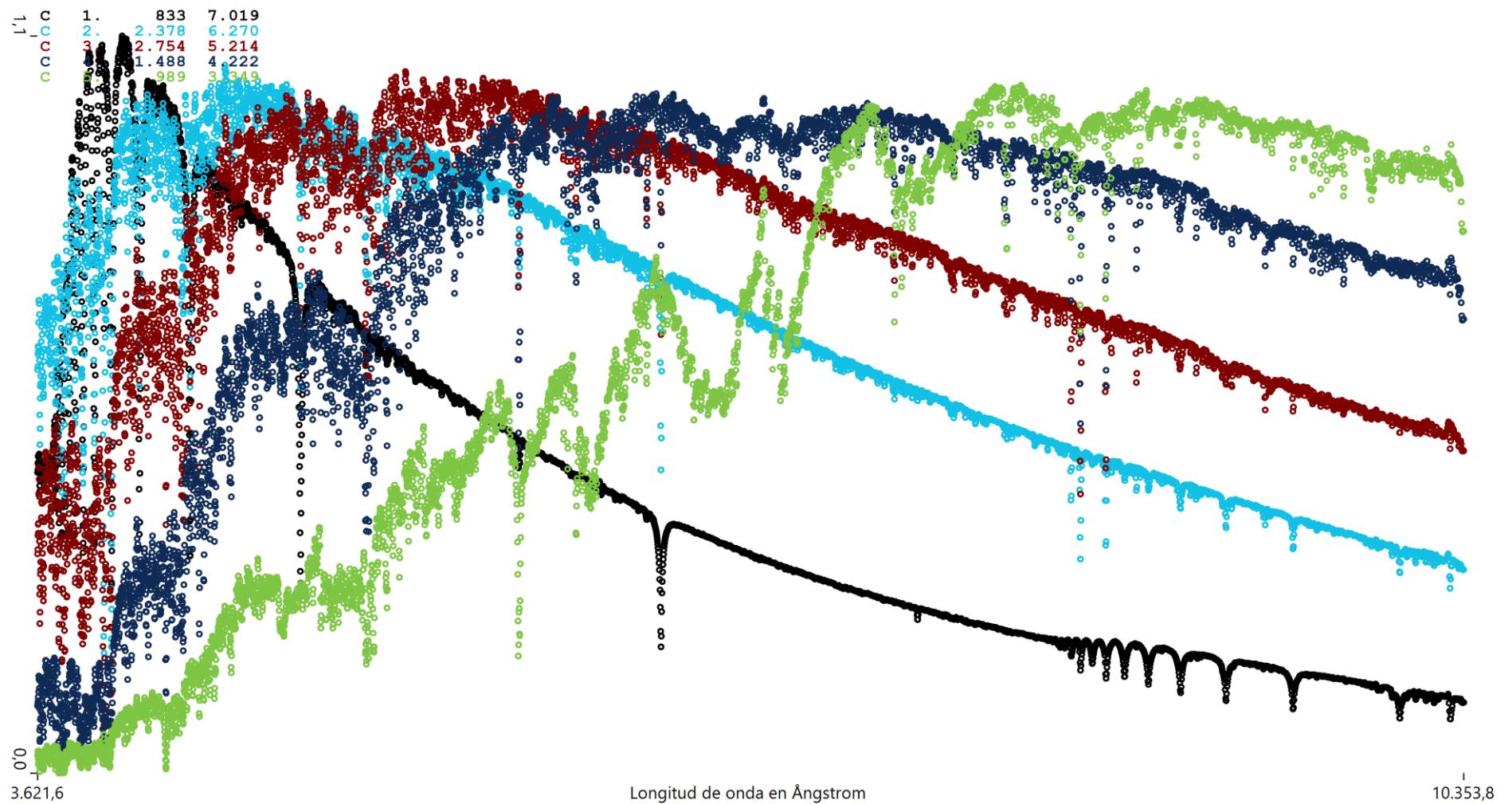


Figura 33. Centroides de la clasificación K-means con datos normalizados y distancia euclidiana.

Para conocer el origen de las diferencias entre espectros de la misma temperatura asignados a grupos distintos hemos calculado el espectro medio de dos conjuntos de espectros dentro del mismo rango de temperaturas pero clasificados en grupos diferentes. En las figuras 34 a 37 se comparan para la clasificación K-means con datos originales y distancia igual a 1-covarianza y en las 38 a 41 para K-means con datos normalizados y distancia euclidiana.

Salvo en el grupo 5 (estrellas variables), las diferencias entre los espectros en el mismo rango de temperatura pero asignados a grupos distintos no parecen cualitativamente determinantes, aunque cuantitativamente lo sean y por eso unos espectros pertenecen a un grupo y el resto a otros, en el caso de datos normalizados y distancia euclidiana, las diferencias son aún menos perceptibles.

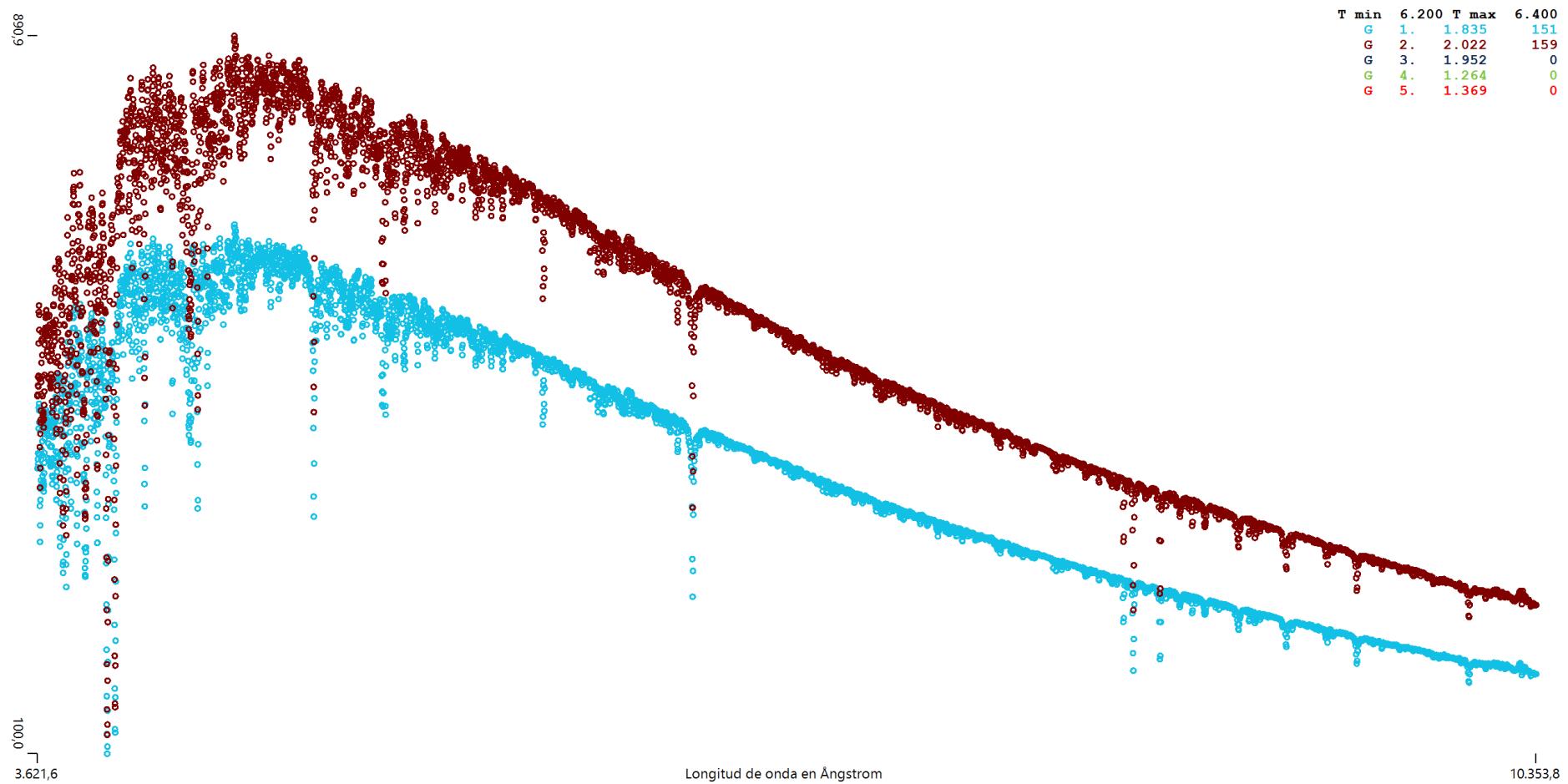


Figura 34. Clasificación K-means con datos originales y distancia igual a 1-covarianza. Espectros en el rango de temperaturas 6.200-6.400 K asignados a dos grupos distintos.

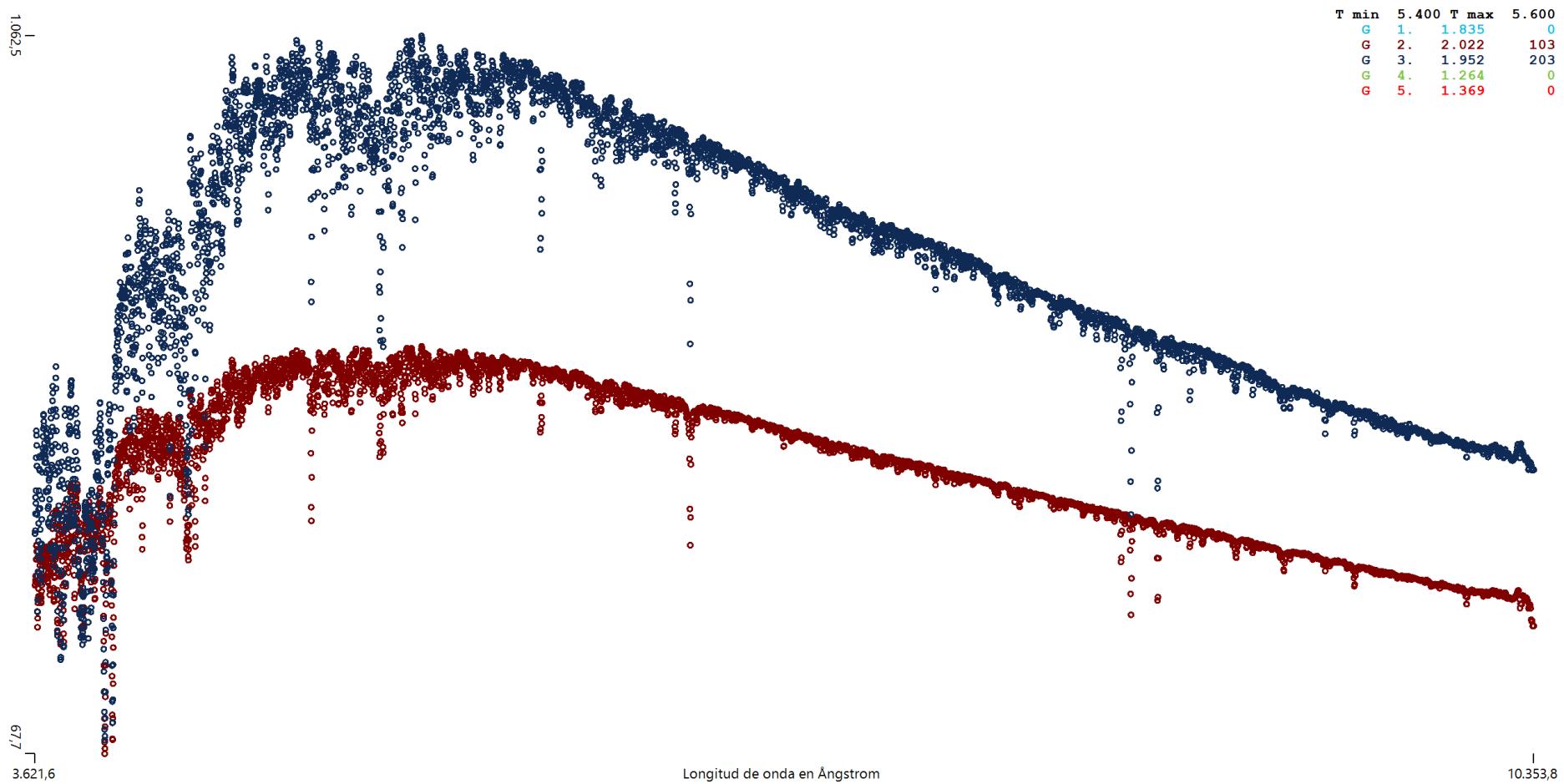


Figura 35. Clasificación K-means con datos originales y distancia igual a 1-covarianza. Espectros en el rango de temperaturas 5.400-5.600 K asignados a dos grupos distintos.

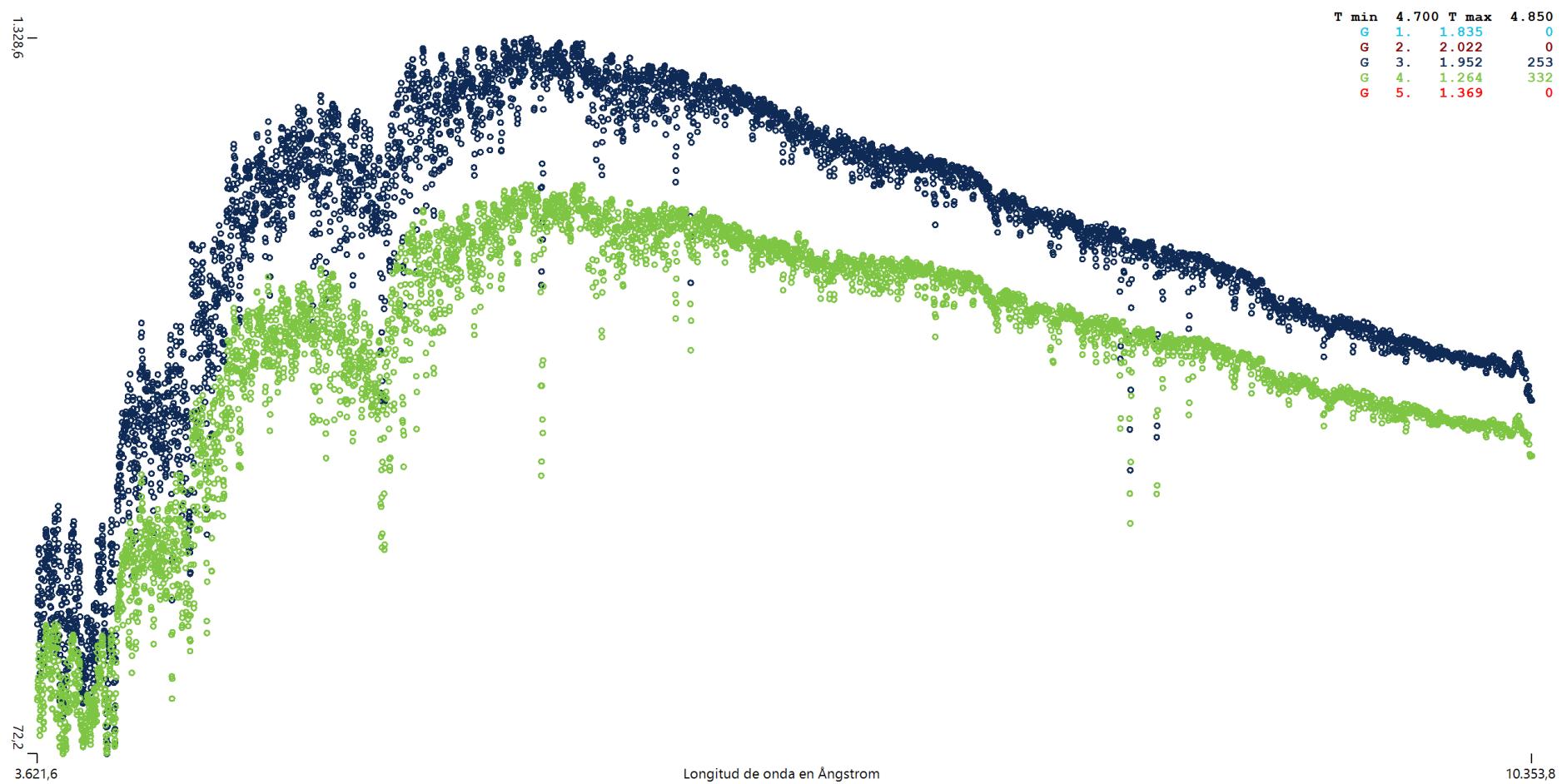


Figura 36. Clasificación K-means con datos originales y distancia igual a 1-covarianza. Espectros en el rango de temperaturas 5.400-5.600 K asignados a dos grupos distintos.

