UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Maestría en Explotación de Datos y Descubrimiento de Conocimiento

"Agrupamiento de datos de seguimiento ocular durante el test neuropsicológico Trail Making Test-A"

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires
en el área de Explotación de Datos y Descubrimiento de Conocimiento

Viviana Alejandra Diaz

Director: Dr. Gustavo Gasaneo

Co-director: Dr. Marcelo A. Soria

Lugar de trabajo: Departamento de Matemática
Universidad Nacional del Sur

Fecha de defensa: 19 de diciembre 2024

"Nos estamos ahogando en información hambrientos de conocimiento."
–Rutherford D. Roger

Índice general

Índice de figuras **5**

1. Introducción **2**
1.1. Presentación del problema . . . 2
1.2. Objetivos del trabajo . . . 5
1.3. Contenido y estructura de la tesis . . . 6

2. Datos de estudio **8**

3. Exploración inicial **10**
3.1. La variable tiempo . . . 11
3.2. Variables de posición . . . 12
3.3. Recorrido en pantalla . . . 14

4. Variables derivadas **15**
4.1. Tiempos de resolución. . . . 15
4.2. Longitud, área y velocidad de resolución . . . 16
4.3. Otras variables derivadas . . . 20
4.4. Normalización . . . 22
4.5. Tendencia al agrupamiento . . . 23
4.6. Reducción de la dimensión . . . 23
4.6.1. Selección de variables . . . 24
4.6.2. Correlación . . . 26
4.7. Detección de atípicos . . . 27
4.8. Determinación del número de grupos . . . 29
4.9. Agrupamiento . . . 31
4.9.1. Obtención de grupos . . . 33
4.9.2. Análisis de los agrupamientos . . . 34
4.9.3. Ensambles . . . 40

5. Series temporales **44**
5.1. Series de coordenadas de posición . . . 45
5.1.1. Análisis exploratorio . . . 45
5.1.2. Agrupamiento . . . 46
5.2. Series adaptadas a la resolución del test . . . 49
5.2.1. Series de trials . . . 50
5.2.2. Series de zonas . . . 64

6. Secuencias vectoriales y Multimatch **73**
6.1. Fijaciones y sacadas . . . 73
6.2. El método Multimatch . . . 74
6.2.1. Agrupamiento con Multimatch . . . 75
6.2.2. Multimatch de datos homogéneos . . . 79

7. Aportes del trabajo **83**

8. Trabajo futuro **84**

A. Métodos de selección de variables **85**

B. Técnicas de detección de atípicos **87**

C. Métodos para determinar el número de grupos **89**

D. Técnicas de agrupamiento **91**

E. Métodos de comprensión de clusterings **95**

F. Algoritmo de clasificación de fijaciones y sacadas **96**

G. Método Multimatch **97**

Bibliografía **99**

Índice de figuras

1.1. Imagen de la versión informatizada del TMT-A 4

2.1. Esquema de la posición del sujeto frente a la pantalla durante el TMT 8

3.1. Ejemplo de una trayectoria en el panel diseñado para consulta visual de los datos 11
3.2. Gráficos de caja de las variables de estudio para todos los sujetos 11
3.3. Distribución de las variables espaciales en pantalla 12
3.4. Posiciones medias de recorrido de cada individuo y la mediana de todas ellas 12
3.5. Gráficos de distribución de las variables espaciales del conjunto de individuos 13
3.6. Evolución temporal (en ms) de las coordenadas espaciales de un sujeto 14
3.7. Ejemplo de trayectoria de resolución 14

4.1. Boxplot de tiempos de resolución 16
4.2. Distribución de tiempos de resolución 16
4.3. Valores resumen de las longitudes totales recorridas 18
4.4. Distribución de longitudes recorridas en la resolución 18
4.5. Gráfico de caja de las velocidades de resolución del test 19
4.6. Distribución de velocidades de resolución 19
4.7. Gráfico de caja de las áreas recorridas en la resolución 20
4.8. Distribución de áreas recorridas en la resolución 20
4.9. Medidas estadísticas de los valores de las variables derivadas 21
4.10. Importancia de variables con LASSO 25
4.11. Encabezado del dataset final de variables seleccionadas 26
4.12. Mapa de correlación de Pearson entre las variables derivadas 27
4.13. Codo de silueta para K-Means. 30
4.14. Gap para número de grupos. 30
4.15. Resultado de uno de los agrupamientos jerárquicos. 31
4.16. Resultados de los agrupamientos considerados 34
4.17. Boxplots comparativos de variables en el agrupamiento aglomerativo 36
4.18. Importancia relativa de variables en el agrupamiento aglomerativo 36
4.19. Proyecciones tridimensionales del agrupamiento aglomerativo 37
4.20. Valores centrales de las variables en cada grupo 37
4.21. Árbol de decisión agrupamiento aglomerativo 38
4.22. Distribución por edad en los grupos del método aglomerativo 39
4.23. Cantidad de personas por sexo biológico en los grupos 39
4.24. Distribución de valores de las variables en cada grupo del ensamble 42
4.25. Valores centrales de las variables en cada grupo del ensamble 42
4.26. Proyección de grupos en variables relevantes del ensamble respecto al aglomerativo . . . 42
4.27. Árbol de decisión del agrupamiento por ensamble 43

5.1. Representación gráfica de cuatro series temporales de abscisas 45
5.2. Foto del video de resolución en tiempo real de un sujeto 46
5.3. Encabezado de la tabla con los resultados de los agrupamientos de abscisas 47

5.4. Series y centroides en cada grupo de abscisas . . . 48
5.5. Series y medianas de cada grupo de abscisas con KShape . . . 48
5.6. Series medianas de cada grupo de abscisas . . . 49
5.7. Número de registros por trial . . . 51
5.8. Valores faltantes por sujeto y trial . . . 52
5.9. Estadística de tiempos insumidos por trial . . . 53
5.10. Mediana de los tiempos de resolución por trial . . . 53
5.11. Estadística de velocidades de resolución por trial . . . 54
5.12. Matriz de correlación entre trials . . . 55
5.13. Boxplots de agrupamiento de tiempos de resolución por trial . . . 56
5.14. Centroides por grupo de las series de tiempos de resolución . . . 56
5.15. Boxplots del agrupamiento Time Series K-Means en los trials con grupos distinguibles . 57
5.16. Visualización de todos los sujetos de cada grupo por tiempos de resolución . . . 57
5.17. Imagen del test TMT-A hasta el cuarto trial . . . 58
5.18. Árbol de decisión del agrupamiento de tiempos de resolución . . . 59
5.19. Agrupamiento de tiempos por trial . . . 60
5.20. Estadística de los valores de velocidades de resolución por trial . . . 61
5.21. Medianas de los grupos de velocidades por trial . . . 62
5.22. Gráfico de boxplots de los grupos de velocidades por trial . . . 62
5.23. Gráficos de caja de velocidades por grupo en trials distinguidos . . . 63
5.24. Agrupamientos de velocidades medias por trial . . . 63
5.25. Estadística de los tiempos de resolución por zonas . . . 64
5.26. Medianas de tiempos de resolución por zonas . . . 65
5.27. Comparación estadística de tiempos de resolución por grupos . . . 65
5.28. Gráfico de boxplots de tiempos de resolución en la Zona 1 . . . 66
5.29. Visualizaciones de los agrupamientos de tiempos por zonas . . . 66
5.30. Árbol de decisión del agrupamiento de tiempos por zonas . . . 67
5.31. Estadística de los valores de velocidades medias por zonas . . . 68
5.32. Gráficos de caja de los grupos de velocidades medias por zonas . . . 68
5.33. Visualización de grupos de velocidades medias por zonas . . . 69
5.34. Árbol de decisión del agrupamiento de velocidades medias por trial en cada zona . . . 70
5.35. Resumen estadístico de las velocidades de avance por zona . . . 70
5.36. Resumen estadístico de las velocidades de avance por zona en cada grupo . . . 71
5.37. Visualización completa del agrupamiento con distancia euclídea en cuatro grupos . . . 71
5.38. Resumen estadístico de las velocidades de avance por zona en cada uno de los tres grupos 72
5.39. Resumen estadístico de las velocidades de avance por zona en cada uno de los dos grupos 72

6.1. Ejemplo de scanpath . . . 75
6.2. Mapa de calor y dendogramas de la matrices de forma y duración de fijaciones . . . 76
6.3. Mapa de calor y dendogramas de las matrices de dirección entre sacadas, posición de fijaciones y longitud de sacadas . . . 76
6.4. Mapa de calor de correlación cruzada de variables . . . 78
6.5. Variables separadoras en grupos según forma de la ruta de escaneo visual . . . 80
6.6. Variables separadoras en método espectral de la matriz correspondiente a longitud de sacadas . . . 80
6.7. Gráficos de boxplots de los grupos de la matriz de dirección de sacadas . . . 81
6.8. Boxplots de agrupamientos de la matriz correspondiente a la posición de fijaciones . . . 81
6.9. Variables separadoras del agrupamiento de la matriz asociada a la comparación de acuerdo a la posición de las fijaciones . . . 82
6.10. Variables destacadas en grupos de acuerdo a duración de las fijaciones . . . 82

Resumen

En este trabajo de tesis se realiza un estudio de posibles agrupamientos de datos de seguimiento ocular de 108 personas registrados durante la realización del test atencional conocido como "Trail Making Test-A" (TMT-A). Se desarrolla un análisis de clustering desde tres distintos enfoques. En primer lugar agrupan las trayectorias visuales a partir de una serie de variables obtenidas de los registros de cada individuo que resumen, en algún sentido, la información global de la ruta de escaneo ocular. En el segundo enfoque, se agrupan las posiciones de la mirada en la pantalla de cada individuo como una serie de tiempo bidimensional, y dos series temporales unidimensionales. El tercer punto de vista considera las trayectorias oculares como secuencias de vectores geométricos definidos por las fijaciones y sacadas del escaneo visual. En este caso se utiliza el método multidimensional conocido como MultiMatch para generar medidas de similitud entre las trayectorias espacio-temporales y se agrupan las resoluciones del test a partir de ellas.

El análisis que hemos llevado a cabo brinda un panorama amplio de agrupamiento de este tipo de datos de seguimiento ocular desde distintas ópticas y se han desarrollado métodos de utilidad para detectar trayectorias anómalas que pueden ser de interés para los especialistas. Además, el estudio incluye una perspectiva general de la resolución del test y una más detallada, lo que proporciona su utilización como herramienta clínica en contextos con y sin capacidad de medición electrónica.

“Clustering of Eye-Tracking Data during the neuropsychological Trail Making Test-A”

Abstract

This thesis work conducts a study of potential groupings of eye-tracking data from 108 individuals recorded during the administration of the attentional test known as the ’Trail Making Test-A’ (TMT-A). The analysis of clustering is performed using three different approaches. Firstly, visual trajectories are grouped based on a set of variables derived from each individual’s records, summarizing, in some way, the overall information of the eye-scanning route. Secondly, gaze positions on the screen for each individual are grouped as a bidimensional time series, and as two unidimensional time series. The third perspective considers ocular pathways as sequences of geometric vectors defined by fixations and saccades in the visual scanning. In this case, the multidimensional method known as MultiMatch is employed to generate similarity measures between spatio-temporal trajectories, which are then used for grouping test resolutions.

The analysis conducted offers a comprehensive overview of grouping eye-tracking data from different perspectives, and utility methods have been developed to detect anomalous scanpaths that might be of interest to specialists. Furthermore, the study encompasses a general perspective of test resolution and a more detailed one, thus enabling its use as a clinical tool in both electronic and non-electronic measurement contexts.”

Capítulo 1

Introducción

1.1. Presentación del problema

Es un hecho que la mayoría de las tareas cognitivas implican el uso del sistema visual humano y una gran cantidad de investigaciones avalan la hipótesis de que el estudio del sistema oculomotor brinda información respecto de los procesos mentales y la actividad cerebral [26].

En este contexto, la técnica conocida como seguimiento ocular o "eye tracking" es un método no invasivo que posibilita la detección y el registro del movimiento de los ojos y la dirección visual de una persona mientras ejecuta una tarea. Ésta técnica permite, por ejemplo, analizar el comportamiento ocular, revelar estrategias de búsqueda visual y determinar patrones de atención del sujeto sobre un estímulo dado, y es utilizada para diversos propósitos en una amplia variedad de disciplinas como neurociencia, medicina, psicología, ingeniería, ergonomía, informática, marketing, publicidad, etc. Las primeras estructuraciones de datos de seguimiento ocular se realizaron entre los años 1879 y 1920 cuando los registros de posición visual fueron agrupados según su comportamiento, clasificándose en fijaciones (espacios donde el ojo se mantiene relativamente quieto, dirigiendo la fóvea hacia el área de interés) y sacadas (movimientos rápidos que llevan el ojo de un área de interés a otra durante los cuales la entrada visual está suprimida). Actualmente, un importante número de investigaciones en neurociencia están implementando técnicas de eye tracking como herramienta de análisis o diagnóstico de patologías como el autismo, el Parkinson, el Alzheimer o la esclerosis lateral amiotrófica, entre otras [8, 16, 29, 38].