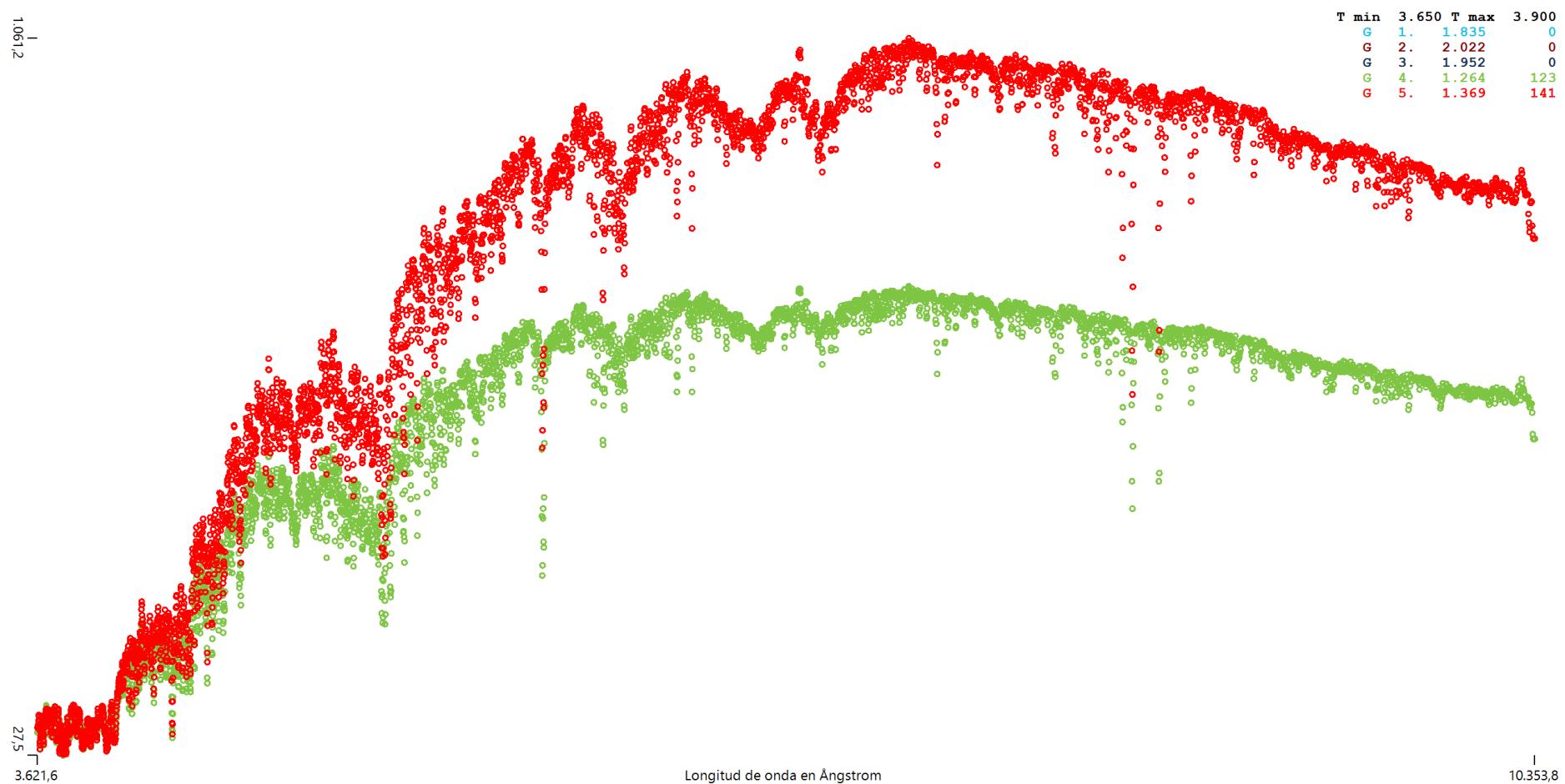


Figura 37. Clasificación K-means con datos originales y distancia igual a 1-covarianza. Espectros en el rango de temperaturas 3.650-3.900 K asignados a dos grupos distintos.

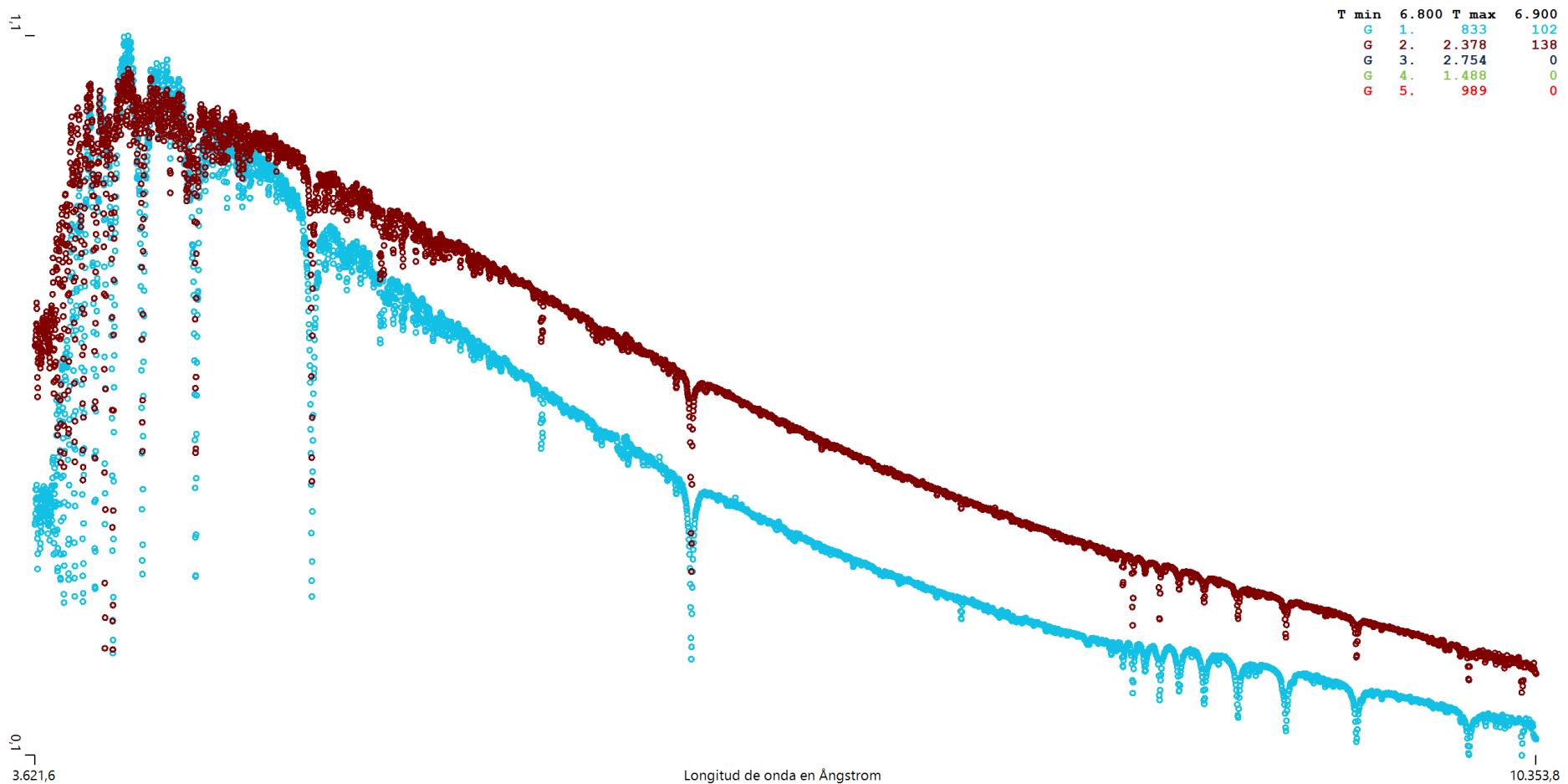


Figura 38. Clasificación K-means con datos normalizados y distancia euclíadiana. Espectros en el rango de temperaturas 6.800-6.900 K asignados a dos grupos distintos.

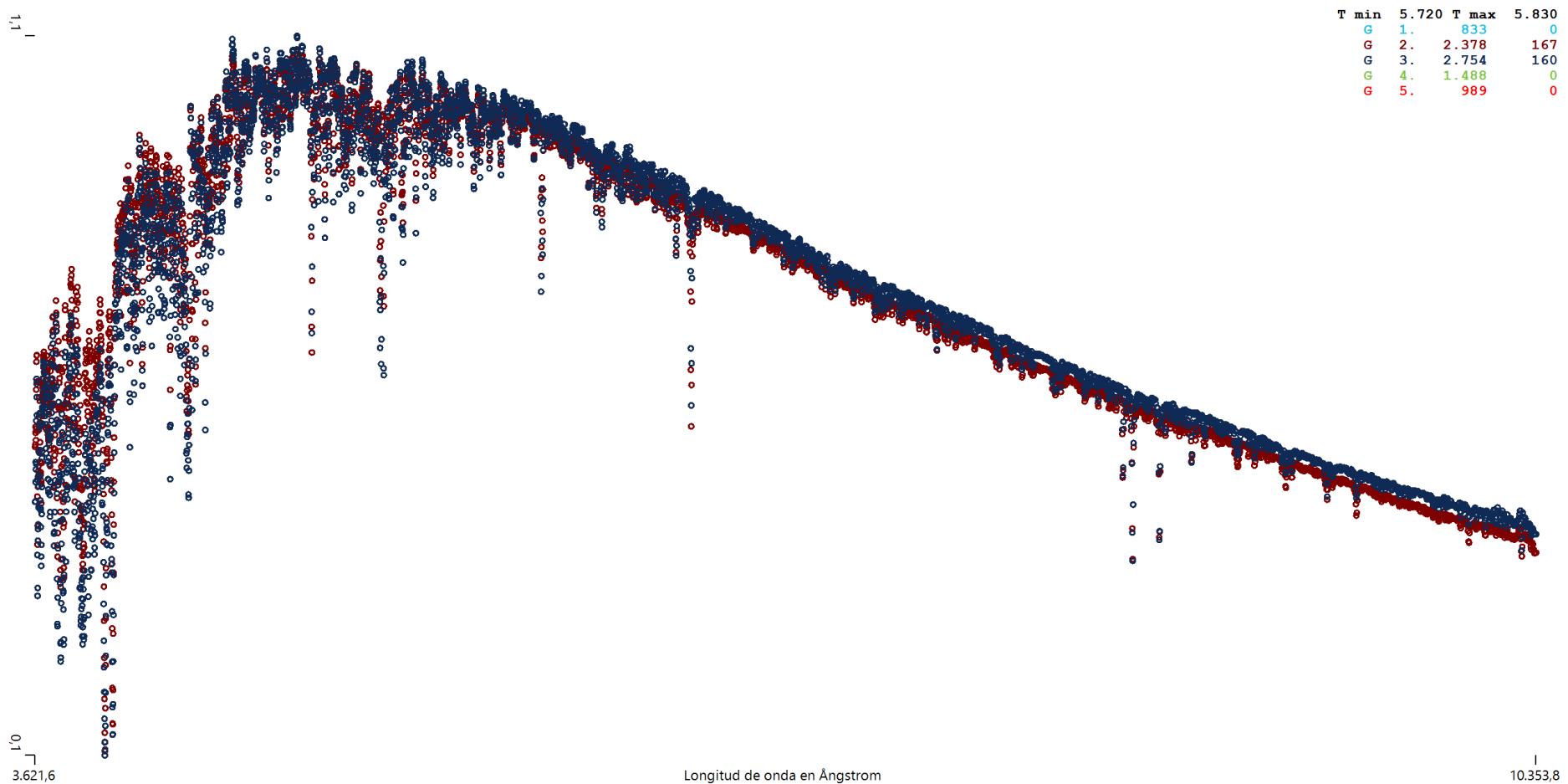


Figura 39. Clasificación K-means con datos normalizados y distancia euclidiana. Espectros en el rango de temperaturas 5.720-5.830 K asignados a dos grupos distintos.

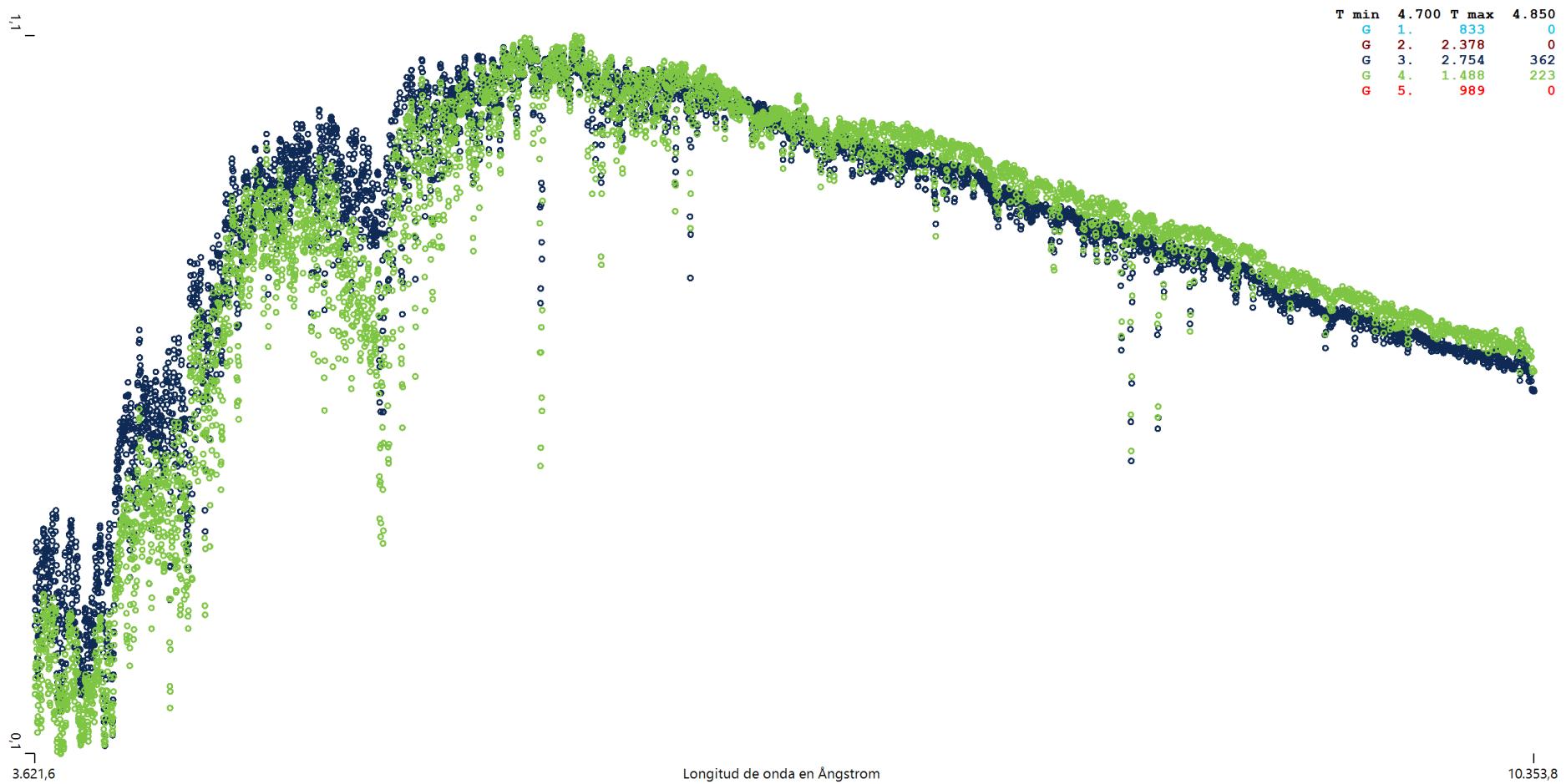


Figura 40. Clasificación K-means con datos normalizados y distancia euclidiana. Espectros en el rango de temperaturas 4.700-4.850 K asignados a dos grupos distintos.

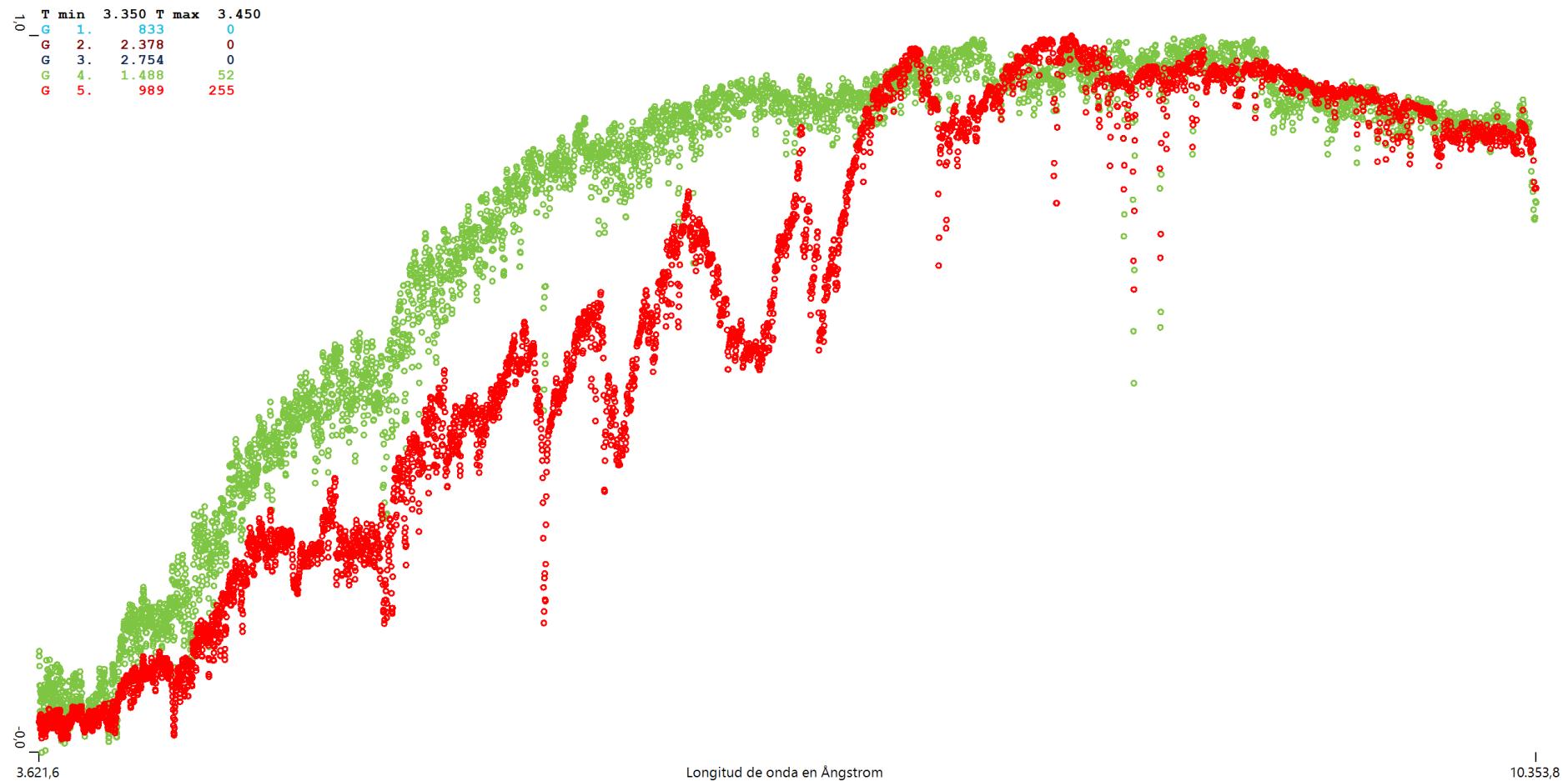


Figura 41. Clasificación K-means con datos normalizados y distancia euclidiana. Espectros en el rango de temperaturas 3.350-3.450 K asignados a dos grupos distintos.

Para completar la panorámica de resultados, la aplicación del algoritmo genético utilizando como cromosomas los 1.000 resultados de las simulaciones Monte Carlo conduce a resultados muy similares en los que apenas se mejora la distancia media y la desviación porcentual por dato (DPD) cuando el criterio de adaptación al medio es la propia distancia media y aunque las mejoras son algo más apreciables en la dispersión de la temperatura por grupos cuando se utiliza esta como criterio de adaptación, la clasificación empeora la distancia media entre los espectros y sus centroides aunque los resultados no cambian de forma apreciable respecto a lo aquí expuesto para la clasificación K-means sin algoritmo genético. Además la mejora en la dispersión por temperatura sólo se produce en el caso de agrupar con datos normalizados y distancia euclíadiana (desviación estándar de la temperatura = 321 en lugar de 347 y distancia media de 5,25 en lugar de 5,07), con datos originales y distancia igual a 1-covarianza no se produce ninguna mejora.

Para terminar, se muestran los resultados del escenario 3 de clasificación: Utilizar la diferencia entre el dato del espectro y el valor del polinomio ajustado. (tabla 9 y figura 42).

Eliminar la radiación térmica del espectro, hace que estos sean menos discriminatorios y la clasificación los diferencia menos, concentrando 3.164 espectros en un solo grupo. Además, la dispersión en temperatura dentro de cada grupo es muy grande, como cabía esperar al eliminar el efecto directo de la temperatura en la clasificación.

grupo	número espectros	grupo						temperatura K			Desviación por dato	
		distancia			desviación		más próximo	índice DB	L.o.Max Å	media	d.estándar T	%media
		d_min	d_max	media	estándar							
1	963	1,28	20,34	4,14	2,42		2	2,089	4.176	6.955	274	2,712
2	3.164	0,88	23,16	3,19	2,39		1	2,089	5.013	5.891	723	2,492
3	2.294	1,02	18,23	3,81	2,37		2	1,616	5.858	5.025	580	1,330
4	1.398	2,76	20,63	5,81	2,16		3	1,733	7.171	4.123	581	1,288
5	623	3,80	27,12	11,90	4,04		4	1,356	8.696	3.347	233	2,541
	8.442			4,54	2,51			1,776		5.297	597	1,799

Tabla 9. Clasificación K-mean de las diferencias entre los datos originales y el polinomio ajustado mediante distancia euclíadiana.

G	1.	963
G	2.	3.164
G	3.	2.294
G	4.	1.398
G	5.	623

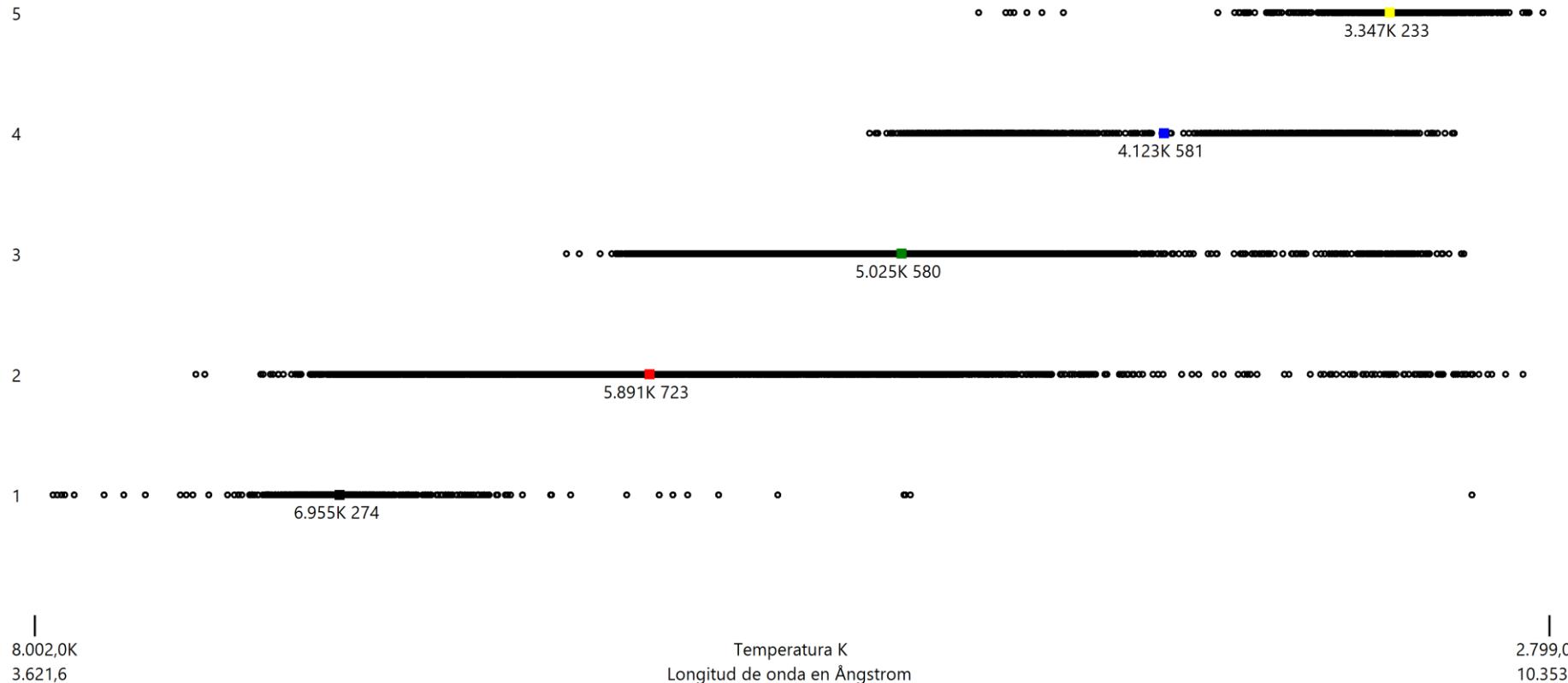


Figura 42. Clasificación K-mean de las diferencias entre los datos originales y el polinomio ajustado mediante distancia euclíadiana.

6. Conclusiones.

El algoritmo K-means es capaz de clasificar cuantitativamente a las estrellas por su espectro. La clasificación resultante presenta métricas mejores que la basada en la temperatura y pone de manifiesto que hay estrellas cuyo espectro es más parecido a los de otras con una temperatura diferente, que a otros de temperatura similar. No obstante, un primer análisis de las diferencias entre espectros en un mismo rango de temperaturas pero asignados a grupos distintos no permite identificar esas diferencias con características físicas reconocibles lo que pone en duda que la clasificación K-means sea preferible a la tradicional.

Lo que sí parece posible, siempre que la muestra de estrellas utilizada en este trabajo sea representativa de la generalidad, es redefinir los límites de temperatura en la clasificación Harvard teniendo en cuenta los resultados de la clasificación K-means con datos originales y distancia igual a 1-covarianza. A partir de la tabla 7, se pueden redefinir dichos rangos, obteniéndose los valores de la tabla 10:

Harvard	K-means
7.500–10.000 K	6.400-10.000 K
6.000–7.500 K	5.500-6.400 K
5.200–6.000 K	4.800-5.500 K
3.700–5.200 K	4.000-4.800 K
≤ 3.700 K	≤ 4.000 K

Tabla 10. Rango de temperaturas propuesto para la clasificación de las estrellas.

La tabla 11 muestra los resultados de la clasificación por temperatura con los nuevos rangos y la figura 43 compara la temperatura media y dispersión de esta, dentro de cada grupo, para las clasificaciones Harvard y Harvard con los nuevos rangos.

grupo	número	grupo					temperatura K			Desviación por dato	
	espectros	distancia		desviación	índice más próximo	índice DB	L.o.Max Å	media	d.estándar T	%media	
		d_min	d_max	media	estándar						
1	1.689	0,00	0,70	0,03	0,03	2	0,269	4.227	6.863	223	0,000
2	2.259	0,00	0,24	0,04	0,04	3	0,296	4.913	5.908	232	0,000
3	1.840	0,00	0,47	0,04	0,05	4	0,438	5.713	5.080	199	0,000
4	1.105	0,00	0,48	0,03	0,04	3	0,438	6.333	4.583	176	0,000
5	1.549	0,00	0,48	0,04	0,04	4	0,339	8.402	3.463	222	0,000
	8.442		0,04	0,04		0,356		5.297		215	1,294E-06

Tabla 11. Clasificación por temperatura con los nuevos rangos.

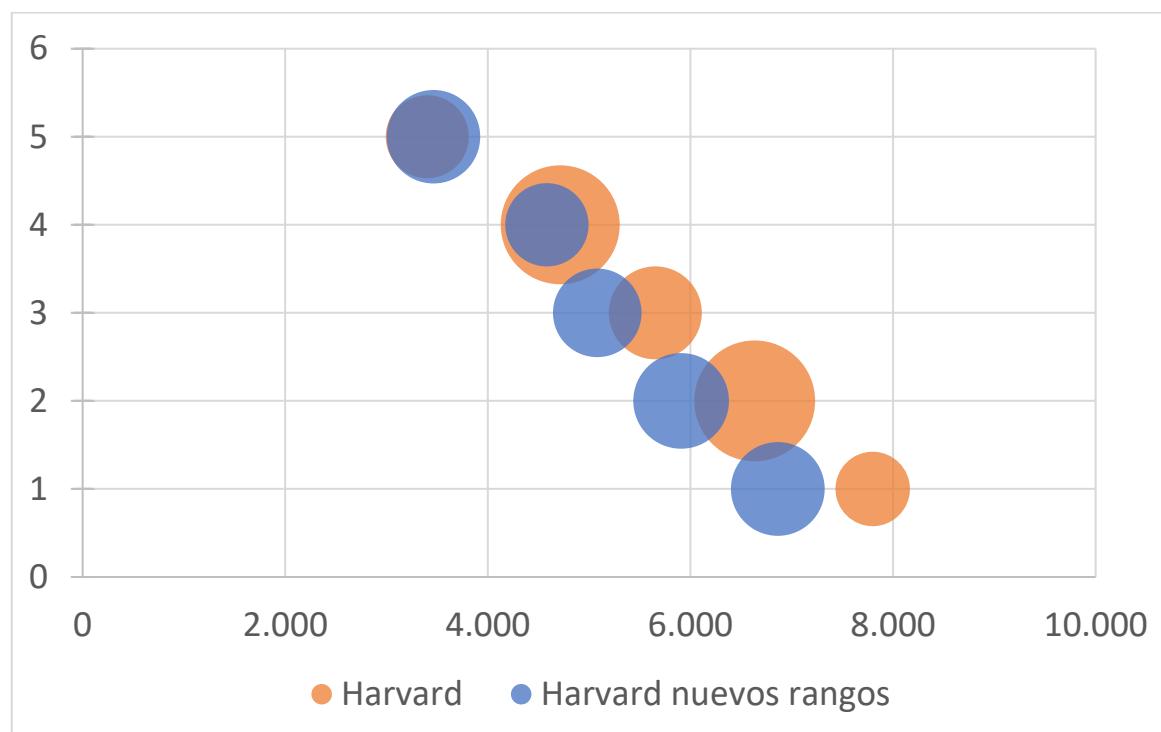


Figura 43. Comparación de la temperatura media y dispersión de esta, dentro de cada grupo, para las clasificaciones Harvard y Harvard con los nuevos rangos.

ANEXO I. Espectrómetro casero.

En la página <https://spectralworkbench.org/> se encuentran las instrucciones bajo el lema “hágaselo usted mismo” para fabricar un espectroscopio con materiales que prácticamente todo el mundo tiene en su casa: un DVD, cartulina y pegamento.

Esta Web, incluso permite calibrar online nuestro espectro y “subirlo” para mostrárselo a la comunidad.

El proceso de construcción es muy sencillo: se imprime y recorta la plantilla que nos facilitan, se le pega un trozo de DVD y listo.

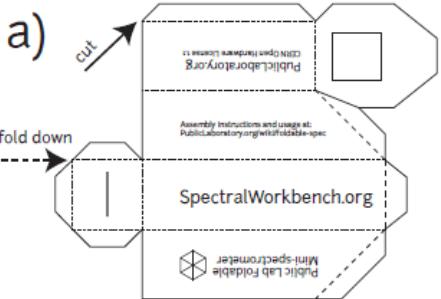
La clave es el DVD. Un DVD tiene un gran número de pistas donde se graba la información, en concreto 1.400 pistas por milímetro. Estas pistas constituyen una red de difracción, basta mirar la cara brillante de un DVD (también de un CD) para observar como la luz se refleja formando un arcoíris.

La clave del proceso de fabricación de nuestro espectrómetro es separar las dos capas de plástico que forman un DVD, cosa que se hace muy fácilmente pasando un cuchillo entre ellas y luego quitando el material reflectante en la capa que tienes las pistas, cosa también extraordinariamente sencilla por medio de cinta aislante, celo o cualquier adhesivo que al pegarse a ella la saca limpiamente.

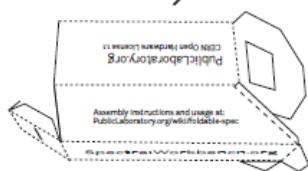
A continuación, la plantilla que explica el proceso y contiene el recortable para el cuerpo del espectrómetro.

1. cut and fold

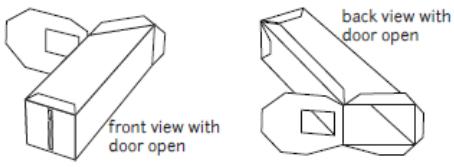
Cut along the outer edge. Fold up or down as indicated by the dotted and dashed lines. All labels should stay on the outside.



b)



Except for the diffraction grating door, glue or tape all flaps down onto the outside.

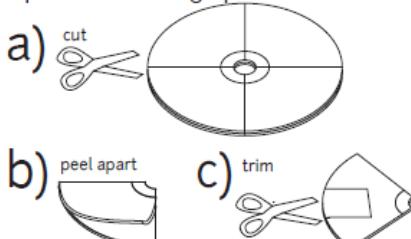


2.

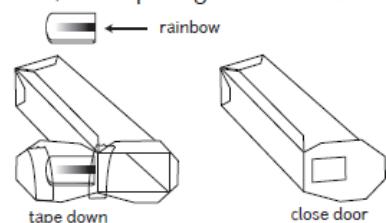
make a diffraction grating from a DVD-R
A diffraction grating is a series of close slits that disperse light.



To make one from a DVD-R, split it into quarters, peel off the reflective layer and trim a small clean square out of the transparent layer. Try to pick a clean piece without fingerprints or scratches.



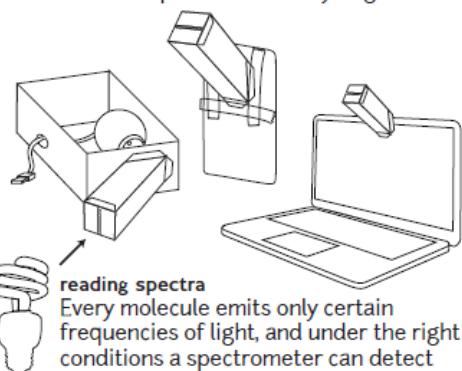
To work as a diffraction grating the DVD-R must be placed so that its grating is vertical, making a horizontal spectral rainbow. Tape your DVD piece to the inside of the spectrometer's door, then tape or glue the door closed.



3.

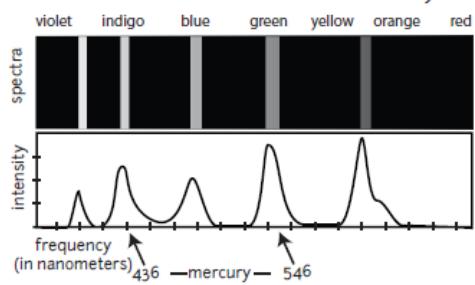
attach to a webcam, phone, or laptop

The spectrometer can be mounted on a camera phone, laptop, or with the help of a box, attached to a webcam. Line up carefully so that the rainbow is in the middle of the image, and tape down firmly so that the spectrometer stays rigid.



reading spectra

Every molecule emits only certain frequencies of light, and under the right conditions a spectrometer can detect these as rainbow bands. With two clear bands, the mercury in compact fluorescents makes calibration easy.



fold down

cut

fold down

flip

This open hardware design was developed by Public Lab contributors; You are free to reproduce, share, & distribute with attribution.