Un trabajo de Lai y colaboradores [22] revisa investigaciones realizadas a lo largo de doce años en las que la técnica de eye tracking se utilizó para estudiar cuestiones vinculadas al aprendizaje. Los autores proponen una clasificación de estas investigaciones de acuerdo al tema estudiado: procesamiento de información, estrategias según instrucciones, revisión de teorías existentes, diferencias individuales, estrategias de aprendizaje, toma de decisiones y desarrollo conceptual. En la misma línea, en un artículo de revisión bibliográfica de Luna y coautores [26] pueden encontrarse referencias a una gran cantidad de investigaciones que permiten asegurar que el estudio de los movimientos oculares puede considerarse una herramienta muy útil en el análisis del desarrollo cognitivo. Varias líneas de investigación trabajan con la hipótesis de que los movimientos oculares realizados por un sujeto al resolver tareas que pongan en juego capacidades atencionales permiten caracterizar su desempeño. Esta caracterización, de la que son de interés algunos aspectos cuantitativos, podría permitir reconocer las fortalezas y dificultades en sus habilidades cognitivas.

Una gran cantidad de investigadores trabajan con el convencimiento de que registrar el comportamiento de los ojos en muchas de las pruebas psicológicas puede proporcionar información sobre cómo las personas procesan la información, sobre cómo funciona el cerebro en relación con muchas de las actividades diarias, y cómo se realizan y organizan las tareas cognitivas. Sin embargo, a pesar de que en muchas de las pruebas neuropsicológicas en las que la visión constituye el canal para ingresar la información externa, en la mayor parte de los casos los movimientos oculares realizados por los sujetos durante el análisis y la ejecución de la tarea no se registran. En este marco, el método de seguimiento ocular o eye tracking permite registrar y analizar el comportamiento ocular de una persona mientras realiza una prueba psicológica con el objetivo de conseguir un entendimiento profundo de los distintos

procesos cognitivos involucrados en la resolución de problemas y la toma de decisiones, razón por la cual el seguimiento ocular se ha comenzado a utilizar en neurociencia como una herramienta de estudio del proceso cognitivo que interviene en el comportamiento atencional, memorístico, perceptivo, lingüístico, etc.

En el ámbito clínico de la psicología y la psicopedagogía se han desarrollado a lo largo del tiempo diversas herramientas con las que se evalúan distintas problemáticas vinculadas al aprendizaje o a capacidades vinculadas a la función cognitiva. Para el estudio de problemas atencionales se han diseñado diversos test que incluyen por ejemplo al test Caras, al Stroop, a las cartas de Wisconsin y al Trail Making Test o Test de Rastreos. Dichos test conforman una batería de evaluación atencional que discrimina distintos aspectos de la atención: atención focalizada, sostenida, selectiva y adaptativa.

En particular, el test conocido como "Trail Making Test" (TMT) [33] se trata de una prueba neuropsicológica clásica cuyo tiempo de realización es muy breve (unas pocas decenas de segundos) creada por Partington en 1938, originalmente para realizarla mediante lápiz y papel. En un primer momento el TMT se construyó como parte de un test de inteligencia para ser utilizado en la distinción de habilidades de los soldados enlistados en la armada estadounidense pero que luego se comprendió que se trataba de una prueba atencional. Posteriormente, su uso se extendió hasta la actualidad debido a que hay consenso en que para llevar a cabo la tarea del TMT se requiere de la indemnidad de diversos aspectos atencionales (concentración o sostenibilidad de atención, atención selectiva y atención alternante).

El Trail Making Test es una prueba relacionada a una tarea de búsqueda visual que está compuesto por dos partes, llamadas A y B. La versión para adultos del TMT-A consiste de una imagen con el conjunto de los primeros 25 números naturales distribuidos con una cierta disposición espiralada (ver figura), tradicionalmente en una hoja de papel, o en la pantalla en la versión informatizada. La tarea consiste en conectar ordenadamente los números de la secuencia lo más rápido posible. La prueba es cronometrada y en la consigna se indica que debe realizarse lo más rápido posible. En el caso de los niños, la versión estándar tiene 15 números.

El TMT es una de las pruebas cognitivas más utilizadas para evaluar los procesos atencionales y su resolución requiere de múltiples aspectos relacionados a la atención que subyacen a la ejecución correcta de la tarea. Este test implica flexibilidad cognitiva añadiendo presión temporal sobre la ejecución. El objetivo del test es evaluar la velocidad de ubicación visual, el escaneo, la velocidad de procesamiento, la atención, la flexibilidad mental, la memoria de trabajo y la función motora. Además, puede proporcionar información sobre la velocidad de búsqueda visual, escaneo, velocidad de procesamiento, funcionamiento ejecutivo. El TMT ha mostrado ser útil para detectar el deterioro cognitivo asociado al mal funcionamiento de la corteza prefrontal y alteraciones cognitivas de la migraña, entre muchas otras patologías. Estos test arrojan resultados numéricos que permiten ubicar al sujeto dentro de un grupo de referencia, sin embargo, nada nos dicen respecto de la forma en que un individuo resuelve el test, esto es, la estrategia seguida para desarrollar la tarea. No es posible saber, por ejemplo, de qué manera el sujeto exploró la imagen, en qué elementos se detuvo o cuáles ignoró, cuánto tiempo observó ciertos elementos, qué elementos revisó varias veces, etc. Es por esto que se desarrolló la idea de registrar los movimientos oculares durantes estas pruebas para obtener información más detallada de la actividad ocular durante la tarea.

Los datos de seguimiento ocular recolectados durante el desarrollo del test TMT-A en su versión en papel han sido estudiados en el artículo [20], donde fueron medidos los movimientos oculares con una cámara con el fin de analizar las diferencias en la aplicación y ejecución de la traza de cada sujeto, y examinar los dibujos de inspección como tareas de la búsqueda visual. Otro trabajo en que fueron estudiados los datos del Trail Making Test es el citado en la referencia [2], donde se producen datos normativos de la población portuguesa en relación al test y se muestra que el sexo, la edad y la educación son variables que están significativamente asociadas al desempeño en el TMT.

Avanzando con esta idea de recoger los movimientos oculares de los sujetos mientras se realiza el Trail Making Test-A, se diseñó una versión informatizada del mismo respetando estrictamente la disposición original de números en el papel. La prueba se presenta en una pantalla conectada a un computadora donde se utiliza un software personalizado para presentar y registrar toda la información a extraer, incluidos los movimientos del ojo. La Figura 1.1 muestra la pantalla que ve el participante al comenzar a realizar el test TMT-A. Antes de comenzar la prueba, se le indica al sujeto que presione la barra

espaciadora cada vez que encuentra un número de la secuencia y se le muestra la posición del número uno en la pantalla. Mientras el sujeto realiza la prueba, se registran sus movimientos oculares, así como el momento en que se presiona la barra espaciadora.

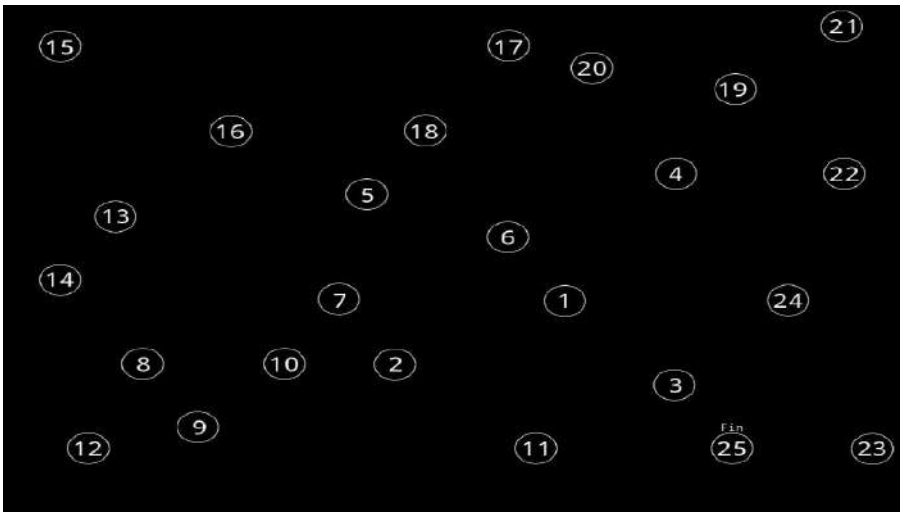


Figura 1.1: Imagen de la versión informatizada del TMT-A

La implementación de un dispositivo de seguimiento ocular en la versión digitalizada de un test psicológico como el TMT permite evaluar muchos elementos más que los tradicionalmente considerados en la aplicación de lápiz y papel. Gracias a la tecnología, es posible saber con alta precisión los tiempos que cada sujeto demora en resolver cada etapa del test (en el caso del TMT, las etapas son los intervalos entre el hallazgo de un número y el siguiente de la serie), o el número de fijaciones que realiza, la ubicación y la duración de las mismas, o incluso es posible conocer la trayectoria que describe con la mirada y calcular, por ejemplo, su longitud para tomarla como un nuevo elemento a analizar. Estos datos permiten un estudio más profundo del desempeño del sujeto, habilitando el análisis de cada etapa por separado o de su evolución a lo largo de la resolución de la prueba.

En el trabajo de Dahmen et al. [4] se muestran estudios que indican que la versión informatizada del TMT es capaz de medir los mismos aspectos de la cognición que la prueba en papel. Además, los datos adicionales del TMT informatizado pueden ayudar a monitorear otros procesos cognitivos no capturados por el TMT en papel. Más recientemente, en el artículo de Linari et al. [23] se muestra una comparación en el rendimiento de ciertos sujetos en las dos versiones del test TMT, la tradicional y la digitalizada.

Otros trabajos que estudiaron la evolución de los movimientos oculares de sujetos durante la realización del TMT-A y permitieron caracterizar de manera cuantitativa los diversos comportamientos encontrados en cuanto a las estrategias de búsqueda, fueron realizados por el grupo NEUFISUR con el que se realizó en colaboración el trabajo de esta tesis. Este grupo de investigación tiene lugar de trabajo de la Universidad Nacional del Sur (UNS) e Instituto de Física del Sur (IFISUR) y estudia los movimientos oculares en relación a cuestiones de psicología cognitiva, neurociencias y salud ocular, entre otros temas. Este grupo ha definido parámetros vinculados con la velocidad de procesamiento y un índice de búsqueda que permitió cuantificar períodos de intensa búsqueda o períodos de resolución más directa. Estos resultados se presentaron en la 18th International Conference on Cognitive Modeling (ICCM, julio de 2020) y se utilizaron para el estudio de la evolución de un paciente, resultados que fueron expuestos por el grupo en el Congreso Anual de la Sociedad Argentina de Investigación en Neurociencias 2022.

La mayoría de los estudios de eye tracking que se interesan por aspectos cognitivos tienen como objetivo identificar y analizar patrones de atención visual de individuos mientras desarrollan tareas específicas (por ejemplo, leer, buscar, escanear una imagen, etc). Para hallar patrones ocultos son muy útiles las técnicas de aprendizaje no supervisado que detectan la estructura subyacente en datos sin etiquetar y muchas veces abren paso para un posterior aprendizaje supervisado, por lo que configuran

un muy buen primer tratamiento de los datos. También resulta útil el aprendizaje no supervisado en el caso en que no se dispone de datos de entrenamiento, como es el caso de este trabajo de tesis. Dentro de las técnicas de aprendizaje no supervisado, en particular el agrupamiento o "clustering" es uno de los métodos más comunes que permite dar una clasificación a cada entrada de datos sin predefinir las diferentes clases. El análisis de clústers o conglomerados es un método propio del análisis exploratorio de datos, que permite descubrir asociaciones y estructuras en los datos que no son evidentes pero que pueden ser útiles una vez que se han detectado. Esta es una técnica esencial en cualquier campo de investigación que involucre analizar o procesar datos multivariados y hay muchas aplicaciones prácticas del agrupamiento, entre las que podemos mencionar minería de datos, taxonomía, reconocimiento de patrones, análisis y segmentación de imágenes, comunicaciones, recuperación de información, bioinformática, psicología, compresión y clasificación de datos, redes sociales, sistemas de recomendación, negocios y marketing, etc. Se recurre a técnicas de agrupamiento cuando se desea identificar grupos naturales en las observaciones de un conjunto de datos no etiquetados, es decir, cuando se busca agrupar registros similares, lo que ayuda a los analistas a realizar una segmentación del conjunto de datos subdividiendo los datos objeto en grupos más pequeños conocidos como clusters o conglomerados, de manera que cada grupo exhiba un alto grado de similitud intra-grupo y un alto grado de disimilitud entre distintos grupos. En otras palabras, el análisis de conglomerados o agrupamientos es la tarea de detectar grupos, desconocidos a priori, en un conjunto de objetos de manera que los objetos del mismo grupo sean más similares o más cercanamente relacionados entre sí (bajo algún criterio) que lo que lo son respecto a los objetos asignados a los otros grupos, conglomerados o clusters. En algún sentido, los elementos del mismo grupo comparten el mayor número posible de características y los objetos en diferentes grupos tienden a ser distintos. Es decir, cada grupo representa objetos con propiedades sustancialmente diferentes.