Join up, calibrate, & share spectra

Go online to Spectralworkbench.org, follow the calibration instructions, and you'll be ready to upload calibrated spectra!

Don't forget to share and publish your research as Research Notes on Publiclaboratory.org, and ask questions through the Public Laboratory Spectrometry mailing list.

SpectralWorkbench.org

Mini-Spectrometer

Una red de difracción no es más que un conjunto de líneas grabadas en algún material de forma paralela y muy juntas. Su funcionamiento se basa en el principio que Christiaan Huygens establecio en 1690: todo punto alcanzado por una onda se convierte en un emisor de ondas en todas las direcciones.

Cuando un frente de ondas (varias ondas viajando paralelas) incide sobre la red de forma oblicua, cada línea se convierte en un repetidor de la onda que le llega, de forma que las emisiones de cada una de las líneas interfiere con la de las demás, en algunos puntos las ondas se refuerzan porque se encuentran en fase y en otras se anulan por estar en fase contraria. Para cada onda del frente, dependiendo de su longitud de onda, los puntos de refuerzo y oscuridad son distintos y de esta forma aparecen separadas en el espectro.

En la imagen I.1 muestro mi propio espectrógrafo y en la I.II el espectro de una bombilla de incandescencia que de milagro quedaba en casa.

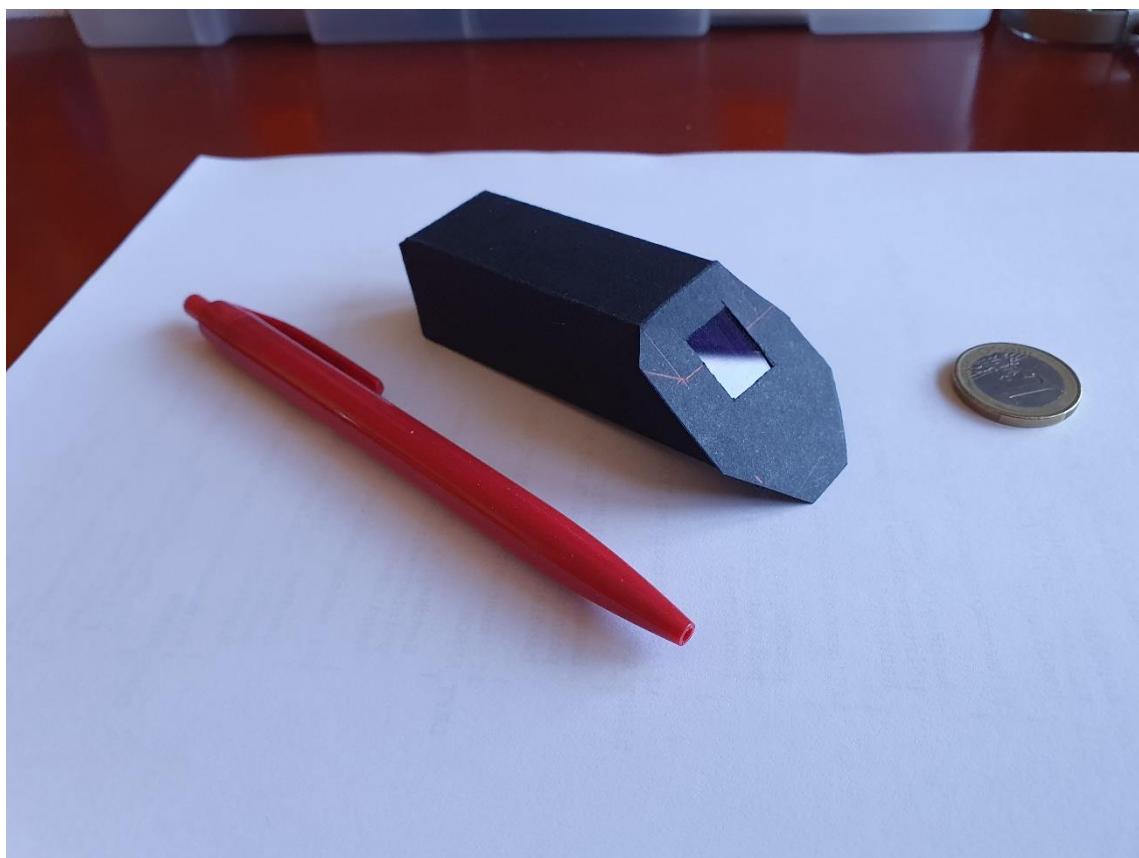


Figura I.1. Espectrógrafo de cartulina con un trozo de DVD como red de difracción.

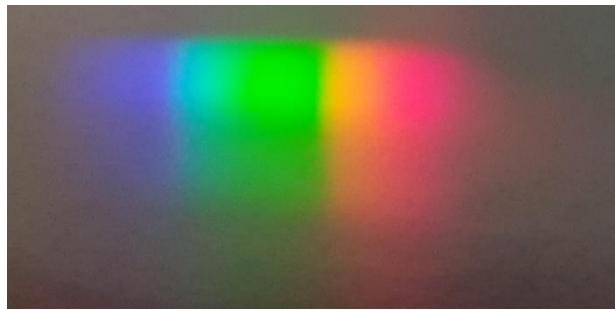


Figura I.2. Espectro de una bombilla incandescente, fotografiado con el teléfono.

ANEXO II. Aplicación Kespectro.

Para realizar los experimentos de clasificación de este trabajo he escrito una aplicación informática para el sistema operativo Windows.

La aplicación lee los ficheros con los datos de los espectros que exporta la aplicación ExploraFits a partir de los ficheros en formato FITS descargados de la Web del proyecto SDSS.

A partir de estos ficheros crea nuevos ficheros que incluyen junto a los datos originales los ajustes polinómicos y las “diferencias” polinomio-espectro que se explican en apartado 2.2 de este documento.

Estos últimos ficheros son lo que emplea la aplicación para realizar la clasificación de los espectros con las opciones que se especifiquen para cada escenario a estudiar.

A continuación se describe la ventana principal de la aplicación (figura II.1) que consta de dos zonas.

La zona izquierda trabaja con los datos originales de los espectros, fundamentalmente para el ajuste de polinomios (figura II.2).

La zona de la derecha es la que se encarga de las clasificación de los espectros (figuras II.3a y II.3b).

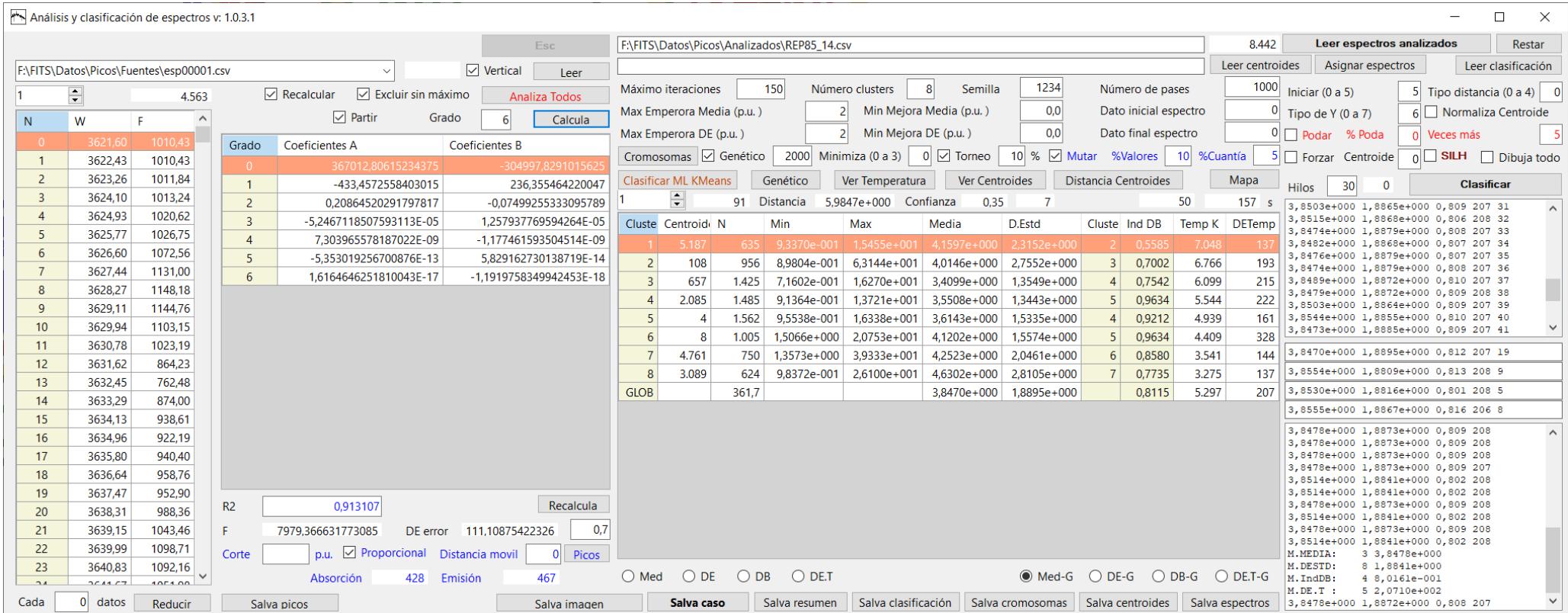


Figura II.1 Ventana principal de la aplicación Kespectros.

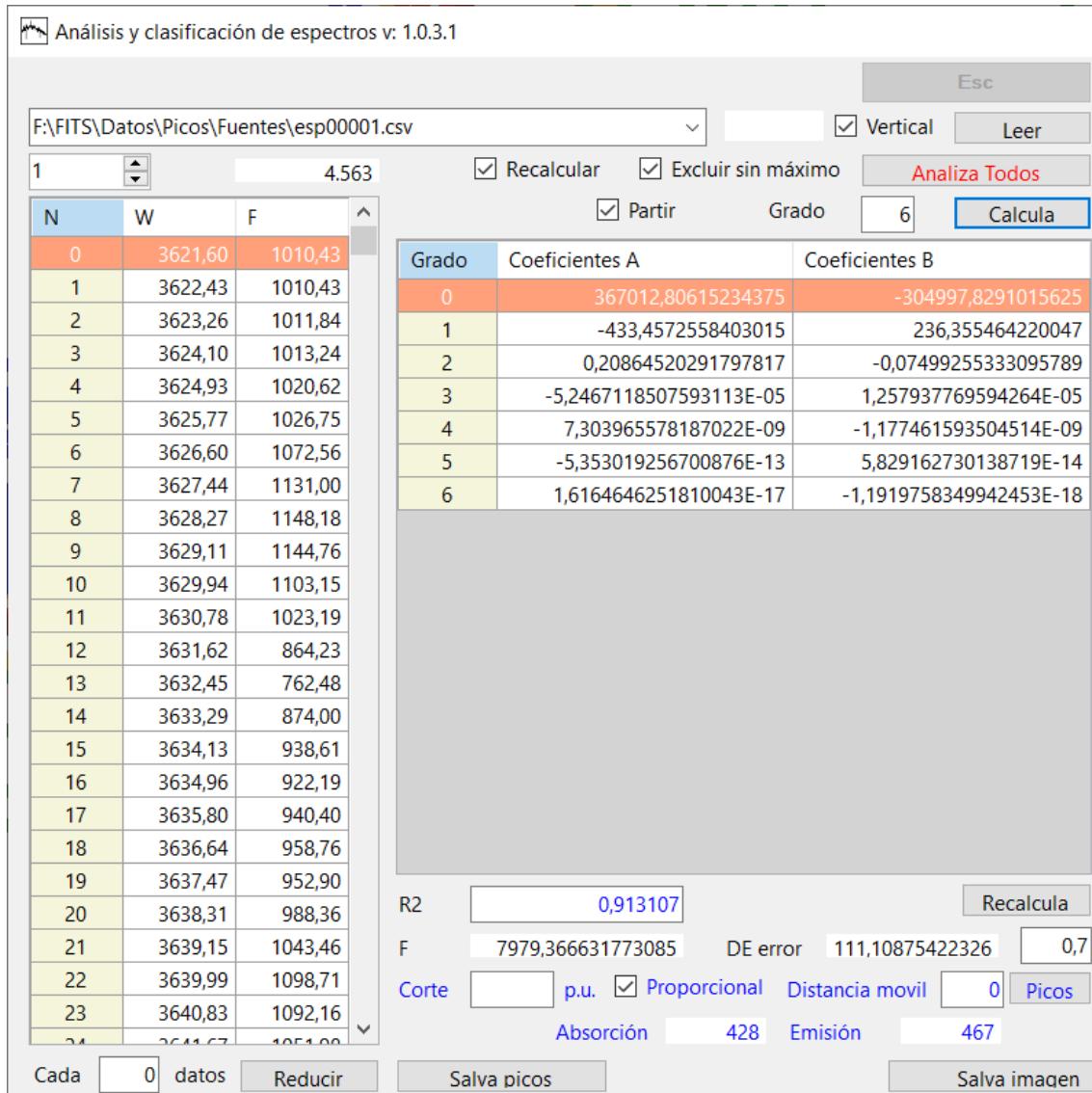


Figura II.2 Zona para el ajuste polinómico de los espectros.

Leer.

Todos los ficheros que maneja la aplicación están en formato CSV, texto separado por caracteres ";". Se puede leer un espectro con los datos en dos columnas: wave y flux, para lo cual hay que marcar el selector "vertical", o se puede leer un fichero con múltiples espectros, un espectro por fila (formato horizontal) con al menos tantas columnas como datos por dos, las columnas con la longitud de onda llevarán las cabeceras WAVE_1, WAVE_2, etc y las radiaciones FLUX_1, FLUX_2, etc. Para leer estos ficheros el selector "vertical" debe estar sin marcar.

El botón "Calcula" realiza el ajuste polinómico de los datos activos en ese momento y el botón "Recalcula" realiza una nueva regresión, eliminando los "picos", que son aquellos datos que están a una distancia superior a un determinado valor de "Corte" del valor ajustado por el polinomio. El valor de Corte se puede establecer manualmente en la

casilla correspondiente (el selector “Proporcional” debe estar sin marcar) o lo puede calcular la aplicación a partir del DE error del ajuste (valor 0,7 en la imagen), este valor de corte variará con la longitud de onda si se marca el selector “Proporcional”, en este caso el valor introducido en la casilla Corte se considera un valor por unidad (p.u.).

Analiza todos.

La tarea principal de esta sección de la aplicación es preparar el fichero con los datos para el algoritmo K-means (fichero de espectros analizados) lo cual se hace mediante el botón “Analiza todos”. Previamente hay que “Leer” un fichero con formato horizontal que contenga todos los espectros. Al pulsar el botón “Analiza todos” la aplicación solicita un nombre de fichero donde escribir los resultados, se realiza el ajuste polinómico de cada espectro y se escribe en el fichero de salida que se haya especificado.

Si el selector “Recalcular” está marcado el ajuste polinómico final se hace excluyendo los picos del espectro.

Si el selector “Excluir sin máximo” está marcado, se eliminarán los espectros cuyo polinomio ajustado no tenga el máximo dentro del rango de longitudes de onda del espectro.

Si el selector “Partir” está marcado, el ajuste polinómico se realiza mediante dos regresiones como se explica más adelante en el apartado “Proceso de ajuste del polinomio”

Grado del polinomio.

A mayor grado del polinomio debe producirse un mejor ajuste, pero hay un problema de precisión numérica. Para realizar los cálculos el ordenador necesita representar los valores numéricos con un formato determinado, cuanto mayor es número de bytes que el formato dedica a cada número mayor es la precisión con la que se representan los datos. Kespectro utiliza el formato de doble precisión (8 bytes por dato, la máxima del compilador de Windows sin tener que recurrir a librerías especializadas). Este formato tiene un máximo de 16 cifras significativas y si tenemos en cuenta que los valores de la longitud de onda de nuestros espectros son del orden de 10^{**3} , el término x^{**10} del polinomio es un número del orden de 10^{**30} , mientras que el término de primer orden x^{**1} , es de 10^{**3} .

Sumar, o restar un número del orden de 10^{**3} a otro del orden de 10^{**30} que sólo tiene 16 cifras significativas no cambia el valor de este último ya que le cambiaría, como mucho la cifra $30 - 3 = 27$.

Esta es la razón por la que no podemos utilizar polinomios de un grado tan alto como queramos, para paliar este problema, la aplicación realiza dos regresiones, la primera, A, con los primeros 2/3 de los datos (desde el dato 0 hasta el dato 2/3 del total) y la segunda, B, con los 2/3 últimos datos (desde el dato 1/3 hasta el último dato). Ambas regresiones se fusionan utilizando la media ponderada para el tercio central de datos

que comparten los dos ajustes: se pondera con peso variable de 0 (primer dato compartido por B) a 1 (último dato compartido por B), de forma lineal.

Proceso de ajuste del polinomio.

Se realiza una primera regresión (en realidad dos regresiones como se ha dicho) de los datos empezando con un polinomio de grado 5 y se va aumentando el grado mientras que R² (coeficiente de correlación al cuadrado) siga aumentando (el menor de las dos regresiones parciales), en el momento que retrocede significa que hemos alcanzado el grado máximo que la precisión numérica de la máquina nos permite, por lo que retrocedemos al grado anterior, eliminamos de los datos (sólo a efectos de la regresión) los valores fuera de la franja del ruido (figura 12) y efectuamos un nuevo ajuste que mejora de forma importante el R², en cualquier caso el R² generalmente supera el 0,95. La eliminación de los picos es para que el ajuste represente mejor la radiación térmica.