La técnica de agrupamiento pretende la subdivisión del conjunto de datos en subgrupos representativos. Sus objetivos radican en obtener una representación más compacta de los datos, realizar una clasificación o, simplemente, alcanzar una mayor comprensión de la estructura de los mismos a partir de la clasificación provista por el agrupamiento. Luego de determinado el número de grupos y divididos los datos, interesará analizar los rasgos característicos que distinguen a cada uno de ellos. De esta manera, una vez que se identifican los distintos grupos, la tarea del analista es explorar qué hace que cada uno de ellos sean especiales y diferentes de otros grupos.

1.2. Objetivos del trabajo

En el contexto mencionado en la introducción, en este trabajo de tesis de maestría se trata de estudiar posibles agrupamientos de los datos de seguimiento ocular producidos durante la realización del test atencional TMT-A en versión digitalizada de ciertos individuos con cuyos registros se cuenta y analizar los resultados en búsqueda de patrones de resolución de la prueba.

En diversos estudios de datos de seguimiento ocular se ha mostrado la dificultad que tiene el tratamiento espacio-temporal de los datos y se ha enfatizado la necesidad de preservar de manera confiable la información espacial y temporal simultáneamente. Como se mencionó, la idea de esta tesis es agrupar los desempeños de los individuos utilizando técnicas de aprendizaje automático no supervisado que permitan descubrir la estructura subyacente de los datos considerando su carácter espacio-temporal. No solo se busca hallar posibles grupos de las trayectorias visuales de resolución del test según su posición en los distintos tiempos de medición, sino también agrupamientos de los datos de cada individuo considerados como una secuencia completa de posiciones en el tiempo de la prueba, para lograr de esta manera identificar posibles patrones de resolución del TMT-A y agrupar los individuos de acuerdo a esos patrones.

Concretamente, el principal objetivo de trabajo de esta tesis es detectar diferencias y similitudes entre las distintas personas durante la resolución del TMT-A. Cabe mencionar que este análisis se plantea considerando la naturaleza espacio-temporal del movimiento visual. Además, se estudian posibles patrones en la dinámica de los movimientos oculares, individuos por fuera de esos patrones, grupos según tiempos y/o métodos de resolución y posibles estrategias de búsqueda.

El hecho de que puedan descubrirse patrones generales de resolución del TMT-A podría colaborar a la construcción de posibles modelos de comportamiento visual. También podría ser de utilidad en la detección de estrategia de resolución del test de evaluación atencional por parte de los sujetos, acortando las distancias camino a una posible caracterización de distintos grupos de individuos según sus métodos de resolución del test con propósitos clínicos, como puede ser el diagnóstico de problemas atencionales u otras patologías neurológicas.

1.3. Contenido y estructura de la tesis

Con el propósito mencionado de caracterizar los datos de seguimiento ocular durante la realización del TMT-A, en esta tesis se aplicaron técnicas de aprendizaje automático no supervisado, básicamente métodos de "clustering", a los registros de las mediciones de 108 personas durante la realización de la versión informatizada del Trail Making Test-A, obteniendo agrupamientos de los sujetos en relación a sus movimientos oculares durante la prueba.

A modo de exploración inicial de los datos de estudio, se realizó una representación visual interactiva de los registros de seguimiento ocular de que se dispone. Esta visualización muestra la información espacio-temporal de las resoluciones observadas del test, tanto en forma particular por individuo como global del conjunto de sujetos. De esta manera, la representación fue de utilidad durante todo el estudio de esta tesis, dando origen a ideas de análisis y facilitando la visualización de patrones.

Se continuó la exploración inicial de los datos llevando a cabo un análisis cuantitativo de las medidas estadísticas y distribuciones de los tiempos y posiciones registrados.

Luego de esta exploración, entendiendo más profundamente la naturaleza de los datos de estudio y atendiendo al objetivo de detectar patrones globales de resolución, se definieron variables derivadas a partir de los datos originales de tiempo de medición y posición de la mirada que registró el dispositivo de seguimiento ocular. En el Capítulo 4 se consideran estas variables derivadas que, de alguna manera, resumen ciertas características del comportamiento espacio-temporal del movimiento ocular durante el test, y se procede a realizar el estudio de agrupamiento en relación a éstas variables derivadas utilizando varios algoritmos de agrupamiento en búsqueda de posibles patrones de estrategias de resolución distinguibles. Previamente al análisis de agrupamiento con este enfoque, se realiza un estudio de posibles elementos atípicos en relación a estas variables derivadas con el objetivo de detectar la posible existencia de individuos alejados de los patrones de comportamiento visual de búsqueda, que se diferencien de la mayoría de los individuos estudiados en su movimiento ocular al resolver el test TMT-A.

El segundo punto de vista que se abordó para el agrupamiento es la consideración de las secuencias de resolución del test de los sujetos como series temporales. En primer lugar, se consideraron las series de las variables espaciales de posición horizontal y vertical en pantalla para cada tiempo de medición (series $x(t)$ e $y(t)$) y se realiza un análisis de clustering de éstas series temporales. Luego, en un segundo tratamiento dentro del enfoque de series temporales, se analizan agrupamientos de las series de tiempos y velocidades por etapas y zonas características del test. En este caso, la variable tiempo y velocidad son función discreta de cada trial y cada zona en que se divide la prueba. Este es el enfoque del Capítulo 5.

En el Capítulo 6 se realiza un estudio de clustering de las trayectorias de los sujetos consideras como un todo a partir de las nociones de fijación y sacada definidas en el estudio de movimiento ocular. En este enfoque, se analiza cada recorrido de la mirada durante la resolución como una secuencia de vectores definidos entre dos fijaciones consecutivas temporalmente. Utilizando el método conocido como MultiMatch, especialmente diseñado para comparar rutas de escaneo de seguimiento ocular en múltiples dimensiones desde un punto de vista geométrico, se obtienen cinco medidas de similitud distintas entre cada par de trayectorias y se agrupan los individuos a partir de los valores de estos índices de similaridad entre sus secuencias de movimientos oculares.

Adicionalmente, en varios casos, los agrupamientos obtenidos en los distintos enfoques fueron combinados y robustecidos mediante técnicas de ensamble, y una vez obtenidos los resultados de los distintos métodos de agrupamiento que se aplicaron, se analizaron posibles características comunes entre las resoluciones de cada grupo en los distintos enfoques, es decir, características distintivas de los grupos

obtenidos a partir de las variables derivadas de los registros de mediciones del dispositivo para cada trayectoria, considerando las posiciones de la mirada en pantalla como series temporales y utilizando las nociones de similaridad entre recorridos que brinda el método MultiMatch.

Capítulo 2

Datos de estudio

Las mediciones de seguimiento ocular que son objeto de estudio en esta tesis fueron obtenidas conjuntamente por el grupo de investigación NEUFISUR de la Universidad Nacional del Sur y profesionales del Centro Integral de Neurociencias Aplicadas de Bahía Blanca (CINA). El registro de los datos se realizó mediante el uso de la plataforma PsiMesh, desarrollada por profesionales de CINA, y utilizando un registrador Tobii disponible en dicho centro. De esta manera se han recolectado los datos de los movimientos oculares de 108 personas, registrando la posición en la pantalla a la que se dirige la mirada de los participantes mientras realizan la versión informatizada del test conocido como TMT-A una única vez. La prueba fue cronometrada y en la consigna se indicó que la tarea debía realizarse lo más rápido posible. La información sobre el sexo biológico de los participantes fue recabada de manera parcial. Se dispone de datos para 56 individuos, de los cuales 26 son hombres y 30 mujeres. En los 52 casos restantes no se cuenta con registro del sexo biológico. Asimismo, se obtuvo información etaria únicamente para 56 individuos, cuyas edades están comprendidas entre 7 y 49 años con la siguiente distribución: 1 persona de 7 años, 2 de 11 años, 2 de 14 años, 1 de 18 años, 4 de 19 años, 13 de 20 años, 13 de 21 años, 4 de 22 años, 3 de 24 años, 4 de 26 años, 3 de 27 años, 2 de 29 años, 1 de 30 años, 1 de 37 años, 1 de 45 años y 1 de 49 años. No se dispone de información etaria para los 52 participantes restantes. Todos los participantes están alfabetizados de acuerdo a su edad.

Cabe mencionar que los individuos participaron voluntaria y gratuitamente en la prueba. Todos ellos, o su tutor legal en el caso de los niños, con anterioridad a la realización del test, firmaron el correspondiente consentimiento informado para que sus datos se utilicen con fines de investigación y los protocolos utilizados para la obtención de los registros han sido avalados por el comité de Bioética del Hospital Dr. Leónidas Lucero de la ciudad de Bahía Blanca. Por otra parte, todos los sujetos fueron tratados de acuerdo con la Declaración de Helsinki que incluye un apartado de privacidad y confidencialidad.

Previamente a la realización de la prueba las personas recibieron una breve explicación oral del funcionamiento del dispositivo de seguimiento ocular y una descripción del test a resolver. A continuación se les dieron indicaciones respecto a la manera adecuada de resolver el test para que las señales puedan ser registradas correctamente. De todos modos, las instrucciones del experimento están indicadas sobre la pantalla al inicio del test. Los participantes realizaron el test sentados frente a una notebook a una distancia aproximada de 45 centímetros de la pantalla (como muestra esquemáticamente la Figura 2).

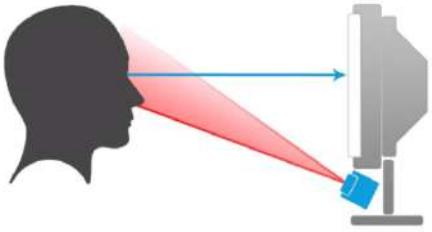


Figura 2.1: Esquema de la posición del sujeto frente a la pantalla durante el TMT

El dispositivo de seguimiento ocular utilizado por los investigadores para la experimentación registra las coordenadas (x, y) de la posición media de ambos ojos de la dirección en la que apuntan las pupilas sobre la pantalla de la notebook, medida en píxeles normalizada, según un sistema de referencia bidimensional fijado en el plano de la pantalla con centro

en el extremo inferior izquierdo. Así, los valores “x” e “y” representan la abscisa y la ordenada de la posición media de los ojos en la pantalla en cada tiempo de medición. Los registros se realizan cada once milisegundos y con una frecuencia de sampleo de 90 Hertz mientras el sujeto está realizando el test. El software también registra el momento (dado por un click con el mouse o una pulsación de la barra espaciadora) en que el participante localiza cada uno de los números en la pantalla y el test se divide en etapas llamadas “trials”, de manera que cada trial incluya el tiempo comprendido entre la localización de un número del test y la localización del siguiente.

A partir de las mediciones que realiza el dispositivo de seguimiento ocular de la manera descrita, para realizar el estudio de esta tesis se cuenta entonces con 108 datasets de estudio, uno por cada persona, con una cantidad variable de registros para cada individuo dada por la cantidad de milisegundos que demoró la persona en la realización del test. Los datasets contienen un número de registros temporales que van entre 1.116 y 10.446 registros de forma que cada uno de ellos contiene el tiempo de la medición en milisegundos y el valor promedio de la posición de los ojos dado por la coordenada horizontal x de la mirada en la pantalla y la coordenada vertical y de la posición promedio al momento del muestreo, ambas medidas a partir del sistema de referencia fijado en la pantalla.

De esta manera, se cuenta con 108 bases de datos cada una de las cuales contiene los valores de tres variables numéricas cuyos nombres son: “t”, “x” e “y” que se corresponden con el tiempo del registro, la abscisa de la posición de la mirada promedio de ambos ojos en ese tiempo “t” y la ordenada de la posición de la mirada promedio de ambos ojos para ese tiempo, respectivamente. A modo ilustrativo, la Tabla 2.1 muestra el encabezado de uno de los datasets de estudio.

Medición	**Tiempo (en miliseg)**	**Absisa (x)**	**Ordenada (y)**
0	0	0.412408	0.563846
1	12	0.412282	0.564447
2	22	0.412155	0.565622
3	33	0.413163	0.567609
4	44	0.412743	0.568215

Tabla 2.1: Ejemplo de los primeros registros de la base de datos correspondiente a un individuo

Capítulo 3

Exploración inicial

En esta sección se realiza un estudio descriptivo preliminar de los datos mencionados en la sección anterior. El estudio incluye algunas representaciones visuales de las variables de estudio con el fin de construir un panorama global de lo registrado. Se trata de generar ideas para realizar análisis más específicos de los datos de seguimiento ocular con que se cuenta y obtener algunas medidas estadísticas de los datos de estudio.

Para tener una idea de los recorridos visuales sobre la pantalla de los sujetos objeto de este análisis durante la resolución del TMT-A y contar con la información bidimensional en forma particular de cada una de las trayectorias durante todo el estudio de esta tesis, se diseñó e implementó una visualización interactiva de consulta de los datos de posición de la mirada en cada registro y de cada individuo. La Figura 3.1 muestra un ejemplo de la información visual proporcionada por el panel de visualización de una trayectoria, que incluye la distribución en pantalla de la coordenada x de la mirada en función del tiempo, la distribución de la ordenada y en cada tiempo de medición t y la posición del par (x, y) en pantalla para cada tiempo de registro. La información temporal de las coordenadas está representada en la gráfica a través de la escala de color dependiente del tiempo de los puntos. Éstas gráficas pueden obtenerse para cada individuo y para cada trial a elección. En caso de que no se elija un trial, en el panel puede apreciarse además la distribución del tiempo insumido en la resolución de cada uno de los trials.

Como primer paso de la exploración analítica de los datos, se obtuvieron las medidas estadísticas resumen de las variables generadas por el dispositivo de seguimiento ocular en el registro de todos los sujetos durante el test, es decir, de los valores de las variables *tiempo*, *abscisa* y *ordenada* de todas las mediciones para todos los individuos de estudio. Estas medidas se muestran en la Tabla 3.1 y las distribuciones de las tres variables pueden verse en los gráficos de boxplots de la Figura 3.2.

Medidas	**Tiempo (en seg)**	**Absisa (x)**	**Ordenada (y)**
Media	20,625	0,515	0,480
Desvío estándar	15,997	0,278	0,261
Valor mínimo	0	−0,0005	0
Primer cuartil 25 %	8,828	0,258	0,262
Segundo cuartil 50 %	18,420	0,544	0,462
Tercer cuartil 75 %	28,376	0,758	0,707
Valor máximo	114,912	0,999	1,001

Tabla 3.1: Valores resumen de las variables de estudio

Cabe mencionar que ninguna de las variables de los distintos datasets presentan valores faltantes ni valores de las coordenadas que puedan atribuirse “a simple vista” a errores de medición.

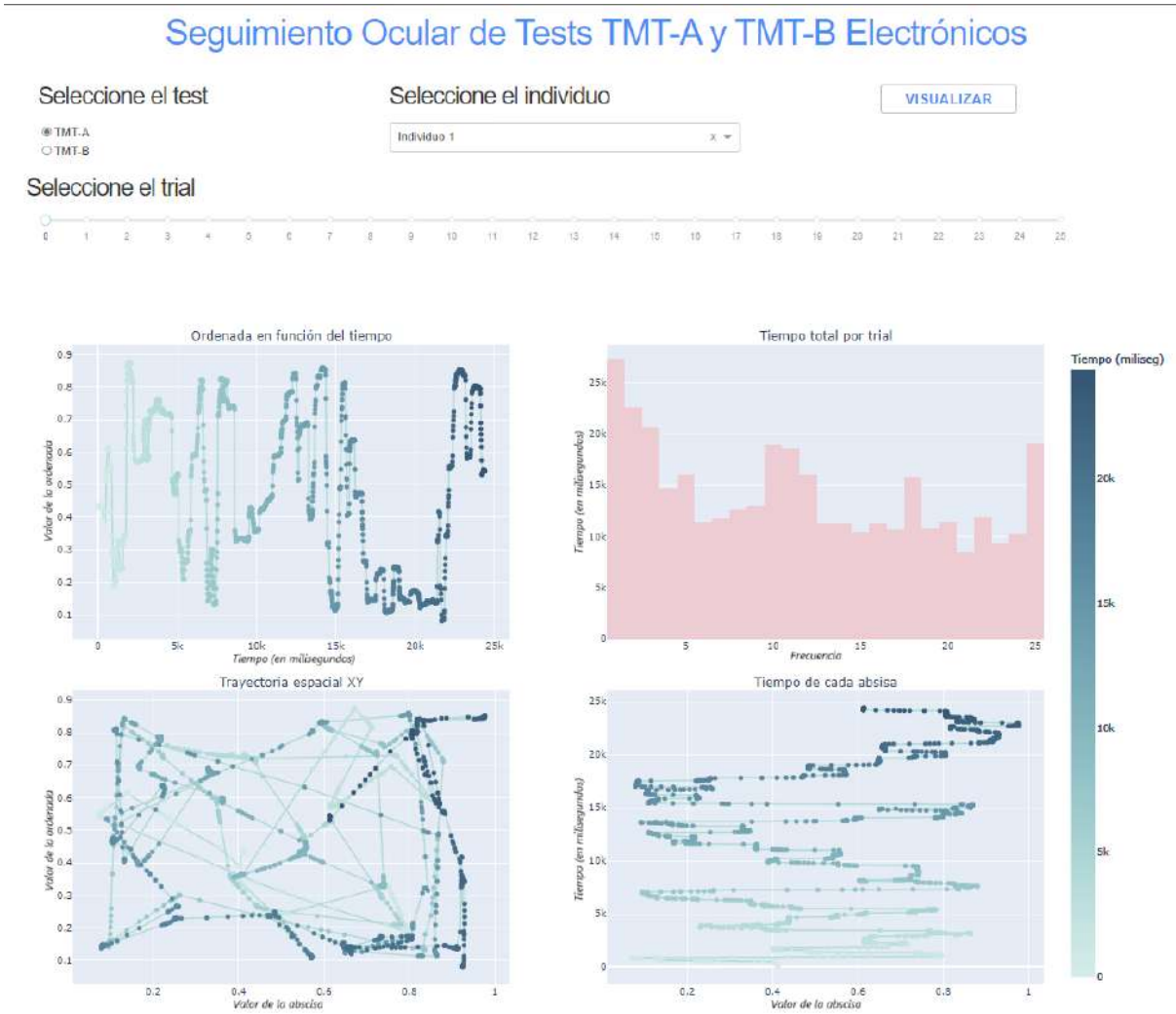


Figura 3.1: Ejemplo de una trayectoria en el panel diseñado para consulta visual de los datos

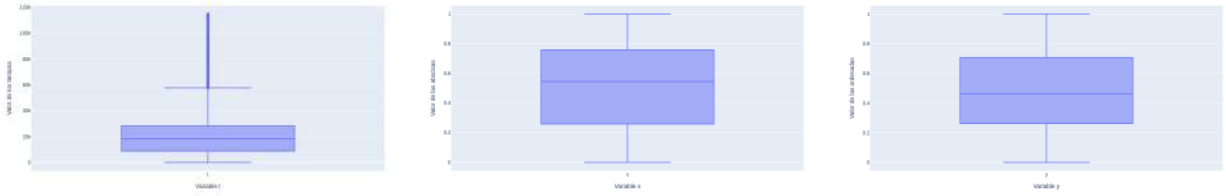


Figura 3.2: Gráficos de caja de las variables de estudio para todos los sujetos

3.1. La variable tiempo

Las medidas resumen de la Tabla 3.1 correspondientes al tiempo no proporcionan información muy relevante porque en ellas están incluídos los registros de todos los individuos, cuyos tiempos siempre comienzan en el valor 0 y avanzan desde ese valor hasta el tiempo total de resolución del test de cada sujeto. De todos modos, un dato que podemos extraer de estas medidas es el tiempo máximo de resolución del test por parte del conjunto de individuos, este máximo indica que todos los sujetos han resuelto el test en menos de 115 segundos, es decir, a grandes rasgos, en menos de dos minutos. Más allá de la observación de la Tabla 3.1, como sabemos que los 108 datasets tienen entre 1.116 y 10.446 registros y las mediciones se realizan cada 11 milisegundos, podemos afirmar que los tiempos de resolución del TMT-A por parte de los individuos bajo estudio van desde un mínimo de $12,27$ segundos

a un máximo de 114, 91 segundos exactamente, con una diferencia importante entre el número de sujetos que finalizan el test en el primer cuarto de ese tiempo de 115 segundos y el resto de los tiempos, es decir, que muchas más personas finalizan el test en 30 segundos, o menos, que las que lo hacen en más de ese tiempo; esto se evidencia en la notoria mayor cantidad de tiempos en el primer cuartil del rango total de valores temporales.

3.2. Variables de posición

En la Figura 3.3 puede observarse la distribución espacial en pantalla (y sus gráficos de distribución correspondientes) de todas las posiciones visuales de los 108 sujetos ocupadas durante la realización del test TMT-A.

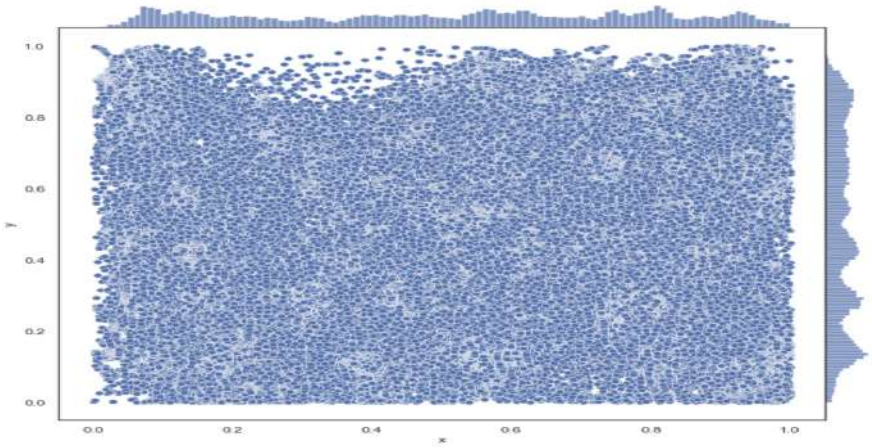


Figura 3.3: Distribución de las variables espaciales en pantalla

Las 108 posiciones medias de todo el recorrido en pantalla durante la resolución del test, por sujeto, están representadas sobre la imagen del TMT-A a continuación en la Figura 3.4, figura en la que puede verse también en rojo la mediana de las posiciones medias de todos los sujetos. El rectángulo coloreado en celeste que se observa en la Figura 3.4 corresponde a la región de la pantalla donde se ubican el 50 % central de todos los datos registrados.

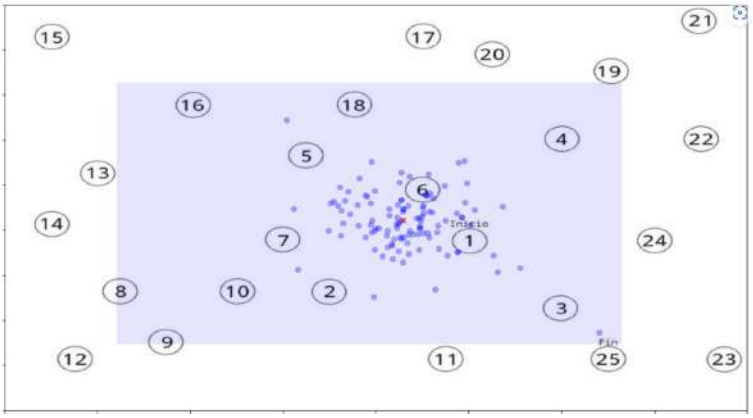


Figura 3.4: Posiciones medias de recorrido de cada individuo y la mediana de todas ellas

Realizando un análisis numérico respecto a las variables x e y, a la luz de las medidas estadísticas de la Tabla 3.1 que consideran la totalidad de los individuos, vemos que las posiciones de las miradas recorren ambas coordenadas x e y prácticamente de extremo a extremo de la pantalla, de 0 a 1 (mínimo de abscisas: 0,0005, mínimo de ordenadas: 0,000006, máximo de abscisas: 0,999 y máximo de ordenadas: 1,001), lo que indica que el conjunto de los individuos ha llegado a explorar toda la pantalla (al menos

uno de ellos ha explorado alguno de los extremos de la pantalla). Por otro lado, vemos que los valores medios de las coordenadas de posición x e y son cercanos al valor 0,5 en ambos casos, las medias son 0,515 y 0,480 respectivamente, lo que muestra que los individuos en su totalidad han recorrido la pantalla observando una cantidad de puntos similar en cada semiplano, tanto en dirección vertical como horizontal. Es decir, no se ve un sesgo de posición hacia alguno de los cuatro semiplanos de la pantalla en la resolución del test por parte del conjunto de sujetos. Más aún, los valores de las medianas de las variables espaciales que pueden observarse claramente en los gráficos de boxplots de la Figura 3.2 y que son de 0,54 y 0,46 en x e y respectivamente, dan indicios de que las miradas se concentran justamente en el centro de la pantalla, con un valor de desvío equivalente a un cuarto de pantalla, y que los sujetos recorren con una asiduidad similar ambos semiplanos verticales y ambos horizontales.

Al observar más detenidamente los graficos de las distribuciones de las variables espaciales x e y de la Figura 3.5, podríamos decir que ambas coordenadas presentan básicamente un rasgo de distribución uniforme aunque con irregularidades que son un poco más pronunciadas en la variable que representa la abscisa en pantalla que en la que representa la posición vertical.

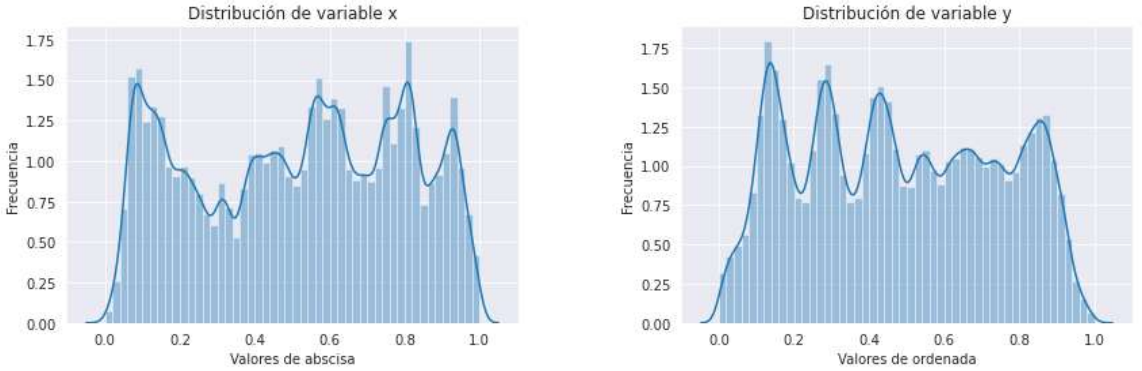


Figura 3.5: Gráficos de distribución de las variables espaciales del conjunto de individuos

Corroborando estas observaciones gráficas con medidas numéricas y, de alguna manera, estudiando la posible aunque poco probable normalidad de la distribución de las variables espaciales, puede mencionarse que se calcularon los valores de "skewness", sesgo o asimetría, y curtosis de las variables x e y y que se obtuvieron un valor de asimetría o skewness de $-0{,}11$ en el caso de la variable de la abscisa y 0,07 en la variable de la ordenada, lo que indica una leve pero mayor asimetría en el recorrido de la pantalla realizado por los sujetos en el eje vertical que en el eje horizontal. Por otro lado, en relación a los valores obtenidos de curtosis, vemos que son muy similares en ambas variables ($-1{,}212$ para x y $-1{,}198$ para y) y negativos, indicando la semejanza de las distribuciones de las dos variables espaciales con una distribución uniforme.