F:\FITS\Datos\Picos\Analizados\REP85_14.csv										8.442		
										<input type="button" value="Leer centroide"/>		
Máximo iteraciones	150	Número clusters	8	Semilla	1234	Número de pasos	1000					
Max Emperador Media (p.u.)	2	Min Mejora Media (p.u.)	0,0	Dato inicial espectro	0							
Max Emperador DE (p.u.)	2	Min Mejora DE (p.u.)	0,0	Dato final espectro	0							
Cromosomas	<input checked="" type="checkbox"/> Genético	2000	Minimiza (0 a 3)	0	<input checked="" type="checkbox"/> Torneo	10 %	<input checked="" type="checkbox"/> Mutar	%Valores	10	%Cuantía	5	
<input type="button" value="Clasificar ML KMeans"/>		<input type="button" value="Genético"/>	<input type="button" value="Ver Temperatura"/>	<input type="button" value="Ver Centroides"/>	<input type="button" value="Distancia Centroides"/>	<input type="button" value="Mapa"/>						
1	<input type="button" value="▼"/>	91	Distancia	5,9847e+000	Confianza	0,35	7		50		157 s	
Cluste	Centroide N	Min	Max	Media	D.Estd	Cluste	Ind DB	Temp K	DETtemp			
1	5.187	635	9,3370e-001	1,5455e+001	4,1597e+000	2,3152e+000	2	0,5585	7.048	137		
2	108	956	8,9804e-001	6,3144e+001	4,0146e+000	2,7552e+000	3	0,7002	6.766	193		
3	657	1.425	7,1602e-001	1,6270e+001	3,4099e+000	1,3549e+000	4	0,7542	6.099	215		
4	2.085	1.485	9,1364e-001	1,3721e+001	3,5508e+000	1,3443e+000	5	0,9634	5.544	222		
5	4	1.562	9,5538e-001	1,6338e+001	3,6143e+000	1,5335e+000	4	0,9212	4.939	161		
6	8	1.005	1,5066e+000	2,0753e+001	4,1202e+000	1,5574e+000	5	0,9634	4.409	328		
7	4.761	750	1,3573e+000	3,9333e+001	4,2523e+000	2,0461e+000	6	0,8580	3.541	144		
8	3.089	624	9,8372e-001	2,6100e+001	4,6302e+000	2,8105e+000	7	0,7735	3.275	137		
GLOB		361,7			3,8470e+000	1,8895e+000		0,8115	5.297	207		
<input type="radio"/> Med <input type="radio"/> DE <input type="radio"/> DB <input type="radio"/> DET <input checked="" type="radio"/> Med-G <input type="radio"/> DE-G <input type="radio"/> DB-G <input type="radio"/> DET-G												
<input type="button" value="Salva caso"/>		<input type="button" value="Salva resumen"/>	<input type="button" value="Salva clasificación"/>	<input type="button" value="Salva cromosomas"/>	<input type="button" value="Salva centroides"/>	<input type="button" value="Salva espectros"/>						

Figura II.3a. Zona para la clasificación de los espectros. Parámetros y resultados.

El máximo de iteraciones es el límite de iteraciones que se impone al algoritmo K-means para encontrar una solución, en nuestro caso el algoritmo converge normalmente en menos de 30 iteraciones, imponer un límite garantiza que no se entrará en un ciclo infinito, por ejemplo, porque se entre en una situación oscilante con uno o varios espectros cambiando entre dos grupos de una iteración a la siguiente.

El número de pasos son el número de clasificaciones que se realizarán en la parte Monte Carlo antes de iniciar el algoritmo genético y por tanto es el número de cromosomas que utilizará este.

Se puede acotar el vector de datos a utilizar en la clasificación especificando el dato inicial y final de los espectros, en definitiva se acota el rango de longitudes de onda a considerar.

La casilla a la derecha del selector “Genético”, que indica si se utilizará o no este algoritmo, es el número de “generaciones” a evolucionar. Mediante el valor “Minimiza” se indica cual es criterio de adaptación al medio, el que establecerá el grado de adaptación.

0. La distancia. A menor valor mejor adaptado.
1. La desviación estándar. A menor valor mejor adaptado.
2. El índice DB. A menor valor mejor adaptado.
3. El índice silhouette. A mayor valor mejor adaptado, para 8.000 espectros es inviable por el tiempo de cálculo que necesita.

Si se marca el selector Torneo, el cromosoma a eliminar, en cada generación, será el peor adaptado de un número de cromosomas elegidos al azar (según el porcentaje que se especifique), en caso contrario se elimina el progenitor peor adaptado.

El selector Mutar da la opción de que una de cada dos generaciones se introduzca una mutación en el % de valores que se indique con una variación máxima (Cuantía) del % especificado.

Cuando se realiza una clasificación se puede escribir un fichero con los resultados de cada uno de los intentos Monte Carlo, que son los cromosomas del algoritmo genético. Este fichero se puede leer posteriormente (botón “Cromosomas”) para volver a aplicar el algoritmo genético (con las mismas o distintas opciones) pulsando el botón “Genético”.

El botón “Ver temperatura” abre una ventana a pantalla completa y muestra los histogramas de temperaturas de la clasificación.

El botón “Ver centroides” abre una ventana a pantalla completa y muestra los centroides de la clasificación, si se hace clic con el ratón, teniendo pulsada la tecla Ctrl, se marca en el gráfico la línea espectral atómica más próxima a la longitud de onda marcada, esta ventana también permite marcar todas las líneas del átomo que se especifique.

El botón “Distancia centroides” genera el gráfico que se muestra en la figura 23.

El botón “Mapa” muestra la ubicación angular (Ascensión recta, Declinación) de los espectros, si es que esta información fue incluida en los ficheros que se le proporcionan a la aplicación. (figura 11)

Los resultados se pueden “Salvar” en ficheros, el botón “Salva Caso” es el más general ya que guarda todo tipo de resultados y gráficos en la carpeta que se especifique. Dentro de esa carpeta se creará una nueva carpeta con el nombre formado por las opciones empleadas en la clasificación y dentro de ellas todos los ficheros de datos y gráficos.

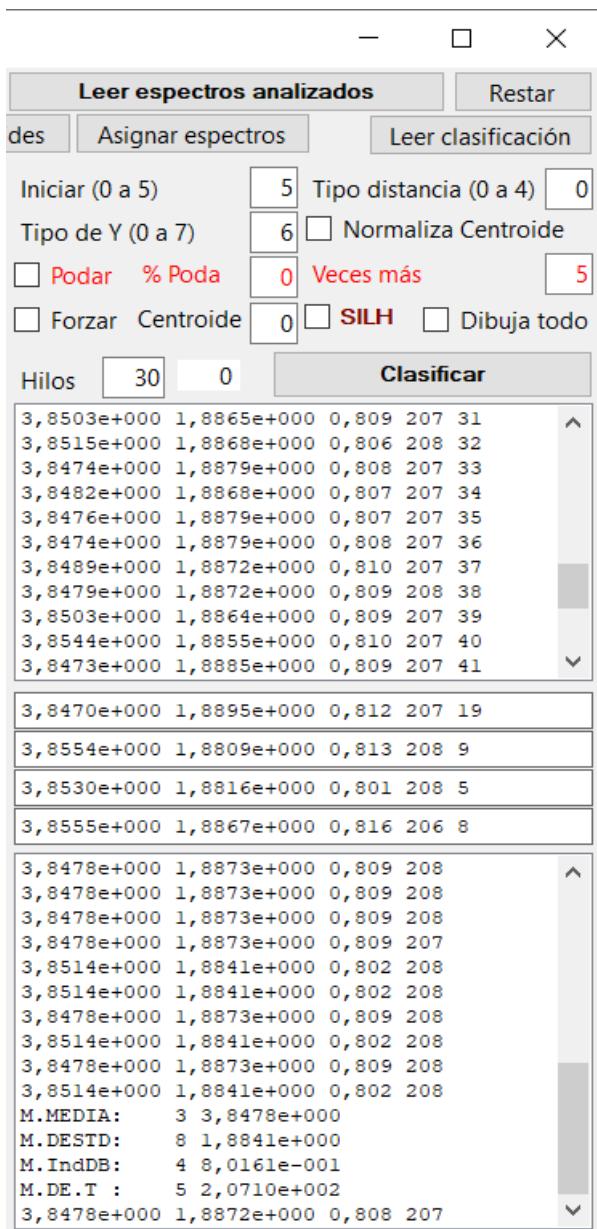


Figura II.3b. Zona para la clasificación de los espectros. Otros parámetros y evolución de los resultados.

La clasificación K-means se puede iniciar de las siguientes formas:

0. Centroides predefinidos. Leídos de un fichero, probablemente salvado en una clasificación anterior.
1. Espectros asignados al azar.
2. Centroides al azar, todos con el mismo número de espectros. El primer centroide es un espectro elegido al azar, se completa el grupo con los espectros más cercanos, luego se elige al azar un espectro aun no asignado como centroide del siguiente grupo y se repite el proceso.
3. Igual que el anterior, salvo que una vez completado un grupo el espectro que se elige como centroide del siguiente grupo, se elige al azar.
4. Se asignan al azar todos los espectros a la mitad de los grupos, luego, en cada grupo se elige al espectro más alejado del centroide y se asigna a un nuevo grupo como centroide.
5. Kmeans++

El tipo de Y es el valor (para una de las longitudes de onda) de los vectores que se van a clasificar:

0. Datos reales del espectro
1. Datos ajustados dentro del corte.
2. Diferencias relativas.
3. Diferencias absolutas.
4. Diferencias reducidas a 0,1 (si/no).

El tipo de distancia puede ser el siguiente:

0. Euclidiana
1. Manhattan
2. Covarianza.
3. Euclidiana^{**2}.
4. Manhattan^{**2}.

Normalizar el centroide sólo se aplica a la clasificación con tipo de Y número 4 y si se marca los centroides también se normalizan a 0,1.

La opción de podar permite que cuando se calcula el centroide, no se tengan en cuenta los espectros más discrepantes, el % de espectros del grupo que se indique y sólo si la distancia del espectro al resto de centroides es inferior al número de veces que se especifica, se trata de no descartar un espectro muy alejado de todos los demás grupos.

Si se marca la opción de Forzar Centroide se sustituirá el centroide de cada grupo por el espectro más próximo a él.

Centroide permite dos opciones:

0. El centroide se calcula como el valor medio de los espectros del grupo.
1. El centroide tiene para cada longitud de onda el valor máximo de los espectros del grupo.

El procedo Monte Carlo se puede realizar en paralelo utilizando todos los procesadores de la máquina, realizando simultáneamente una clasificación en cada uno de ellos. Mediante el Número de hilos, se indica cuantos usar.

Una vez realizada una clasificación se puede salvar a fichero mucha información, entre ellas los centroides y la clasificación (ordinal de los espectros asignados a cada grupo), esta información puede leerse posteriormente.

Si se lee:

- Fichero de espectros analizados
- Centroides

Se puede iniciar una nueva clasificación usando los centroides en lugar de que la aplicación genere unos centroides iniciales, pero también se puede hacer una asignación inmediata de los espectros a los centroides mediante el botón “Asignar espectros”. Esta es la forma de **aplicar el modelo a un fichero de espectros** distintos al utilizado en la clasificación que creo los centroides para asignarlos al grupo que les corresponda.

Si se lee:

- Fichero de espectros analizados.
- Clasificación.

Se reconstruye la clasificación realizada en su momento, el “Fichero de espectros analizados” debe ser el mismo que dio lugar a la “Clasificación”.

ANEXO III. Líneas espectrales atómicas.

Para la clasificación de los datos originales de los espectros utilizando la distancia igual a 1-covarianza, se han marcado las líneas espectrales del H (figuras III.1 a III.5), del He (figuras (III.6 a III.10) y otros elementos (III.11 a III. 15) presentes en la mayoría de los centroides y por tanto en los espectros agrupados en torno a ellos.

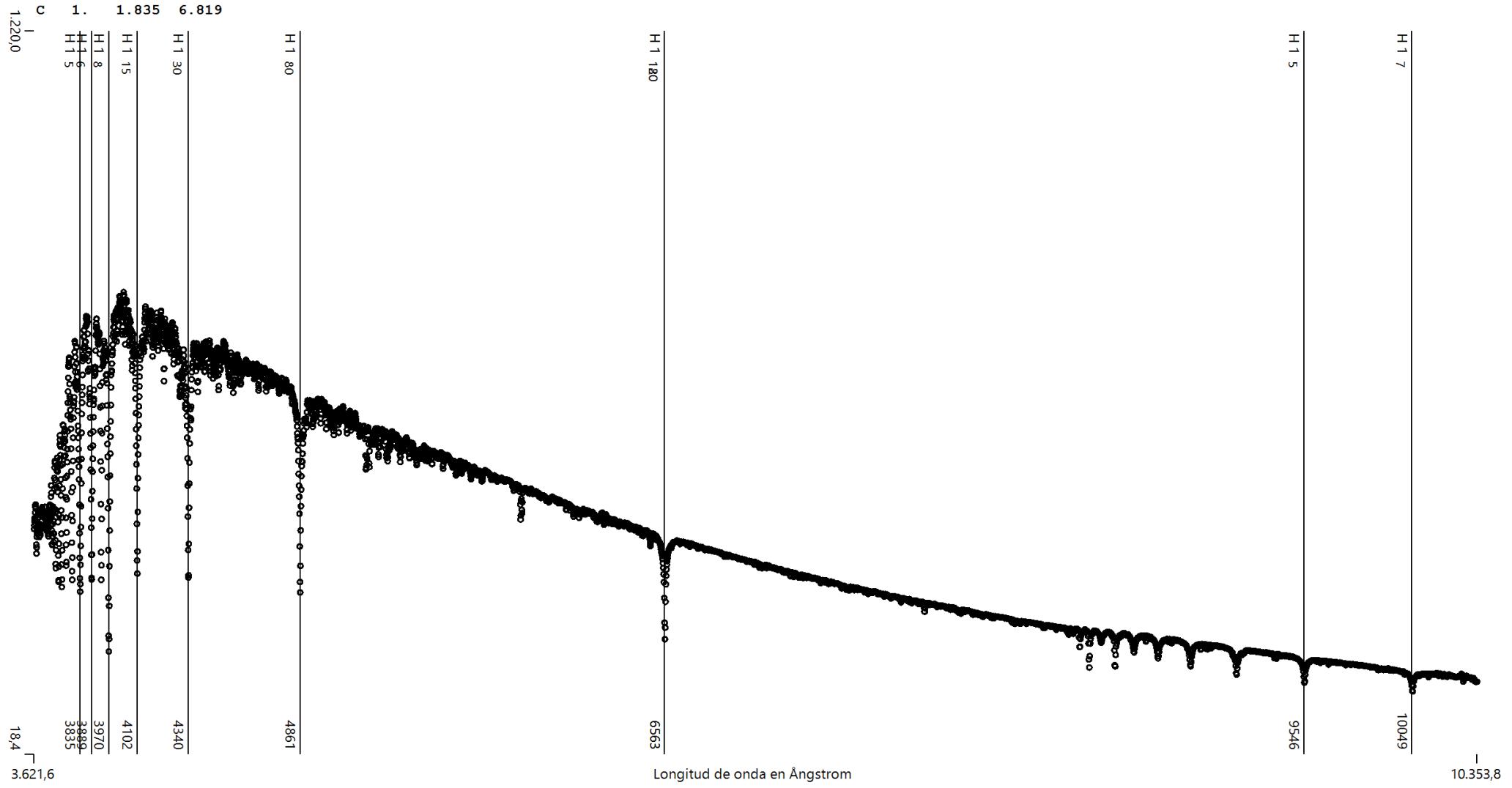


Figura III. 1. Líneas del hidrógeno sobre el centroide 1.