Coordenadas espaciales en función del tiempo

Dada la diferencia en tiempo de realización de la prueba TMT-A por parte de cada persona que mencionamos en la sección previa, la cantidad de registros de cada individuo es muy distinta ya que depende del tiempo que el sujeto haya demorado en la resolución del test. Por esta razón, es difícil pensar en realizar una comparación directa entre individuos en relación a los valores de las coordenadas espaciales x e y en función del tiempo. Además, por la misma razón, los valores resumen de las variables espaciales en su estado original no dan información clara sobre el conjunto de esas variables ni de su evolución temporal. De todos modos, y con el objeto de que a partir de ellos pueda obtenerse información de estos atributos para cada individuo o puedan realizarse otros estudios a futuro (considerar las coordenadas como series temporales, por ejemplo), se incluyen a modo ilustrativo en la Figura 3.6 los gráficos de las variables espaciales x e y en función del tiempo para un individuo en particular. Cabe mencionar también que en este tipo de gráfico puede observarse la magnitud de los cambios en los valores de las coordenadas para un cierto tiempo. El estudio de estos cambios, especialmente el análisis de la magnitud de los cambios en un intervalo de tiempo pequeño, pueden conducirnos a la detección

y posible análisis de las fijaciones y sacadas del individuo durante la resolución del test, elementos de crucial importancia en el estudio de datos de seguimiento ocular.

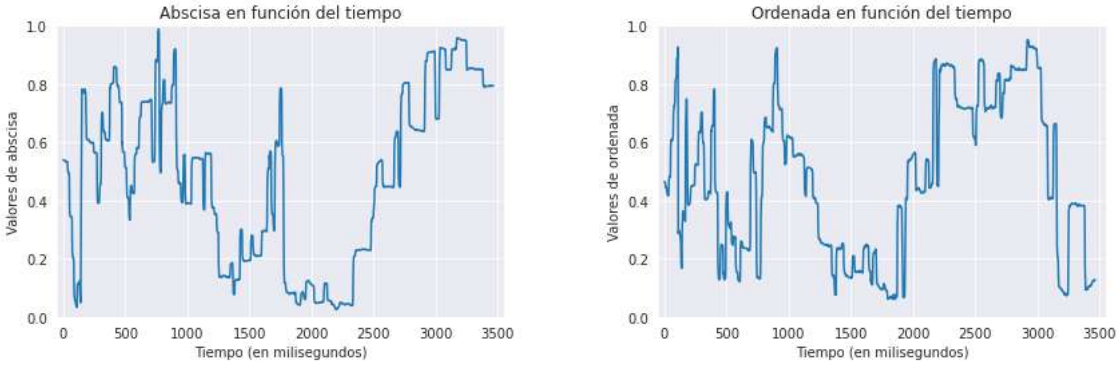


Figura 3.6: Evolución temporal (en ms) de las coordenadas espaciales de un sujeto

3.3. Recorrido en pantalla

La Figura 3.7 muestra la configuración espacial de la mirada de uno de los sujetos cuyo seguimiento ocular se ha registrado en nuestro experimento (sin considerar los tiempos en que alcanzó cada posición). Es decir, en la figura se puede apreciar la representación gráfica de todos los puntos (x, y) registrados como posiciones de la mirada del sujeto, uniendo los puntos de tiempos consecutivos, para todos los registros durante la resolución del test del individuo, sin tener en cuenta su evolución temporal exacta y representados sobre la imagen en pantalla del TMT-A.

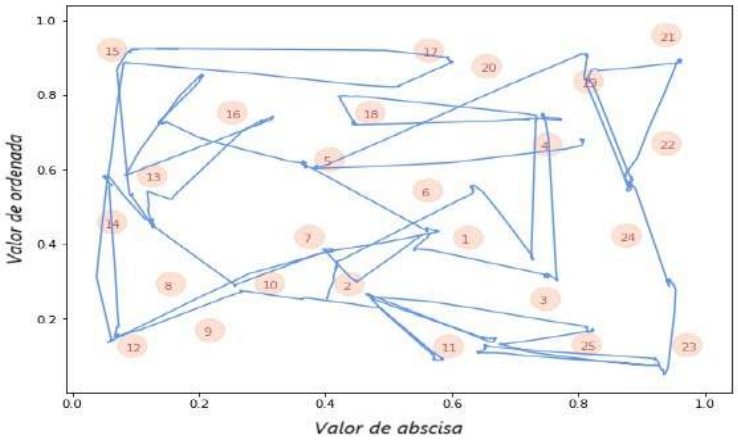


Figura 3.7: Ejemplo de trayectoria de resolución

Si bien esta representación es solo un ejemplo ilustrativo, como es de esperar, en ella puede apreciarse que los puntos del recorrido de la mirada del individuo están en directa relación con la posición de los números en la pantalla del test. Observando las representaciones de las trayectorias de varios individuos puede pensarse que, si bien las trayectorias son muy diferentes de un individuo a otro, quizás puedan encontrarse algunos patrones de recorrido de la pantalla realizando análisis más profundos. El estudio de gráficos de este tipo también puede conducir a la detección de posibles personas con trayectorias muy diferentes a los recorridos más frecuentes, hecho que puede indicar alguna deficiencia de atención o algún trastorno relacionado a los objetivos neuropsicológicos del test TMT-A. Por otro lado, el análisis de las trayectorias de resolución puede permitir la definición de posibles regiones de interés donde los participantes posan la mirada con mayor frecuencia. Las nociones de lugares de interés y posiciones de fijación de la mirada son de una importancia fundamental en el estudio del seguimiento ocular.

Capítulo 4

Variables derivadas

Teniendo en mente el objetivo de analizar la información espacial de las trayectorias visuales de los individuos sin perder de vista la información temporal, en esta sección se estudian posibles agrupamientos de las resoluciones del TMT-A en función de ciertas variables espacio-temporales obtenidas a partir de las mediciones (t, x, y) y que, en algún sentido, resumen o caracterizan el comportamiento visual durante la resolución del test por parte de los sujetos. En otras palabras, a partir de las variables t, x, y surgidas del dispositivo de seguimiento ocular y considerando a las variables espaciales x e y como funciones del tiempo t, es decir, considerando la secuencia de puntos $(x(t), y(t))$ de la trayectoria visual de cada persona, se obtuvieron variables derivadas de los registros que sintetizan o describen la información global de la trayectoria de cada individuo de la manera que se detalla en los siguientes apartados.

4.1. Tiempos de resolución.

En esta subsección se realizará un análisis exploratorio de la variable tiempo total del resolución de test de cada persona, siendo ésta una de las variables más relevantes para los psicólogos y neurólogos al estudiar el resultado del TMT-A (ver [27]).

La variable de estudio en que nos concentraremos ahora es la variable que llamaremos **Tiempo total** de realización de la prueba por parte de cada individuo, es decir, el máximo valor de la variable t de cada sujeto. Las medidas estadísticas de los valores de esta variable en el conjunto de los participantes de nuestro experimento se detallan en la Tabla 4.1 y puede verse su distribución en el boxplot de la Figura 4.1.

Medidas	Tiempo total (en miliseg)
Media	37950,68
Desvío estándar	13300,36
Valor mínimo	17776,00
Primer cuartil 25 %	30107,25
Segundo cuartil 50 %	35554,00
Tercer cuartil 75 %	41877,25
Valor máximo	114912,00

Tabla 4.1: Valores estadísticos de los tiempos totales de resolución

A partir de los valores estadísticos que muestra la Tabla 4.1, ratificamos el análisis de tiempos de la sección anterior y observamos nuevamente que quien más demoró en la resolución del test fue un individuo que tardó casi 115 segundos, casi 2 minutos. Además de este dato, podemos ver que los individuos bajo estudio han resuelto el test en un tiempo promedio de casi 38 segundos y que el desvío estándar de los tiempos es de poco más de 13 segundos, con un coeficiente de variación

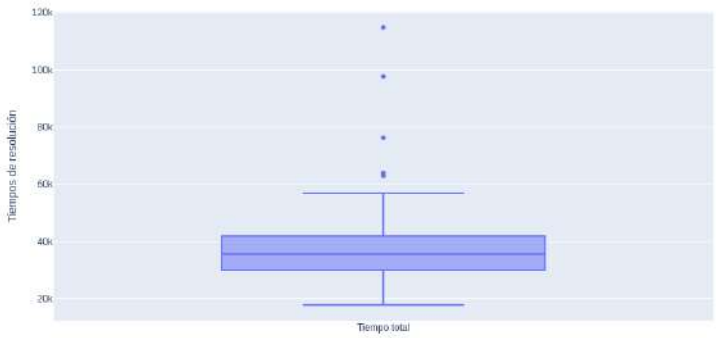


Figura 4.1: Boxplot de tiempos de resolución

$\left(CV = \frac{desvio}{media}.100\right)$ de 35 respecto al tiempo medio. En otras palabras, se observa una alta variabilidad en el tiempo total de resolución de la prueba por parte de las distintas personas. Al realizar el cálculo de la mediana del tiempo total para todos los individuos, se obtiene que ésta es de $35,55$ segundos, lo que indica una distribución bastante balanceada de los valores de los tiempos a ambos lados de la media. Esta situación puede observarse claramente en el gráfico de caja de los tiempos totales de resolución de los 108 individuos que se muestra en la Figura 4.1.

La Figura 4.2 exhibe la distribución de los tiempos de resolución de los sujetos y en ese gráfico puede verse que la amplia mayoría de las personas han resuelto el TMT-A en un tiempo que va de 25 a 45 segundos. Otro dato relevante que puede extraerse de esa representación gráfica es que solo 3 individuos de los 108 han demorado más de 65 segundos en resolver el test.

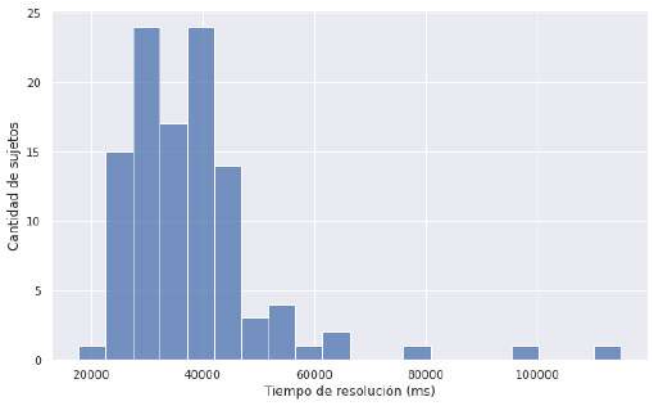


Figura 4.2: Distribución de tiempos de resolución

Si analizamos numéricamente los extremos de esta distribución obteniendo los primeros diez tiempos más cortos y los diez tiempos más largos de resolución del conjunto de sujetos, datos que se muestran en las Tablas 4.2 y 4.3, vemos que el menor tiempo de 17 segundos es un tiempo notoriamente pequeño, dado que los siguientes cuatro mejores tiempos son de 24 segundos y los siguientes seis registros son de 25 segundos de resolución total. Con respecto a los mayores tiempos registrados, el máximo de 115 segundos también es muy diferente de los otros tiempos puesto que el segundo mayor tiempo es de 97 segundos, que también es bastante mayor que el registro que le sigue que es de 76 segundos. Luego, el ranking de mayores tiempos lo siguen valores de 63, 62 y 56 segundos, y a partir de ese valor se tienen tiempos con un segundo a lo más de diferencia con el siguiente.

4.2. Longitud, área y velocidad de resolución

En esta sección se estudiarán otras variables que sintetizan cierta información global de la trayectoria recorrida y la manera en que se recorrió, estas variables son la longitud total recorrida por la mirada,

Registro	Tiempo (ms)
16	17776
14	24372
57	24595
6	24839
80	24960
84	25118
28	25294
94	25660
24	25705
43	25793

Tabla 4.2: Registros más bajos de tiempos de resolución

Registro	Tiempo (ms)
0	111912
1	97719
107	76313
86	63971
44	62989
105	56948
45	54909
104	53550
29	52828
87	51952

Tabla 4.3: Registros de mayores tiempos de resolución

el área total de pantalla explorada y la velocidad de resolución del test. Más concretamente, a partir de los valores de las variables espaciales x e y en cada tiempo de medición t_i definimos, para cada sujeto, las siguientes variables:

- **Longitud total** de la trayectoria, definida por $\sum_{i=0}^{T} \sqrt{(x(t_{i+1}) - x_t)^2 + (y(t_{i+1}) - y(t_i))^2}$, siendo T el tiempo total de resolución del test.
- **Área total** de pantalla explorada, definida como $[max(x(t)) - min(x(t))].[max(y(t)) - min(y(t))]$.
- **Velocidad media** de la trayectoria, dada por $\sum_{i=0}^{T} \frac{1}{T} \sqrt{(x(t_{i+1}) - x_t)^2 + (y(t_{i+1}) - y(t_i))^2}$, siendo T el tiempo de resolución del test.

A continuación, se muestran las medidas estadísticas de éstas variables derivadas y se realiza una breve descripción de ellas.

La variable correspondiente a la longitud recorrida en el trayecto visual de resolución del test por parte de los sujetos tiene las medidas estadísticas que exhibe la Tabla 4.4 y un panorama más global del comportamiento de esta variable longitud recorrida en la resolución del TMT-A en el conjunto de sujetos puede verse en el gráfico de boxplot de la Figura 4.3.

A la vista de estos resultados, puede mencionarse que las longitudes de resolución varían entre un mínimo de 24,71 píxeles normalizados (PN) y un máximo de 141,39 PN, lo que indica un rango importante de longitudes recorridas en pantalla por parte de los distintos individuos. Dentro de este rango, la longitud media de recorrido en el conjunto de participantes es de 59,94 PN, ligeramente diferente del valor de la mediana establecido en 55,17 PN. También se observa una variabilidad media

Medidas	Longitud total recorrida
Media	59,943
Desvío estándar	22,735
Valor mínimo	24,710
Primer cuartil 25 %	43,362
Segundo cuartil 50 %	55,176
Tercer cuartil 75 %	71,392
Valor máximo	141,395

Tabla 4.4: Valores resumen de las longitudes recorridas

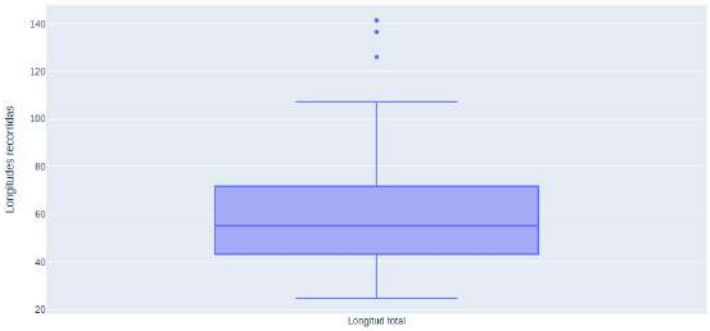


Figura 4.3: Valores resumen de las longitudes totales recorridas

de los valores de la longitud recorrida, dado que el valor del desvío estándar tiene un coeficiente de variación de casi 38 de los valores de las longitudes.

En relación a la distribución de los valores de las longitudes recorridas en la pantalla, que puede observarse tanto el gráfico de boxplot de la Figura 4.3 como el histograma de la Figura 4.4, muestra una escasa diferencia entre la media y la mediana de las longitudes, aunque con una mayor frecuencia de valores en las longitudes de la primer mitad de la campana de Gauss, con la presencia de tres longitudes totales recorridas con valores atípicamente muy altos.

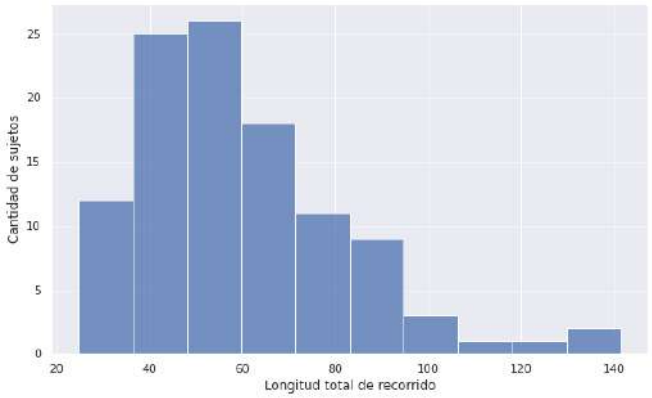


Figura 4.4: Distribución de longitudes recorridas en la resolución

Realizando un análisis de la misma naturaleza, observando la Tabla 4.5 de medidas estadísticas de las velocidades medias de resolución del test y las Figuras 4.5 y 4.6 que exhiben el boxplot y la distribución de las velocidades, lo que puede comentarse sobre la media de velocidad de recorrido de pantalla en el conjunto de individuos bajo estudio es muy similar a lo observado en la variable longitud.

La media y la mediana de los valores de velocidad de recorrido son prácticamente iguales (0,00158 y 0,00156, respectivamente), lo que indica un comportamiento similar a una distribución normal con una diferencia entre el mínimo y el máximo de 0,00189 píxeles normalizados sobre milisegundo.

Medidas	Velocidad del recorrido
Media	$1{,}585x10^{-3}$
Desvío estándar	$3{,}6x10^{-4}$
Valor mínimo	$8{,}79x10^{-4}$
Primer cuartil 25 %	$1{,}331x10^{-3}$
Segundo cuartil 50 %	$1{,}566x10^{-3}$
Tercer cuartil 75 %	$1{,}797x10^{-3}$
Valor máximo	$2{,}769x10^{-3}$

Tabla 4.5: Valores resumen de las velocidades del recorrido en pantalla

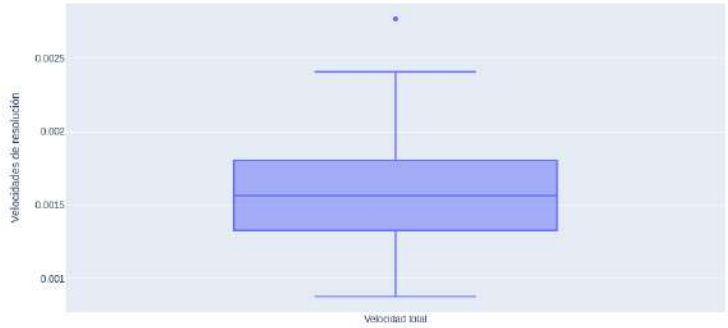


Figura 4.5: Gráfico de caja de las velocidades de resolución del test

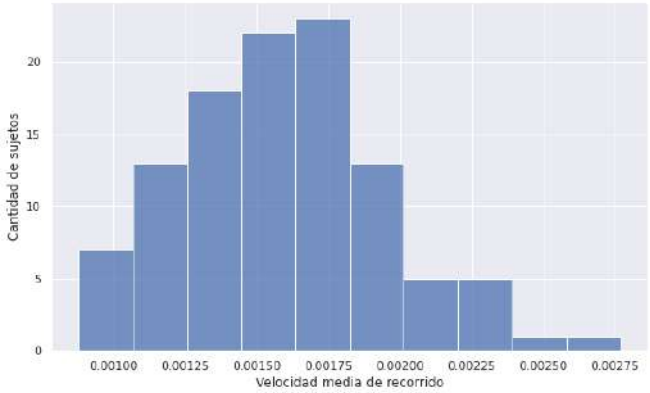


Figura 4.6: Distribución de velocidades de resolución

En referencia al análisis de las áreas totales escaneadas visualmente durante la resolución del test, a partir de los valores estadísticos que presenta esta variable y que pueden consultarse en la Tabla 4.6, vemos que los sujetos recorren en promedio un 85 % del área total de la pantalla con un mínimo de recorrido en la resolución por parte de un individuo del 50 %, siendo el máximo de área escaneada prácticamente el total de la pantalla.

En cuanto a la distribución de los valores de las áreas recorridas que pueden apreciarse en el gráfico de boxplot de la Figura 4.7 y especialmente en el histograma de la Figura 4.8, podemos ver que prácticamente la totalidad de los individuos han recorrido un área de pantalla mayor al 70 % de la misma y que la amplia mayoría de los sujetos han escaneado más del 80 % del área de la pantalla al resolver el test, en general, entre el 80 y el 90 % del área total.

Medidas	Área total recorrida
Media	0,854369
Desvío estándar	0,072812
Valor mínimo	0,507214
Primer cuartil 25 %	0,823623
Segundo cuartil 50 %	0,853359
Tercer cuartil 75 %	0,891754
Valor máximo	0,995286

Tabla 4.6: Valores resumen de las áreas recorridas en pantalla

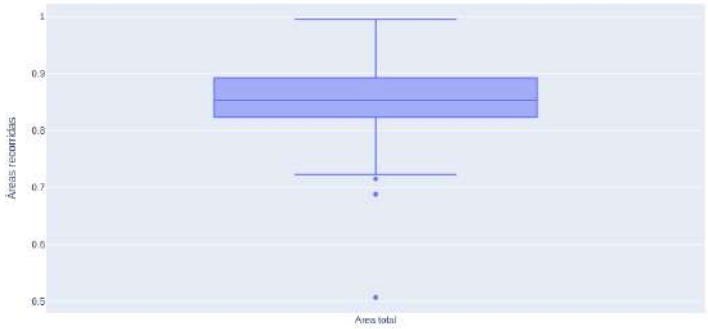


Figura 4.7: Gráfico de caja de las áreas recorridas en la resolución

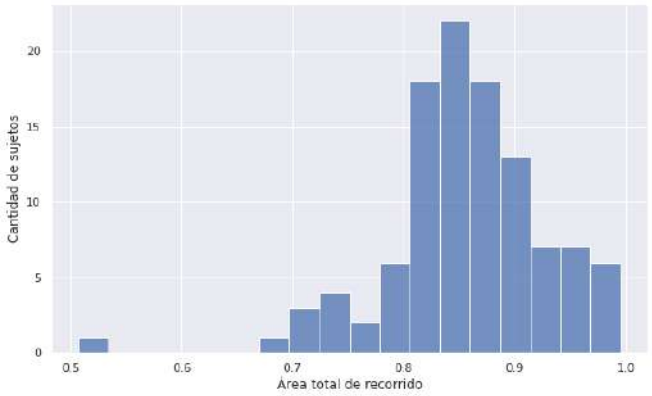


Figura 4.8: Distribución de áreas recorridas en la resolución

4.3. Otras variables derivadas

A las variables generales analizadas especialmente en las dos secciones previas, y con miras al estudio de agrupamiento de las resoluciones, agregaremos ahora la consideración de algunas otras variables de carácter estadístico que dan información de la trayectoria de resolución llevada a cabo por los sujetos, esas variables son las siguientes.

- **Media horizontal**: el valor promedio de la variable x de cada individuo.
- **Media vertical**: media de la variable y por individuo.
- **Mediana horizontal**: la mediana de la variable x de cada sujeto.
- **Mediana vertical**: la mediana de la variable y del sujeto.
- **Media de velocidades medias puntuales en x**: la media de los valores $\dfrac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i}$ del sujeto, siendo $0 \leq i \leq T - 1$ si T es el tiempo total de resolución.

- **Media de velocidades medias puntuales en y**: la media de los valores $\frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}$ por sujeto, siendo $0 \leq i \leq T-1$ si se nota T al tiempo total de resolución.
- **Mediana de velocidades medias puntuales en x**: la mediana de los valores $\frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i}$, siendo $0 \leq i \leq T-1$ si T es el tiempo total de resolución del individuo.
- **Mediana de las velocidades medias puntuales en y**: la mediana de los valores $\frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}$, si $0 \leq i \leq T-1$ con T el tiempo total de resolución.
- **Media de velocidades medias puntuales bidimensionales**, es decir, la media de los valores $\frac{\sqrt{(x(t_{i+1}) - x_t)^2 + (y(t_{i+1}) - y(t_i))^2}}{t_{i+1} - t_i}$ por sujeto, siendo $0 \leq i \leq T-1$ con T el tiempo total de resolución.
- **Mediana de velocidades medias puntuales bidimensionales**, es decir, la mediana de los valores $\frac{\sqrt{(x(t_{i+1}) - x_t)^2 + (y(t_{i+1}) - y(t_i))^2}}{t_{i+1} - t_i}$, siendo $0 \leq i \leq T-1$ con T el tiempo total de resolución.

Solo a título descriptivo y para formar una idea de los valores de estas variables que serán con las que se realizará el análisis de clustering, en la Tabla 4.9 se muestran las medidas estadísticas de éstas variables derivadas en el conjunto de los sujetos de estudio. En ella se presentan distintas medidas estadísticas descriptivas correspondientes a las variables analizadas. La columna "Cantidad" indica la cantidad de sujetos o registros considerados en el análisis; La columna "media" representa la media aritmética de los valores observados, mientras que "Desvío" corresponde al valor del desvío estándar, que cuantifica la dispersión de los datos en torno a la media. El valor "Mínimo" señala el mínimo registrado entre los valores, y los porcentajes cuartiles (25 %, 50 % y 75 %) indican los valores por debajo de los cuales se encuentran, respectivamente, el 25 %, el 50 % (mediana) y el 75 % de los datos. Finalmente, "Máximo" corresponde al valor máximo observado en la distribución de datos.