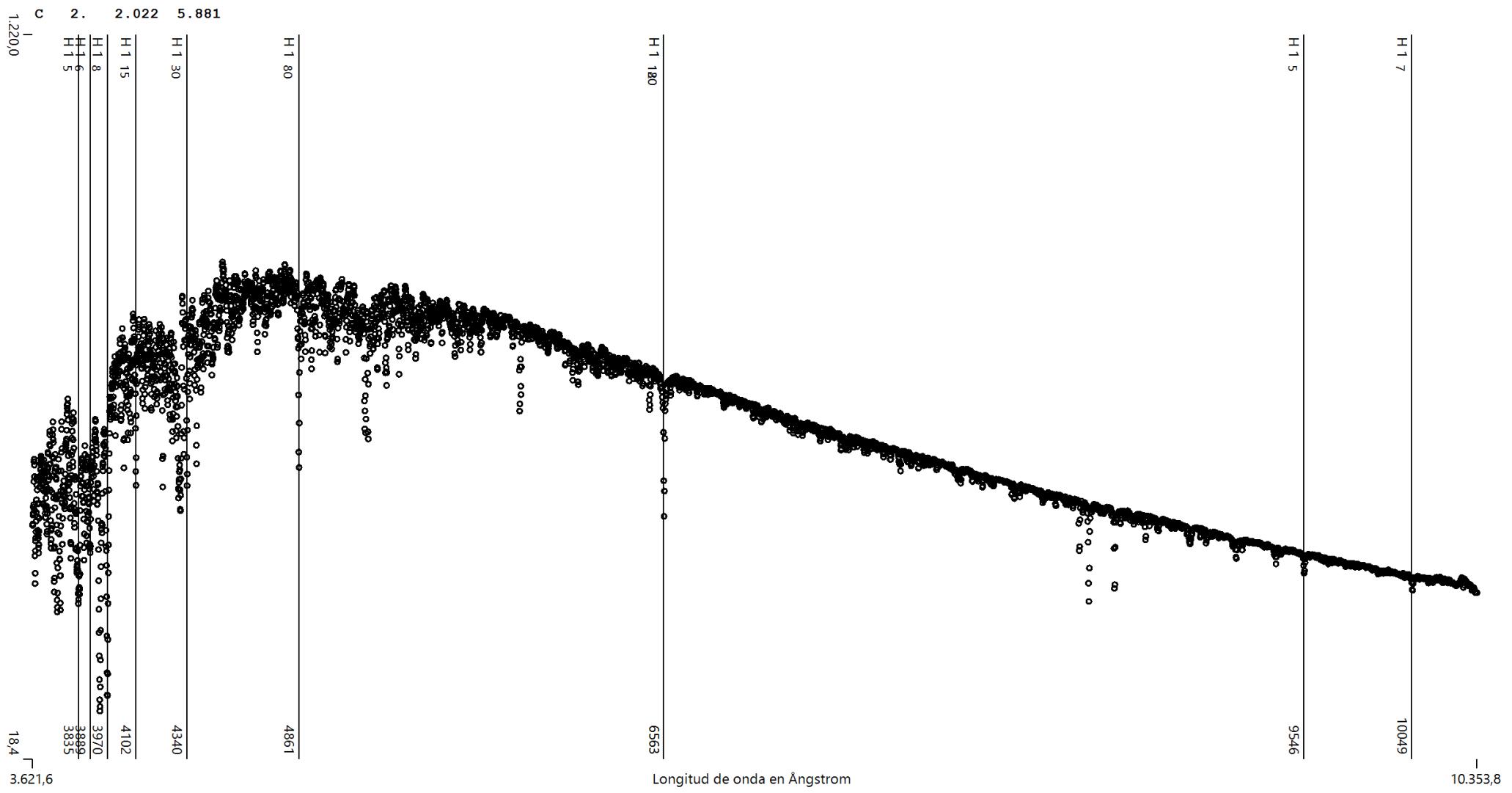


Figura III. 2. Líneas del hidrógeno sobre el centroide 2.

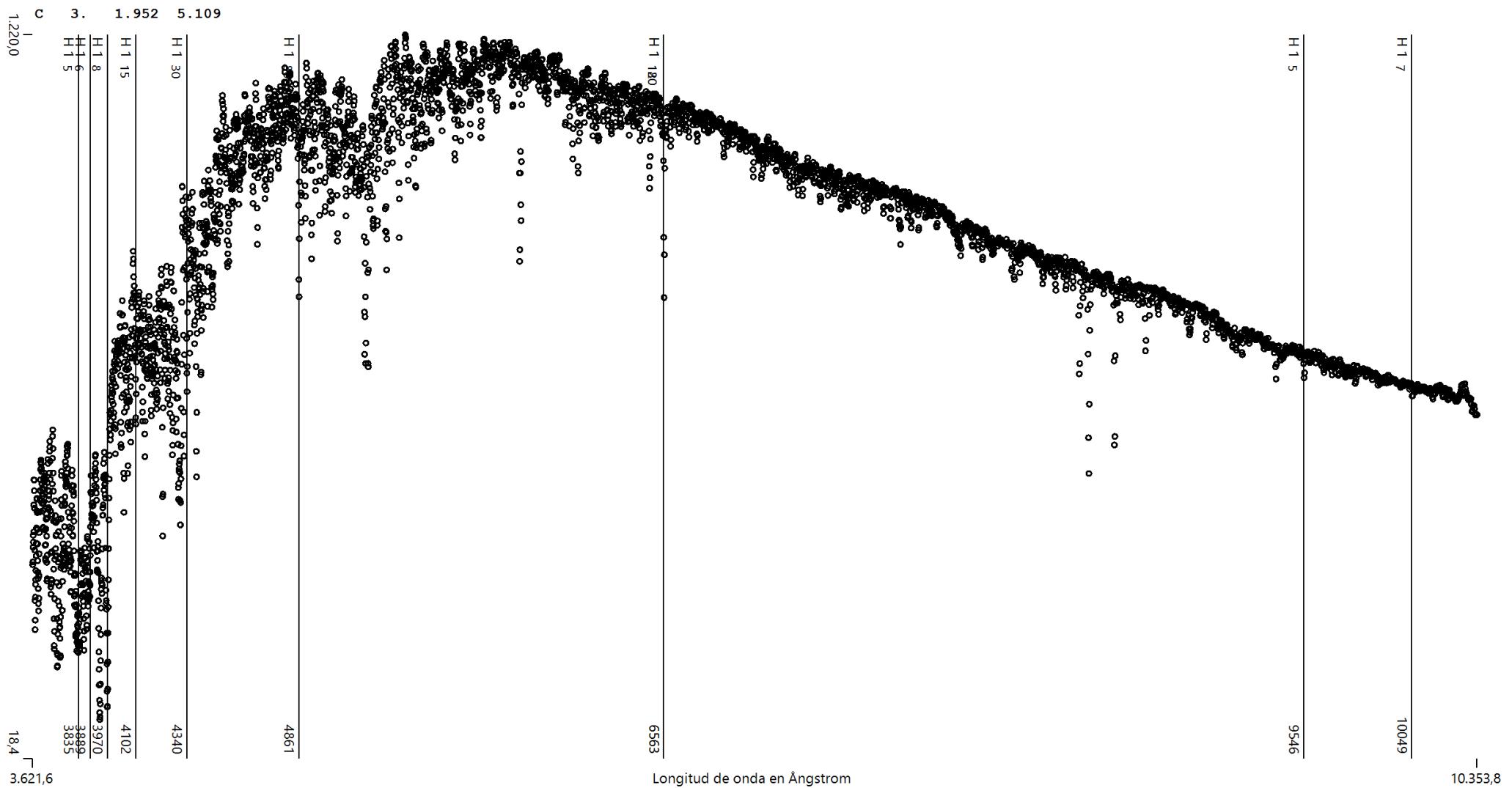


Figura III. 3. Líneas del hidrógeno sobre el centroide 3.

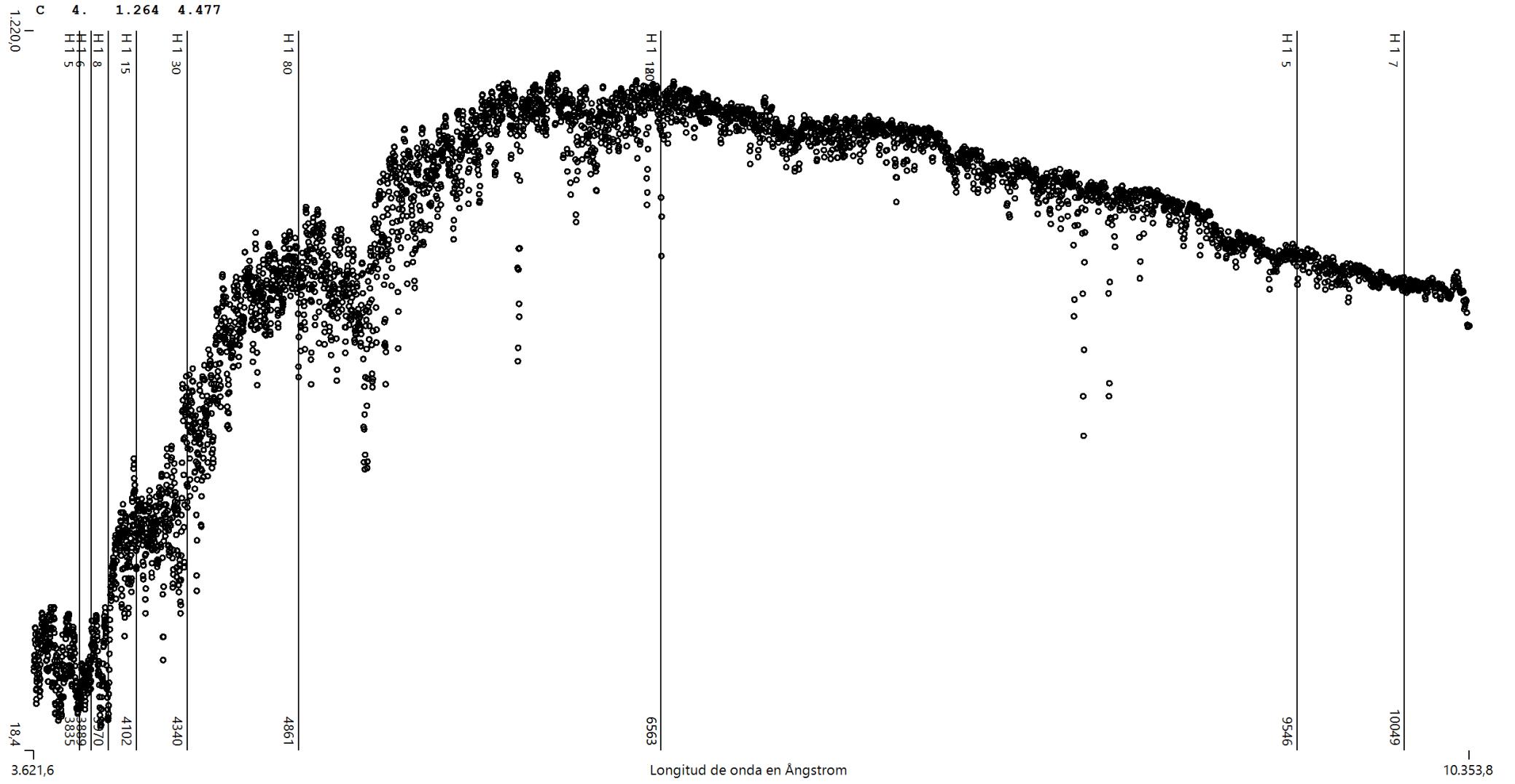


Figura III. 4. Líneas del hidrógeno sobre el centroide 4.

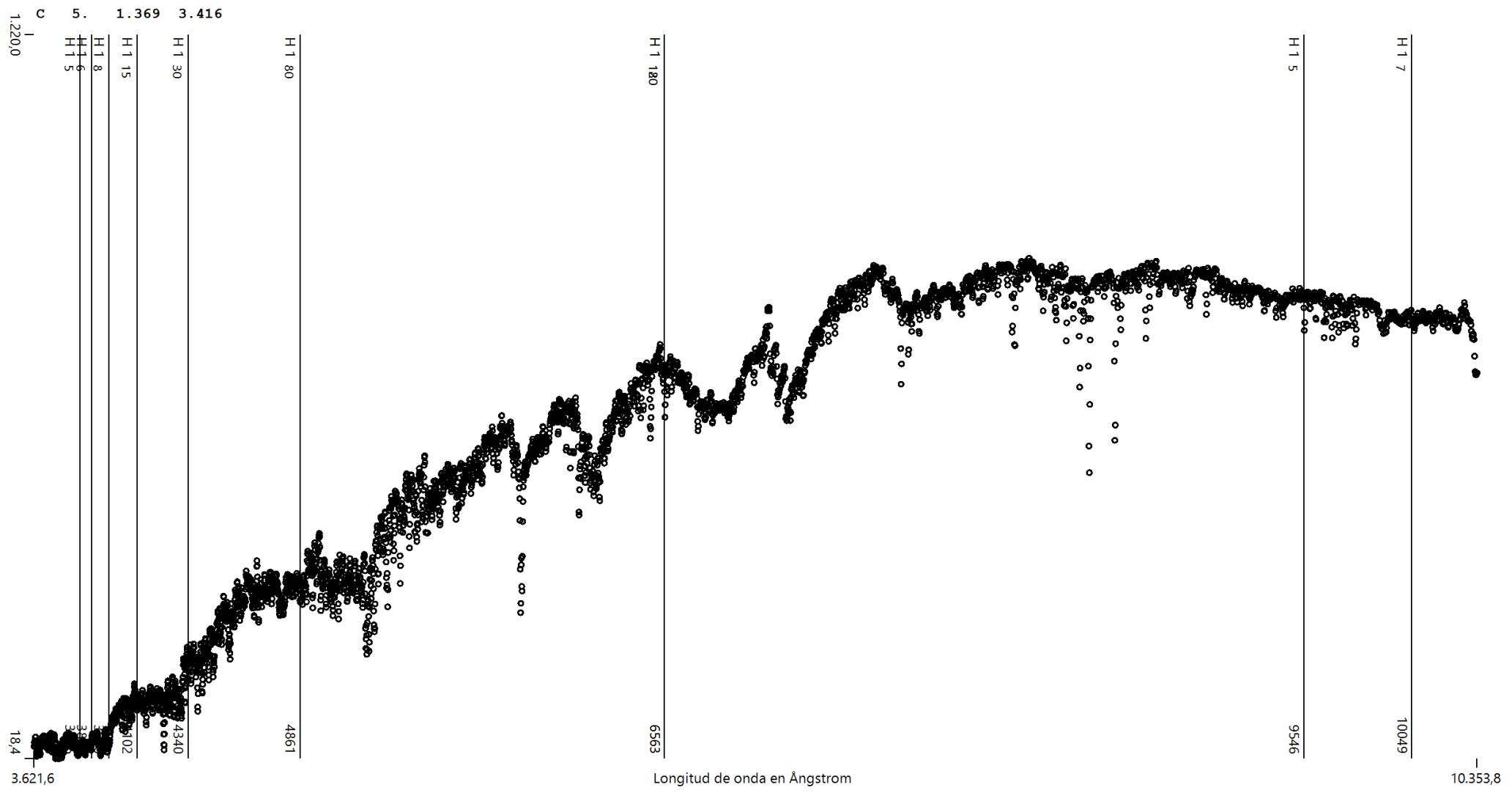


Figura III. 5. Líneas del hidrógeno sobre el centroide 5.

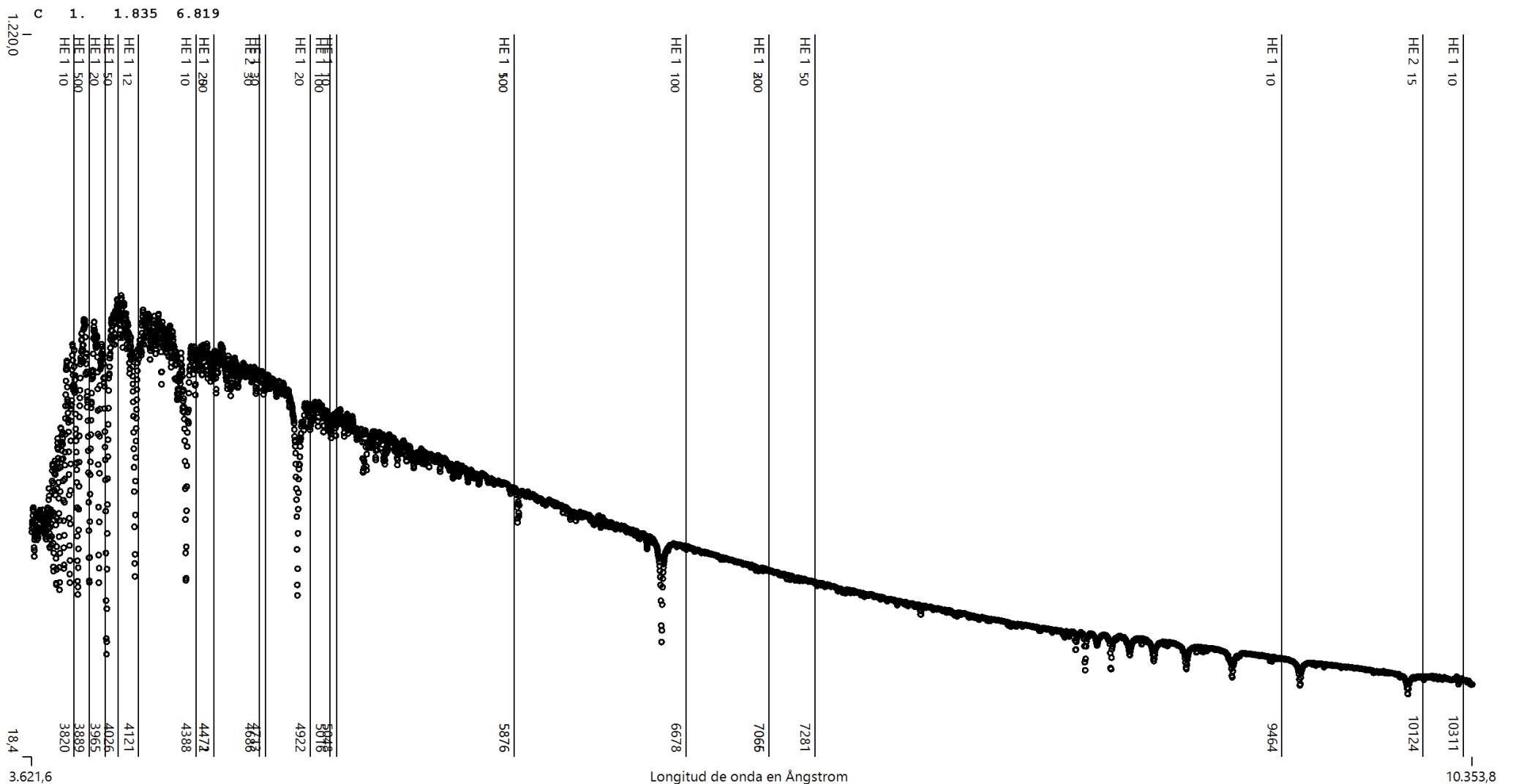


Figura III. 6. Líneas del helio sobre el centroide 1.