	Tiempo total	Media x	Mediana x	Media y	Mediana y	Media desplazamiento x
Cantidad	108.000000	108.000000	108.000000	108.000000	108.000000	108.000000
Media	37950.685185	0.517542	0.538292	0.476859	0.467572	0.000174
Desvío	13300.366254	0.041663	0.063356	0.054933	0.073073	0.000167
Mínimo	17776.000000	0.428527	0.379903	0.259579	0.192472	-0.000162
25%	30107.250000	0.487358	0.498190	0.444571	0.430253	0.000067
50%	35554.000000	0.516754	0.536019	0.478802	0.466837	0.000141
75%	41877.250000	0.540344	0.570168	0.513163	0.508146	0.000233
Máximo	114912.000000	0.715656	0.800539	0.632784	0.715523	0.000632

	Mediana desplazamiento x	Mediana desplazamiento y	Media desplazamiento xy	Mediana desplazamiento xy	Longitud total	Velocidad total
Cantidad	108.000000	108.000000	108.000000	108.000000	108.000000	108.000000
Media	0.000005	-0.000026	0.010661	0.002167	59.943703	0.001585
Desvío	0.000041	0.000090	0.003387	0.001387	22.735421	0.000360
Mínimo	-0.000086	-0.000245	0.005053	0.000736	24.710930	0.000879
25%	-0.000019	-0.000079	0.008311	0.001374	43.362238	0.001331
50%	-0.000005	-0.000025	0.009980	0.001664	55.176155	0.001566
75%	0.000025	0.000010	0.012880	0.002391	71.392615	0.001797
Máximo	0.000169	0.000468	0.023081	0.008717	141.395827	0.002769

Tabla 4.9: Medidas estadísticas de los valores de las variables derivadas

Observando el resumen estadístico de las variables, podemos mencionar que las medias y medianas horizontales, es decir, medias y medianas de las abscisas, son muy cercanas entre sí y centrales en el eje horizontal, con valores de 0,51 y 0,53, respectivamente, y con desvíos muy pequeños de aproximadamente 0,05. En el caso de las ordenadas sobre el eje vertical, también la media y mediana son muy cercanas pero ambas por debajo del valor correspondiente a la mitad exacta de la pantalla 0,5. Estos valores de medias y medianas de las variables espaciales por individuo son, previsiblemente, muy cercanos a los valores de medias y medianas considerando las mediciones de todos los sujetos bajo estudio como un conjunto de datos único.

Respecto a los valores de las velocidades unidimensionales y bidimensionales, éstos valores son muy pequeños como para hacer algún análisis comparativo a simple vista y las áreas recorridas por las trayectorias son del 85 % de la pantalla en promedio.

4.4. Normalización

Dada la muy diferente escala de valores en la que se mueven las distintas variables derivadas que se describieron en las secciones previas del presente capítulo (por ejemplo, la variable "Tiempo total" tiene un rango de 115.000 milisegundos y la mediana de las velocidades medias puntuales en x tiene un rango de 0,00024) y dado que serán los atributos respecto a los que se realizará el estudio de agrupamiento, realizaremos una normalización de las variables para balancear la relevancia de todas las variables en la formación de grupos, evitando que las tan distintas magnitudes alteren el "peso" de cada una de ellas en el proceso de agrupamiento.

En esta dirección, normalizamos los 14 atributos con el método conocido como "Min-Max". Este método somete a los datos de una variable X a una transformación lineal que consiste en un desplazamiento y un re-escalado que mapea los valores de la variable a los de una nueva variable $\overline{X}$ asociada que tiene rango en el intervalo $[0, 1]$, de acuerdo a la fórmula dada por la *ecuación de normalización*:

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{min}},$$

donde X_{max} y X_{min} son los valores máximo y mínimo de la variable X, respectivamente.

En el momento de plantearnos qué normalización era la más adecuada para el problema, analizamos el hecho de que el dataset de estudio no tiene valores faltantes (para cada tiempo de muestreo se registró un valor de abscisa x y uno de ordenada y y las variables derivadas se han definido directamente a partir de esos valores sin datos ausentes), se cuenta con los valores máximos y mínimos de todas las variables, y no se espera tener uan cantidad relevante de valores atípicos dado que se midió una población neurotípica sin casos que puedan resultar demasiado disonantes a priori. En este contexto, preferimos realizar una normalización min-max respecto al rango de valores de nuestros datos en lugar de utilizar el método Z-score. Las principales razones que fundamentan la decisión son:

- Luego de la normalización, se tienen los datos dentro de un rango específico $[0, 1]$ preservando la estructura relativa de los datos que es crucial a los efectos de realizar un proceso de agrupamiento.

- Se preserva la escala original de medición los registros en el intervalo $[0, 1]$, lo que puede ser ventajoso para la interpretación de los resultados del agrupamiento.

- Si bien se utilizarán varios algoritmos de clustering con distintas métricas, por la naturaleza espacial del problema la distancia euclídea será la más utilizada y los algoritmos que trabajan con esta métrica pueden funcionar mejor con datos normalizados en un rango específico $[0, 1]$.

- La idea del método Z-score es analizar los valores de las variables en relación a una media "poblacional" ordinaria y los datos con que contamos para realizar el estudio de agrupamiento distan mucho de ser un muestreo significativo de toda una población, cualquiera que sea en la población que se piense, por lo que consideramos más apropiado normalizar los valores en relación a la propia muestra de datos con que contamos.

- Resignamos una posible pérdida de información de desviaciones estándar y efecto de valores atípicos, considerando que nuestro objetivo es agrupar los datos para obtener información sobre posibles estrategias de resolución del test y no estamos interesados específicamente en los valores de los datos de la muestra.

4.5. Tendencia al agrupamiento

Antes de realizar un análisis de clustering, resulta conveniente analizar la posibilidad de que exista alguna agrupación en el conjunto de datos. A este proceso se lo conoce como "Evaluación de la tendencia al agrupamiento" y puede llevarse a cabo mediante test estadísticos o en forma visual.

Justamente como nuestro objetivo es aplicar métodos de agrupamiento a nuestros datos de seguimiento ocular, a modo de primera exploración de la tendencia al clustering del conjunto de datos de estudio, evaluamos indicios de la existencia de alguna agrupación con la prueba estadística conocida como "Test de Hopkins". El estadístico Hopkins permite evaluar la propensión hacia el clustering de los datos calculando la probabilidad de que los datos provengan de una distribución uniforme, en otras palabras, analiza la distribución aleatoria en el espacio de las observaciones. Valores del estadístico en torno a 0,5 indican que los datos estudiados se distribuyen uniformemente y que por lo tanto no tiene sentido aplicar clustering. Cuanto más se aproxime a 0 el estadístico de Hopkins, más evidencias se tienen a favor de que existen agrupaciones en los datos y de que, si se aplica clustering correctamente, los grupos resultantes serán "reales".

Aplicando el test de Hopkins a nuestros datos, obtenemos un índice de 0,261 que, si bien no es cercano al valor $0,5$ que indicaría un agrupamiento nada significativo, tampoco es muy cercano a 0, por lo que el test muestra que puede haber agrupamientos "reales" de los datos en función de las variables derivadas, pero los agrupamientos no aparecen demasiado claros.

Con la intención de obtener un posible mejor índice de Hopkins y, por ende, un mejor agrupamiento de los datos, realizaremos un proceso de reducción de dimensión de nuestro dataset constituído por las 14 variables definidas a partir de los registros del dispositivo de seguimiento ocular que hemos descrito con anterioridad.

4.6. Reducción de la dimensión

En muchos problemas del mundo real, la reducción de la dimensión es un paso esencial antes de que cualquier análisis de datos pueda realizarse. El criterio general para reducir la dimensión es el deseo de preservar la mayor cantidad de información relevante de los datos originales de acuerdo a algún criterio de optimalidad. El objetivo de la "reducción de la dimensionalidad" es bajar la dimensión del espacio de los datos filtrando las características menos relevantes y manteniendo las que dan la información más determinante.

Las técnicas de reducción de dimensionalidad tienen muchos usos en el aprendizaje automático, ya que la capacidad de extraer la información útil de un conjunto de datos puede proporcionar mejoras de rendimiento en muchos problemas. Éstas técnicas son particularmente útiles en el aprendizaje no supervisado, dado que el conjunto de datos no contiene ninguna clase de etiquetas u objetivos a alcanzar, por lo que el propósito es organizar los datos de una manera apropiada para el problema que se está resolviendo y permiten identificar patrones de manera más efectiva, evitando cercanías de todos los datos entre sí, y resolver más eficientemente problemas computacionalmente costosos. Un gran número de variables no resultará necesariamente útil para un problema de agrupamiento dado que, por un lado la inclusión de variables irrelevantes no puede ser contrastada por el análisis de cluster y, por otro lado, aumenta la posibilidad de errores y puede generar imprecisiones en la conclusión final. Mediante una reducción inteligente de dimensiones, podemos hacer que la agrupación en clústers sea más fácil de interpretar y de comunicar, y que los modelos de aprendizaje sean lo más eficientes posible.

Las técnicas de reducción de dimensión pueden clasificarse en dos diferentes formas según la manera en que se realizan:

- Encontrando un conjunto más pequeño de nuevas variables determinadas como combinaciones de las variables originales, tarea que consiste en hallar una aplicación del espacio original de variables a otro espacio de variables de menor dimensión, resultado de un proceso de proyección, de modo que las nuevas variables contengan básicamente la misma información que las originales siendo un número menor de variables.

- Manteniendo solo los atributos más relevantes del conjunto de datos original, es decir, realizando una selección de un subconjunto de las variables originales para el proceso de aprendizaje automático, removiendo efectivamente algunas de las variables menos relevantes bajo consideración. Esta técnica, conocida como "Selección de variables", ayuda a evitar el sobreajuste al ruido del dataset.

En la sección subsiguiente se someterá al dataset de estudio de este trabajo de tesis a esta técnica de selección de variables para reducir la dimensión del problema de agrupamiento bajo análisis.

4.6.1. Selección de variables

En algunas aplicaciones puede resultar beneficioso hallar un subconjunto de las variables originales en lugar de una aplicación que use todas las variables del problema original, por ejemplo, implicando un menor costo computacional y de sensores en sistemas con mediciones físicas, y un menor número de variables de ruido. Así, los procedimientos de selección de variables han sido muy utilizados para reducir la dimensión en diferentes escenarios. Un estudio completo de las técnicas de selección de variables y un marco general para su categorización puede encontrarse en [24].

En el caso de estudio de esta tesis, y con el objetivo de reducir la dimensión, se decidió trabajar con técnicas de selección de variables y no aplicar métodos de reducción por proyecciones a espacios de dimensión menor ya que provocarían una importante modificación de las variables espacio-temporales de seguimiento ocular que se tienen, ocasionando una sustancial pérdida de capacidad de interpretación de los agrupamientos que puedan obtenerse en función de las variables de interés que se han definido especialmente en relación al problema.

Así, en esta sección se desarrollará un análisis de las trayectorias de seguimiento ocular realizando una reducción de la dimensión del dataset de estudio mediante la aplicación de los métodos de selección de variables conocidos como "Principal Feature Analysis (PFA)", "LASSO", "Spectral feature selection (SPEC)" y "Logistic Regression". Para obtener más información sobre los métodos y sus detalles específicos, puede consultarse el apéndice A.

A continuación se resumen los resultados obtenidos de la aplicación de los mencionados métodos de selección de variables, con vistas al tratamiento del problema de agrupamiento, al dataset de 14 variables derivadas de los registros del dispositivo de seguimiento ocular que es objeto de estudio de este trabajo (ver encabezado del dataset en la Figura 4.9).

- *Lasso*: En orden de importancia, las variables de relevancia destacada a los efectos de agrupar los datos de estudio son: tiempo total, longitud total, velocidad total, mediana y y media y. Le siguen con una importancia muy menor, las variables área total, mediana x, media velocidades medias puntuales bidimensionales xy, media de velocidades medias puntuales en x, mediana de velocidades medias puntuales en x, mediana de velocidades medias puntuales en y, media de velocidades medias puntuales en y, mediana de velocidades medias puntuales bidimensionales y media x. La Figura 4.10 y, más precisamente, la Tabla 4.7 muestran los valores de los coeficientes del método Lasso para cada una de las variables derivadas.

- *Principal feature analysis*: Solicitando al método las diez primeras variables en orden de importancia para el agrupamiento, se obtuvieron como resultado las variables: tiempo total, mediana de velocidades medias puntuales en x, mediana x, velocidad total, mediana y, media de velocidades medias puntuales en x, media de velocidades medias puntuales en y, mediana de velocidades medias puntuales en y, media de velocidades medias puntuales bidimensionales y área total. Al restringir la respuesta del método a ocho variables, los atributos resultantes fueron: tiempo total, media x, media y, media de velocidades medias puntuales en y, área total, mediana de velocidades medias puntuales en y, media de velocidades medias puntuales bidimensionales y velocidad total.