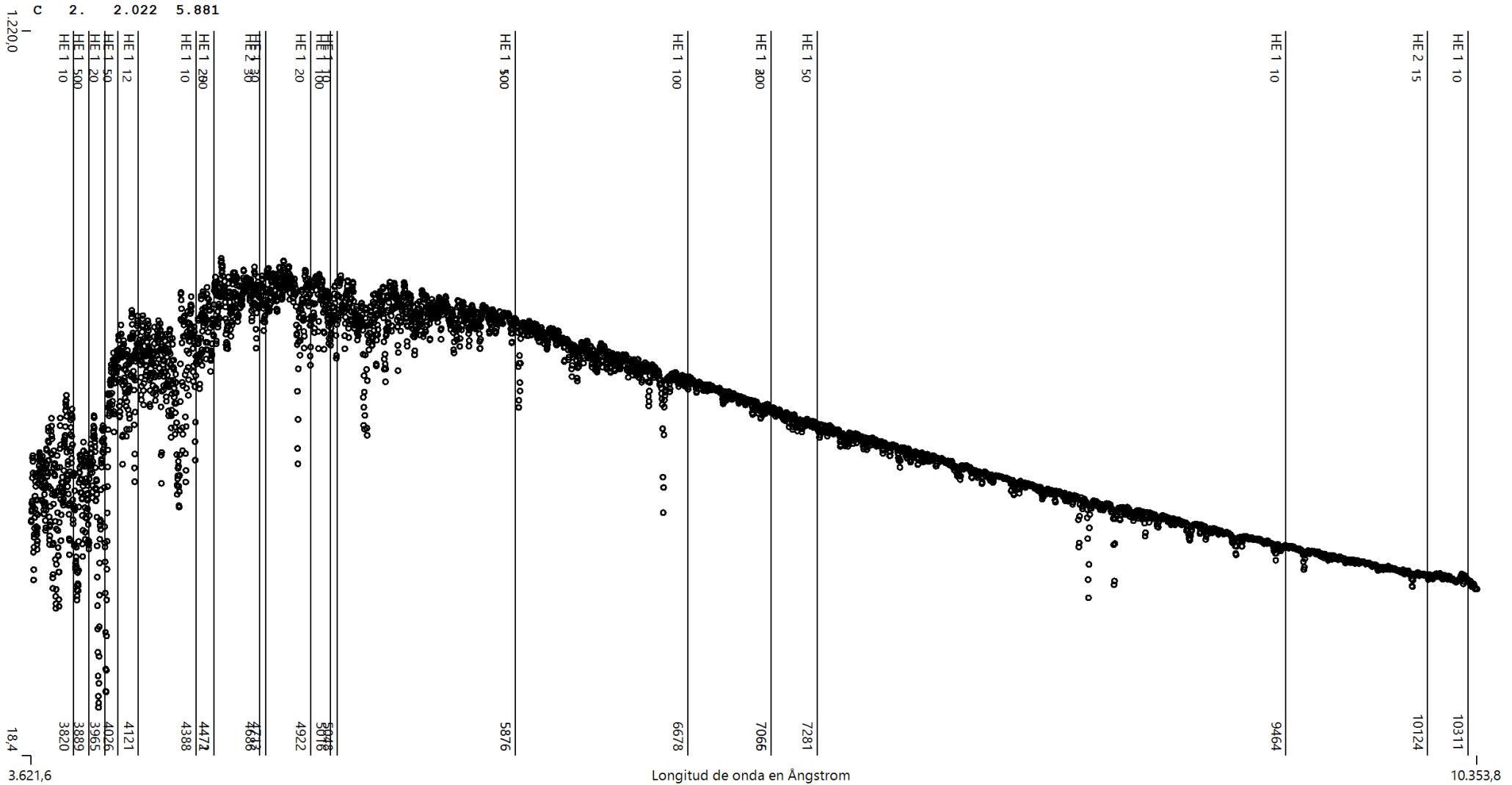


Figura III. 7. Líneas del helio sobre el centroide 2.

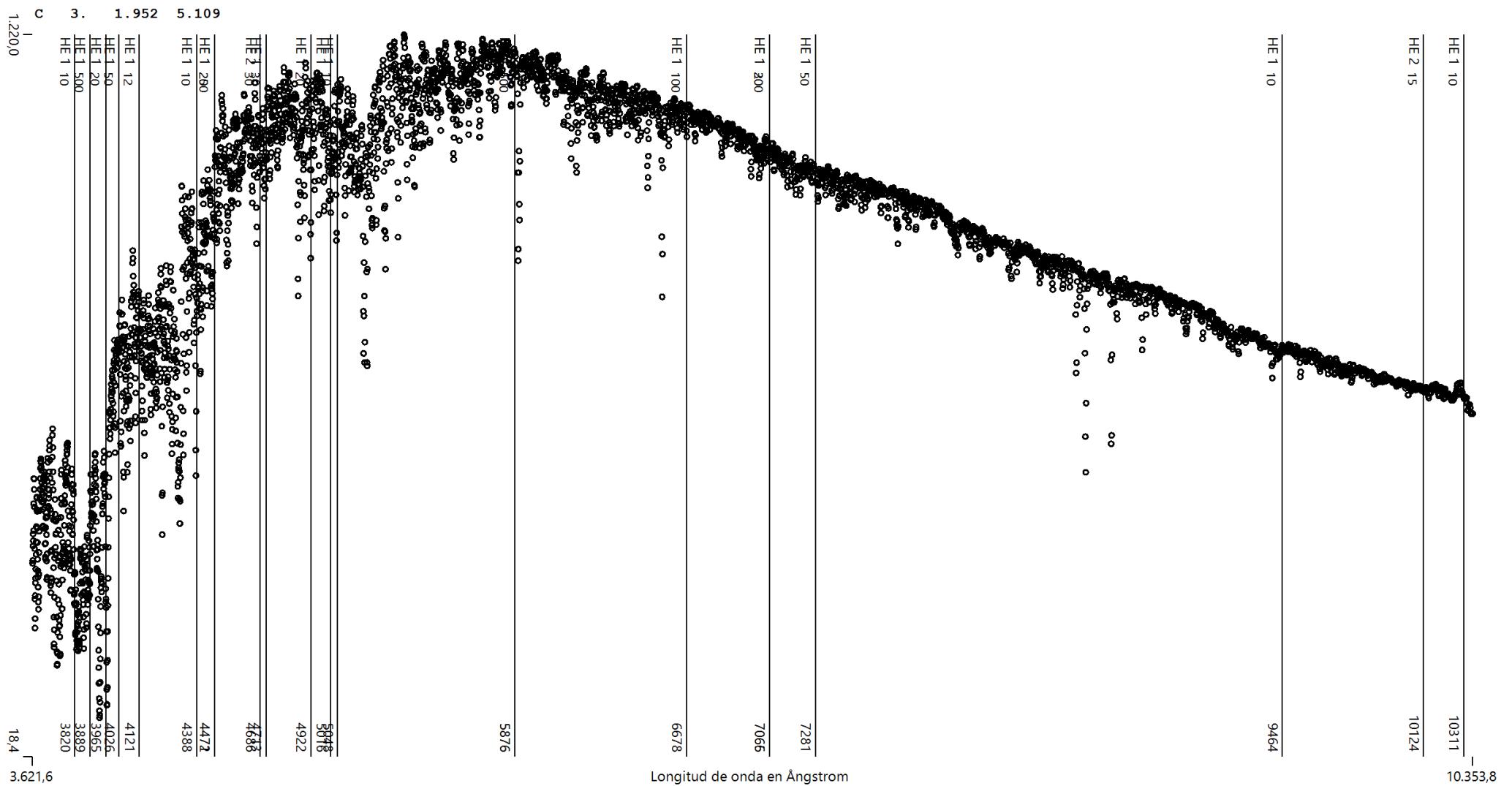


Figura III. 8. Líneas del helio sobre el centroide 3.

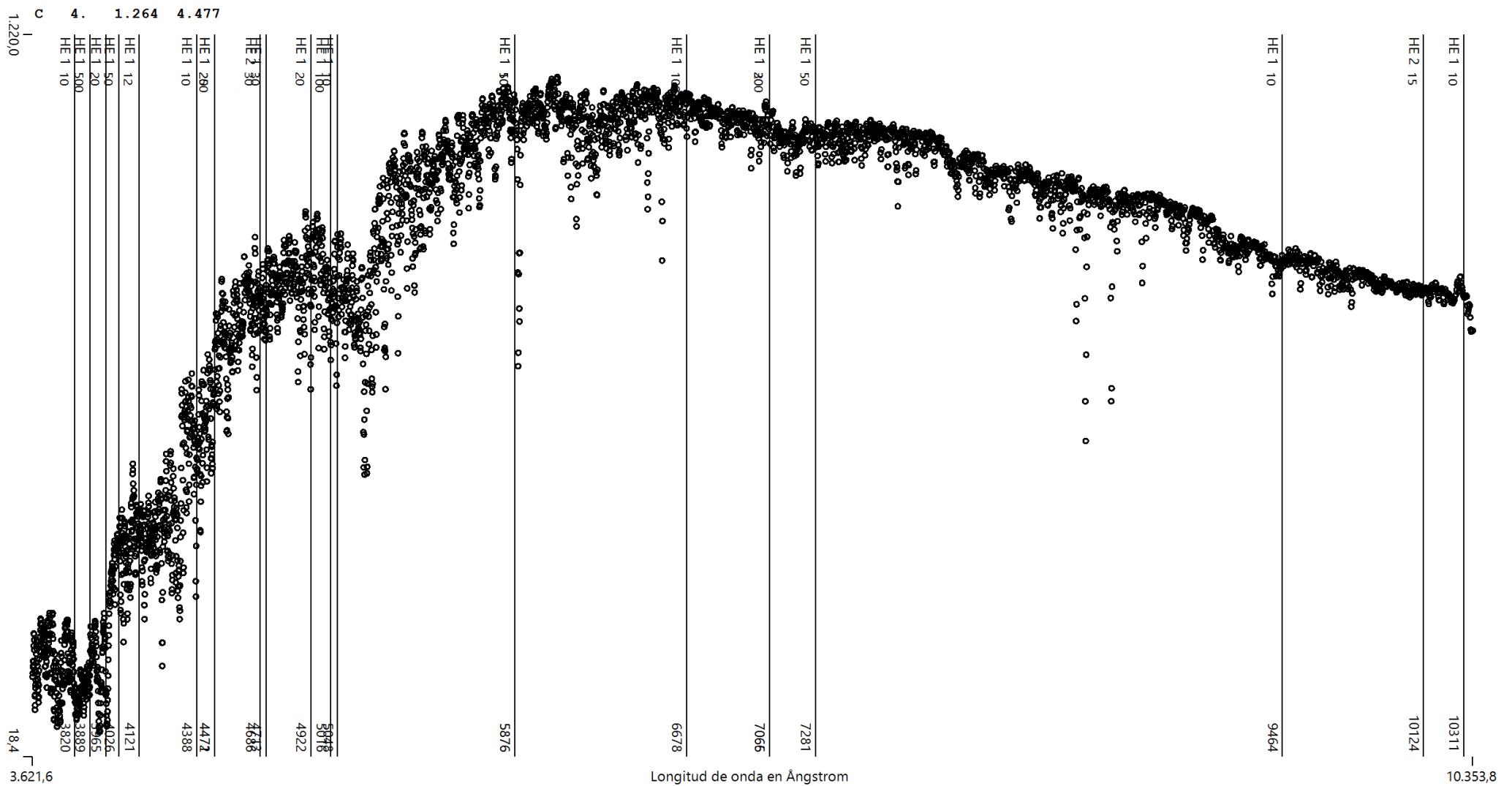


Figura III. 9. Líneas del helio sobre el centroide 4.

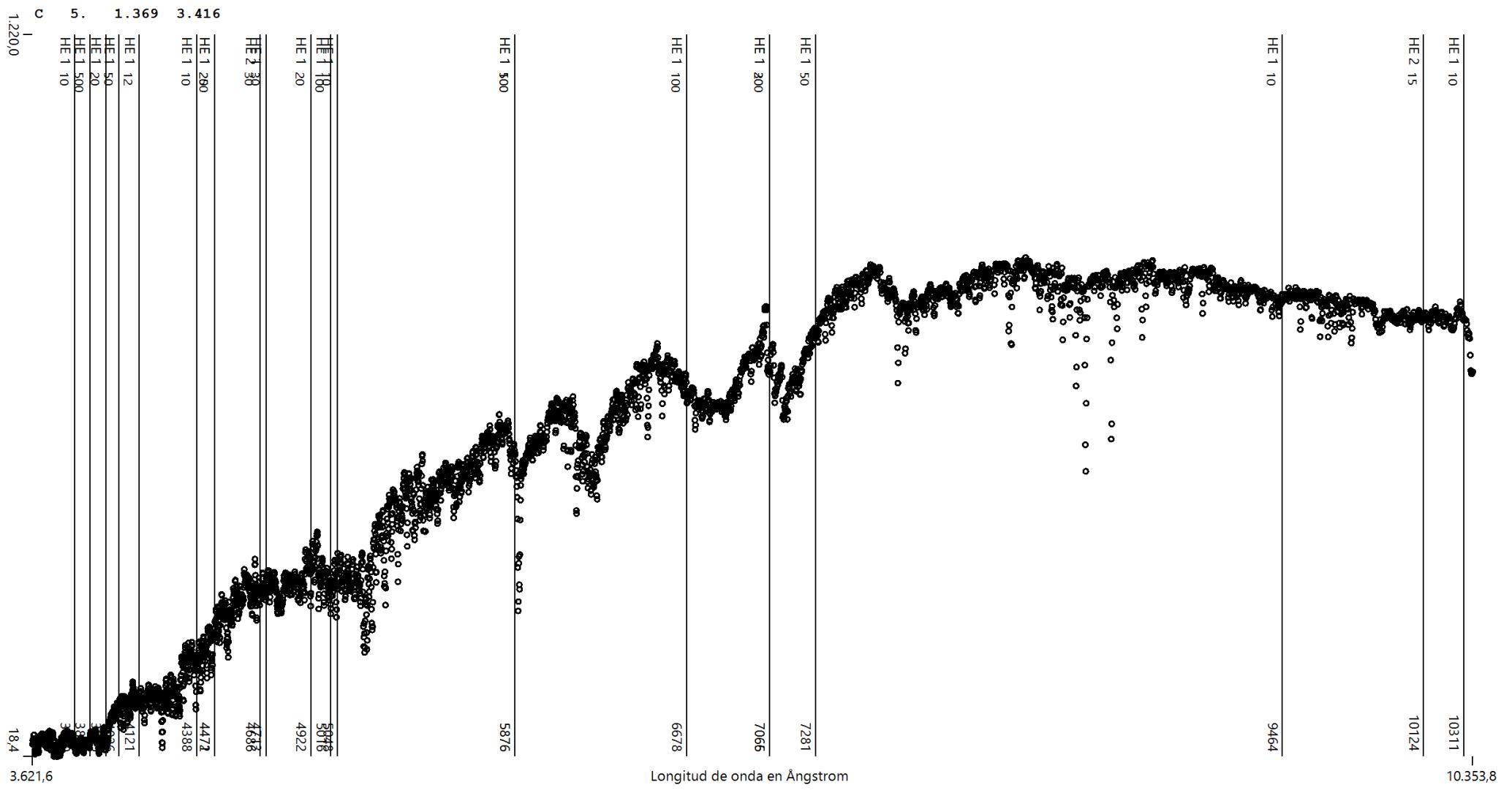


Figura III. 10. Líneas del helio sobre el centroide 5.

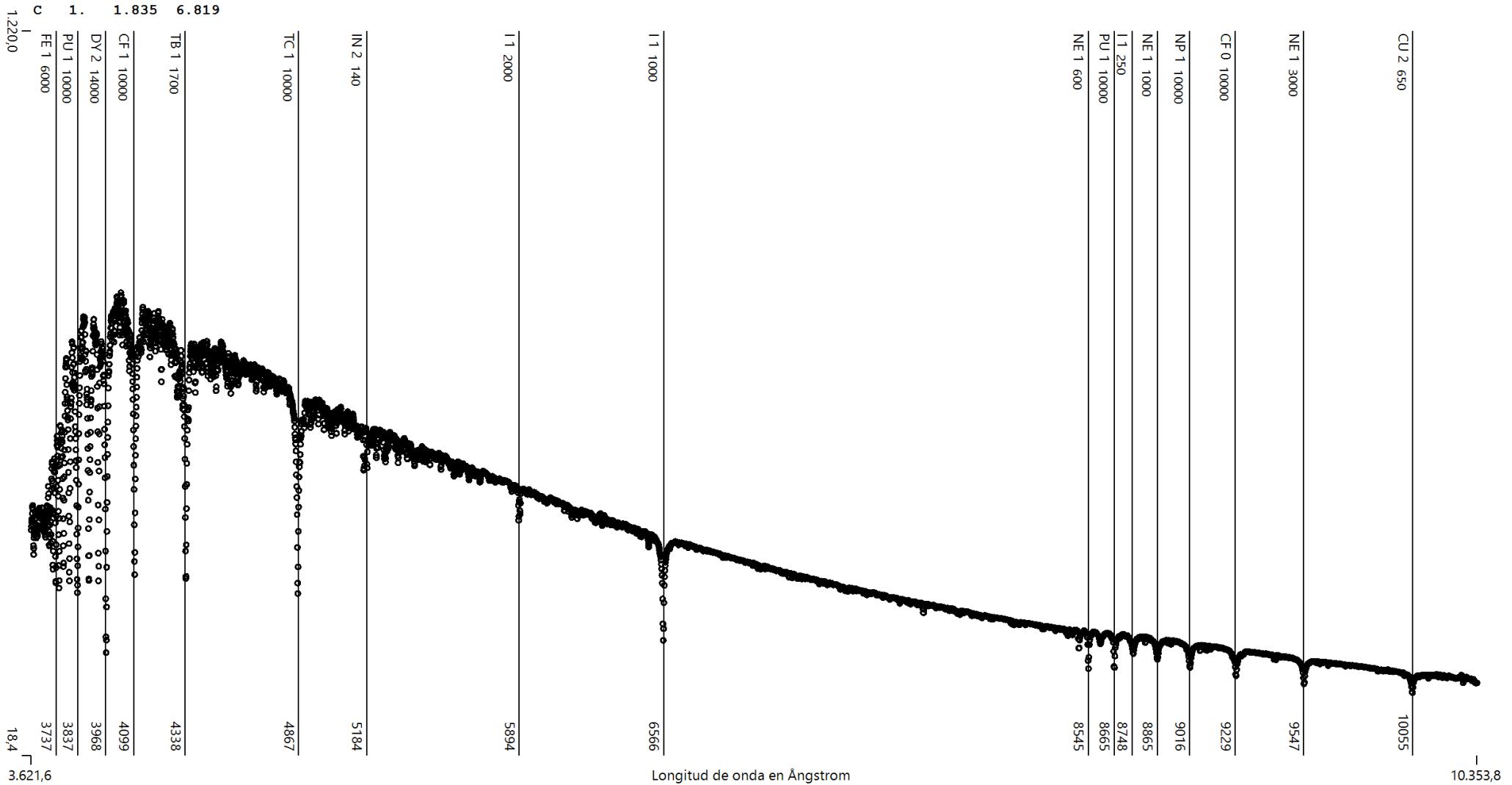


Figura III. 11. Líneas de algunos elementos sobre el centroide 1.

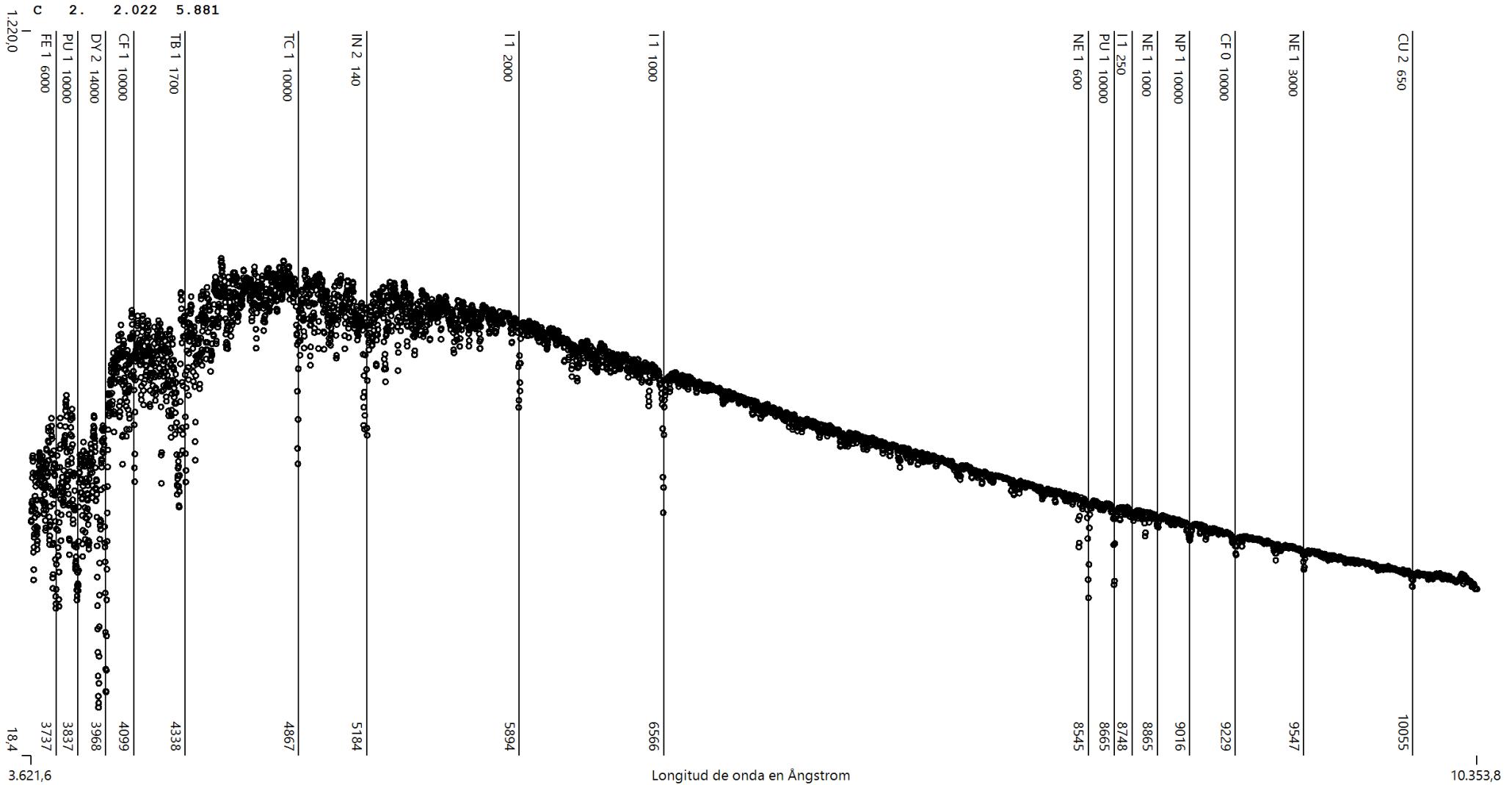


Figura III. 12. Líneas de algunos elementos sobre el centroide 2.

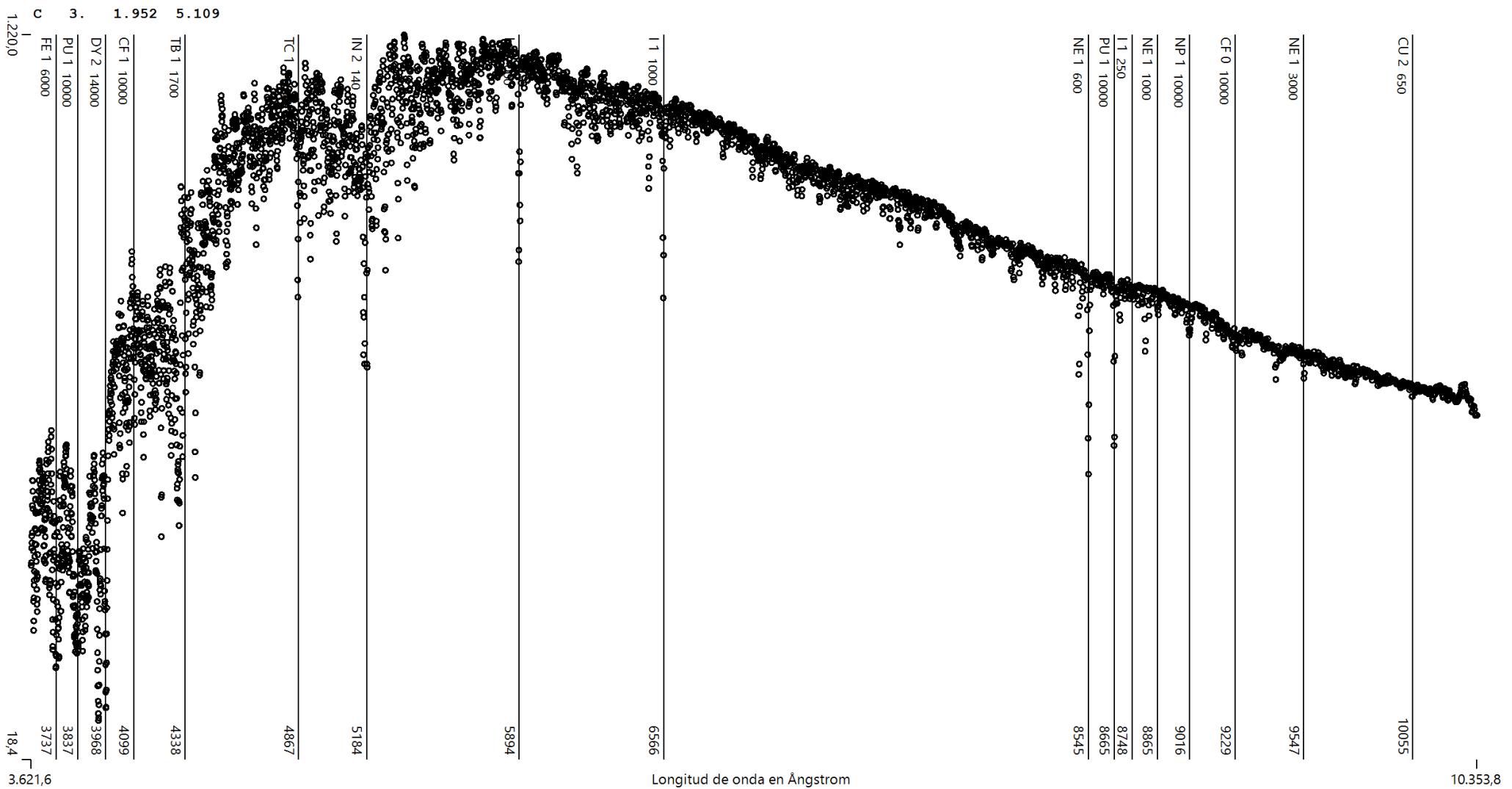


Figura III. 13. Líneas de algunos elementos sobre el centroide 3.

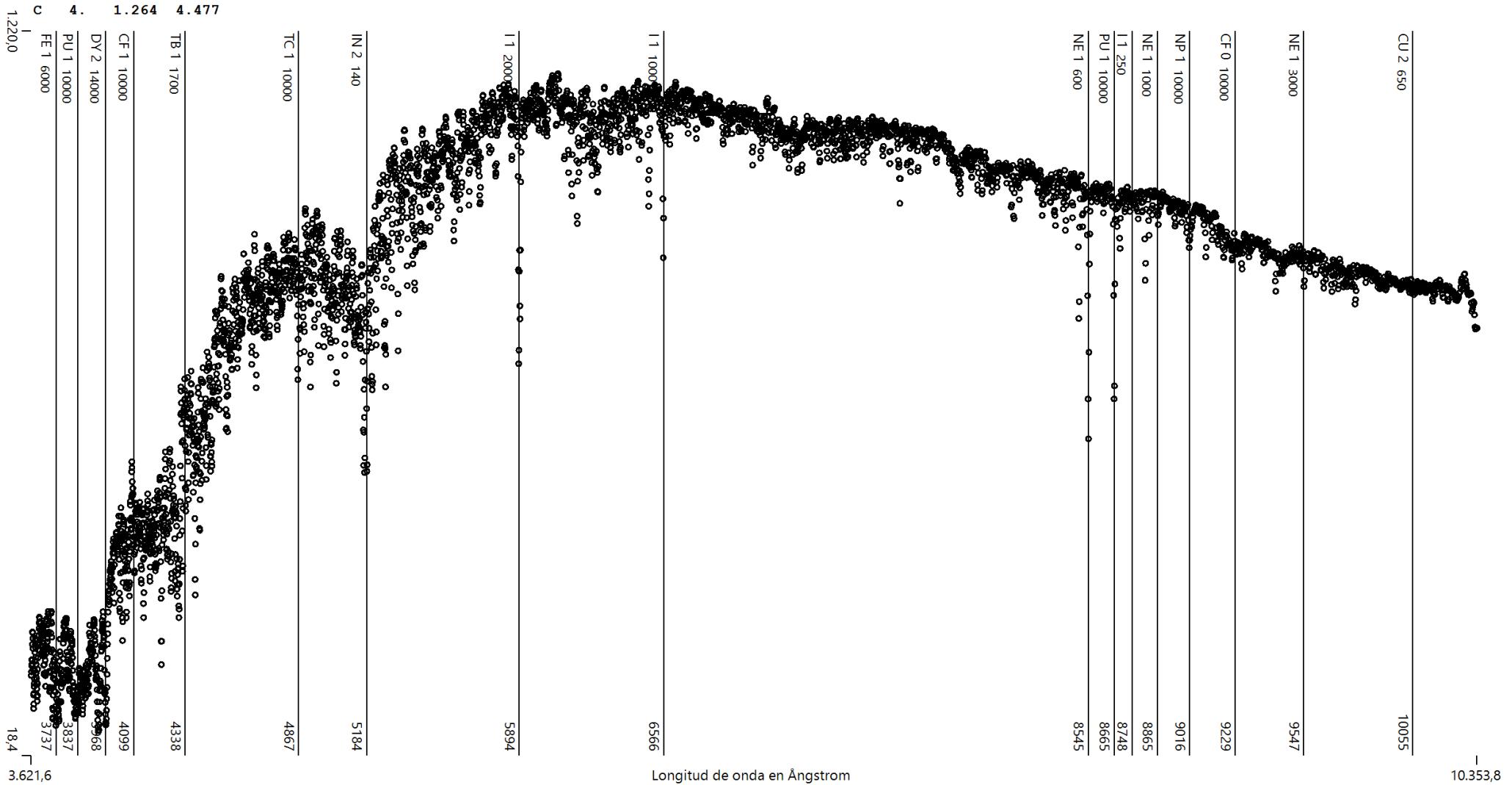


Figura III. 14. Líneas de algunos elementos sobre el centroide 4.

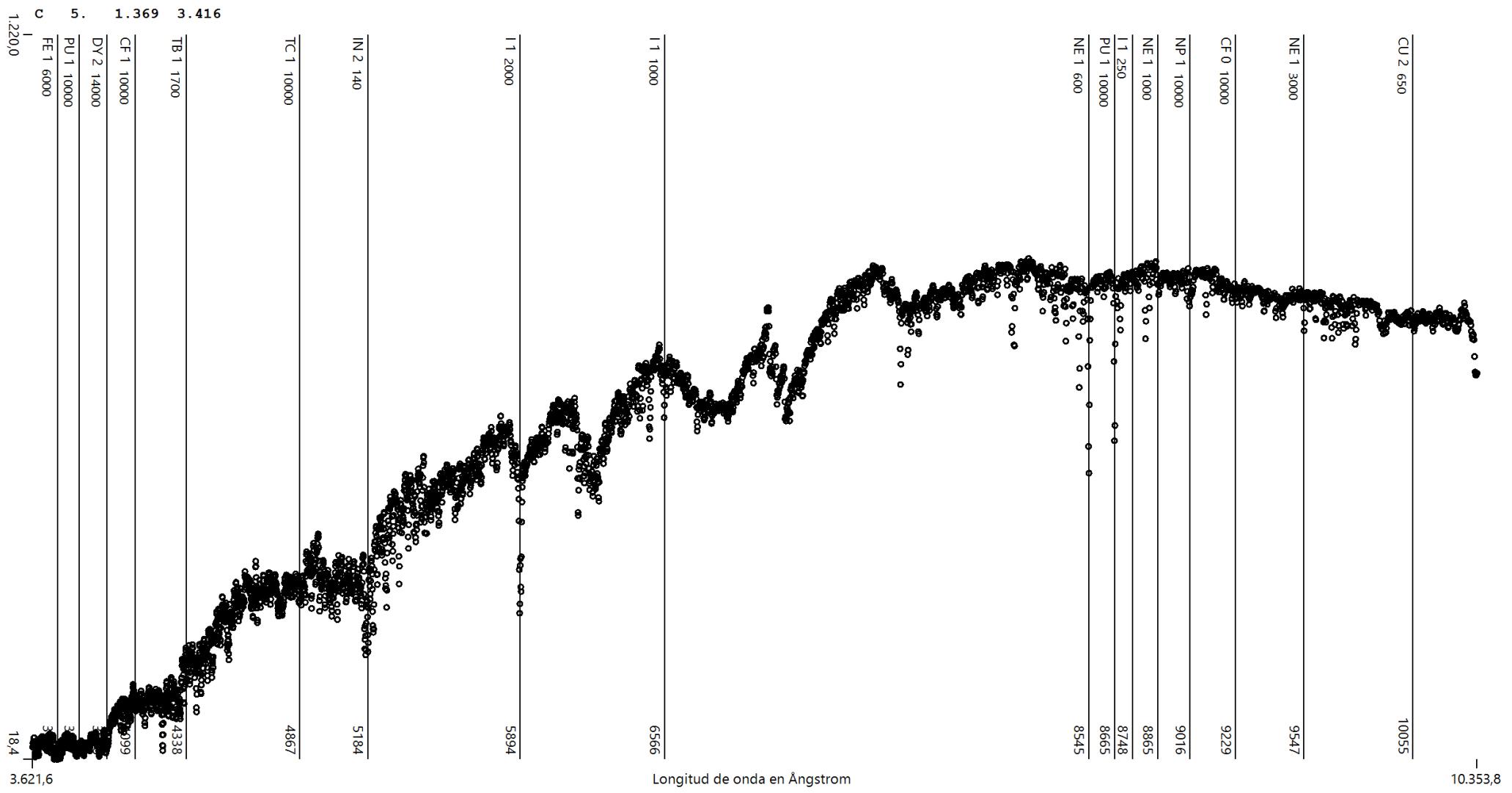


Figura III. 15. Líneas de algunos elementos sobre el centroide 5.