- *Spectral feature selection*: La aplicación de este método arrojó como las diez variables más relevantes en la agrupación de los datos a las variables: velocidad total, tiempo total, mediana y, mediana x, mediana de velocidades medias puntuales en y, mediana de velocidades medias puntuales bidimensionales, mediana de velocidades medias puntuales en x, media y, media x, media de velocidades medias

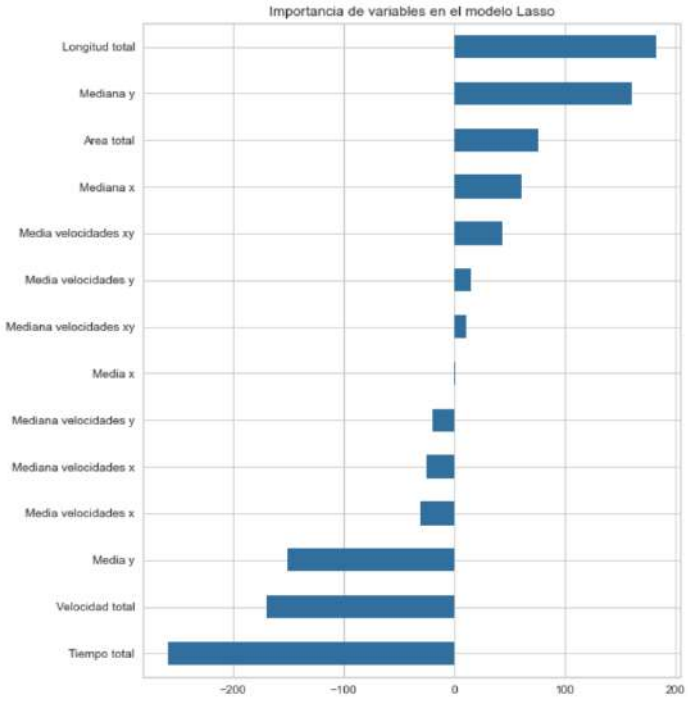


Figura 4.10: Importancia de variables con LASSO

Variable	Coeficiente
Longitud total	182,35
Mediana y	160,86
Area total	75,77
Mediana x	60,68
Media velocidades xy	43,18
Media velocidades y	14,80
Mediana velocidades xy	10,76
Media x	1,18
Mediana velocidades y	−19,63
Mediana velocidades x	−25,28
Media velocidades x	−30,29
Media y	−151,16
Velocidad total	−169,73
Tiempo total	−259,33

Tabla 4.7: Tabla de coeficientes de importancia de las variables según el método LASSO

puntuales en y, media de velocidades medias puntuales bidimensionales, media de velocidades medias puntuales en x, longitud total y área total.

- *Logistic regression*: En este caso, los resultados obtenidos para las diez variables más importantes fueron: mediana x, media y, media de velocidades medias puntuales en x, media de velocidades medias puntuales en y, mediana de velocidades medias puntuales en x, media de velocidades medias puntuales en xy, mediana de velocidades medias puntuales en xy, longitud total, velocidad total y área total. Para las ocho variables más relevantes, el resultado fue: mediana x, media de velocidades medias puntuales en x, media de velocidades medias puntuales en y, mediana de velocidades medias puntuales en x, media de velocidades medias puntuales bidimensionales, mediana de velocidades medias puntuales bidimensionales, longitud total y velocidad total.

A la luz de los resultados de los distintos métodos para las primeras variables más relevantes a los efectos del agrupamiento y para combinar las sugerencias de las distintas técnicas, se seleccionaron las ocho variables con mayor frecuencia de aparición en los cuatro métodos aplicados como conjunto de variables seleccionadas para conformar el dataset de estudio en dimensión reducida. En otras palabras, de aquí en lo que sigue procederemos con el análisis de grupos de las trayectorias de seguimiento ocular solo con la información de las siguientes ocho variables derivadas de los registros del dispositivo: "Velocidad total", "Mediana x", "Media de velocidades medias puntuales en y", "Media de velocidades medias puntuales bidimensionales, "Tiempo total", "Mediana de velocidades medias puntuales en x", "Área total" y "Media y".

La Figura 4.11 muestra el encabezado del dataset de estudio luego de realizada la selección de variables, es decir, luego de reducida su dimensión original.

	Tiempo total	Mediana x	Media y	Media velocidades y	Mediana velocidades x	Media velocidades xy	Velocidad total	Area total
0	1.000000	0.389508	0.587452	0.309343	0.390075	0.177671	0.163914	0.980242
1	0.823001	0.135119	0.498909	0.373818	0.362649	0.075262	0.110271	0.955983
2	0.266564	0.284373	0.568081	0.622241	0.378044	0.349092	0.423589	0.781312
3	0.270353	0.367303	0.473261	0.515940	0.260533	0.241340	0.223643	0.786670
4	0.305994	0.281014	0.440298	0.460387	0.250144	0.483394	0.499302	0.735768

Figura 4.11: Encabezado del dataset final de variables seleccionadas

A los efectos de realizar una verificación de que la reducción de dimensión por selección de variables no empeoró la tendencia al agrupamiento del conjunto de datos con las catorce variables originales, sometimos al dataset reducido al test de Hopkins. El dataset de ocho variables seleccionadas tiene una tendencia al agrupamiento similar al dataset original de las catorce variables, dado que el índice de Hopkins del conjunto de datos con las ocho variables es de $0,278$. De esta manera y de cara a ocuparnos de la formación de grupos en el conjunto de datos, no habiendo perdido posibilidad de agrupamiento, estamos ahora en presencia de un dataset a tratar con una dimensión considerablemente menor que el de variables derivadas original.

4.6.2. Correlación

Un aspecto fuertemente relacionado a la eleccion de variables para reducir la dimensión de un conjunto de datos es el de la eliminación de atributos correlacionados. Así, para complementar el análisis de selección de variables llevado a cabo en la sección anterior y corroborar su resultado, realizaremos en esta subsección un estudio de correlación cruzada en el conjunto de catorce variables derivadas originales, para asegurarnos de que no se hayan incluído en el conjunto de variables seleccionadas pares de atributos fuertemente correlacionados.

El análisis de correlación se llevó a cabo en esta tesis mediante el cálculo del coeficiente de correlación de Pearson. Este coeficiente consiste en un índice de medición de la covariación existente entre un par de variables en relación lineal, es decir, la correlación de Pearson se utiliza para detectar correlación lineal entre las variables. El índice de Pearson, que toma valores dentro del intervalo $[-1, 1]$, es un indicador de cuán asociadas entre sí están dos variables. Si el coeficiente de un par de variables es cero, significa que no es posible determinar alguna covariación o correlación lineal entre las variables. Por el contrario, si el índice es cercano en valor absoluto a 1, significa que existe una fuerte covariación entre las variables y su signo indica si la covariación es positiva o negativa. Es decir, un coeficiente negativo y cercano a al valor -1 da indicios de una correlación inversa entre las variables (un valor alto de una de ellas coexistirá con un valor bajo en la otra, y viceversa), y un coeficiente positivo cercano a 1 indica que las variables tienen correlación directa (existe coocurrencia en los valores altos y bajos de ambas variables).

La Figura 4.12 exhibe la representación gráfica de todos los coeficientes de Pearson obtenidos para cada par de variables de nuestro dataset de catorce variables derivadas, donde puede apreciarse visualmente la correlación existente entre los atributos.

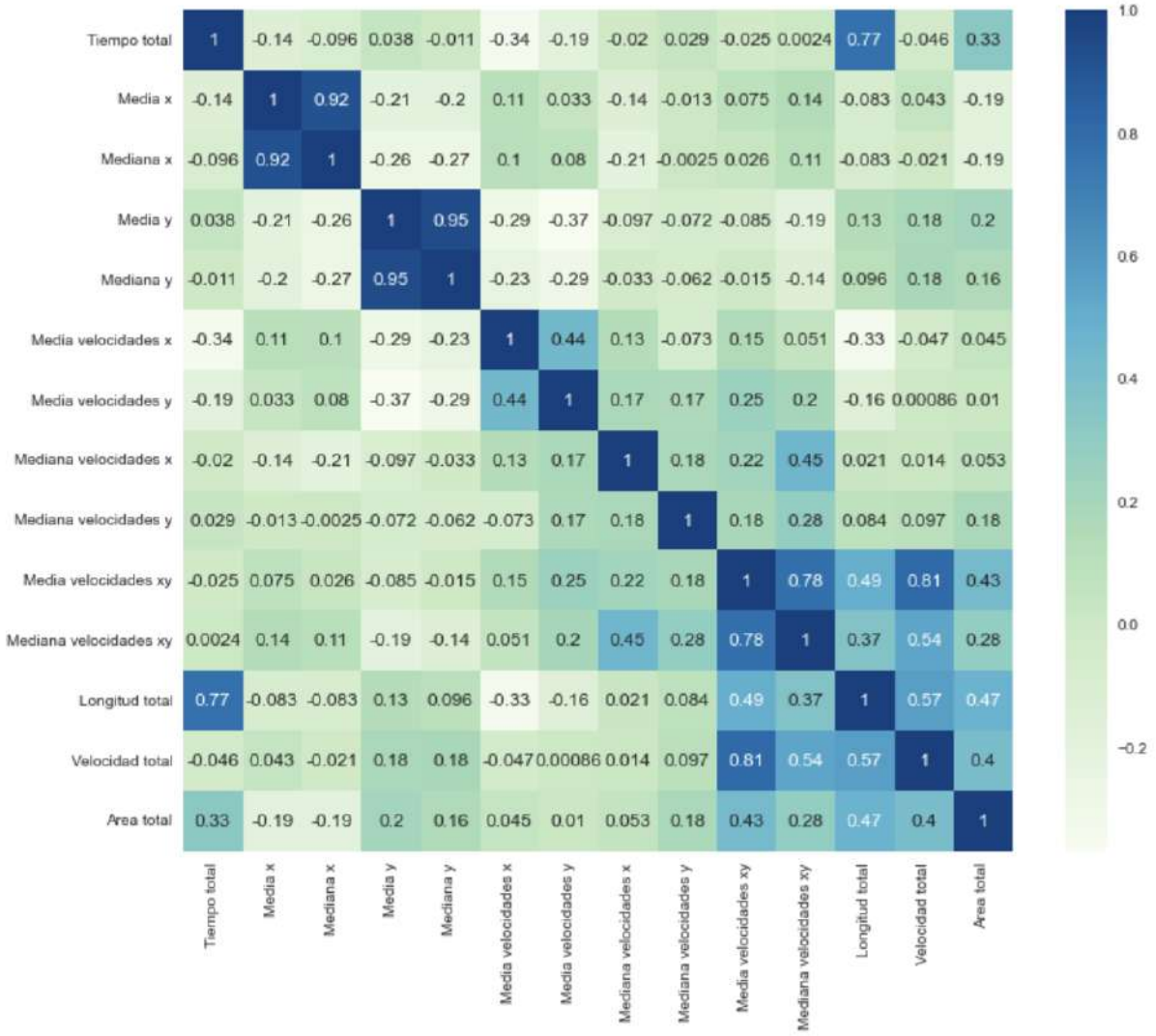


Figura 4.12: Mapa de correlación de Pearson entre las variables derivadas

A partir de este cálculo de la correlación de Pearson, obtenemos información respecto a que los pares de variables altamente correlacionadas son, en orden descendente de correlación: Media x-Mediana x; Media y-Mediana y; Media de velocidades medias puntuales bidimensionales-Velocidad total; Media de velocidades medias puntuales bidimensionales-Mediana de velocidades medias puntuales bidimensionales; Longitud total-Tiempo total.

Cabe mencionar que la selección de ocho variables realizada a partir de los cuatro métodos aplicados en la sección 4.6.1 no incluye a la mayoría de los pares de variables de alta correlación, sino que solo incluye a una variable de cada par. Este hecho ratifica la elección realizada en la sección anterior de variables a considerar para el agrupamiento como un conjunto de variables sin redundancia, al menos en el sentido de la correlación lineal.

4.7. Detección de atípicos

La detección de anomalías o detección de valores atípicos es la tarea de hallar instancias que se desvían fuertemente de la norma del conjunto de muestras, es decir, observaciones significativamente diferentes o inusuales dentro de un conjunto de datos. Estas instancias se llaman anomalías, valores atípicos o outliers, mientras que las instancias normales se denominan inliers. La detección y análisis de datos atípicos se basa en la suposición muy sencilla y efectiva de que las observaciones comunes

son “normales”, mientras que los eventos poco probables son anomalías. Así, mediante técnicas conocidas como de detección de anomalías se trata de identificar, e incluso eliminar, observaciones que son estadísticamente diferentes del resto de las observaciones y esta identificación es útil en una amplia variedad de aplicaciones, ya que la detección y posterior eliminación de los valores atípicos de un conjunto de datos antes de entrenar puede mejorar significativamente el rendimiento del modelo resultante.

En el caso que nos ocupa, de aprendizaje automático no supervisado, los modelos no se basan en etiquetas sino que debemos confiar en las propiedades de todo el conjunto de datos con que contamos para descubrir las similitudes y resaltar las diferencias entre las distintas trayectorias observadas de seguimiento ocular y así detectar posibles registros atípicos.

En esta sección realizaremos un estudio de posibles observaciones anómalas en el dataset de estudio de esta tesis, ya con la dimensión reducida a ocho variables, mediante la aplicación de distintos métodos de detección de registros atípicos. En el apéndice B se incluye una muy breve descripción de las técnicas que utilizaremos para tal fin.

Para realizar la detección de anomalías en nuestros datos fueron aplicadas las siguientes técnicas: Mezcla Gaussiana con tres distintos valores de umbral (0.03, 0,05 y 0,1); HDBSCAN con dos valores de umbrales (0,9 y 0,82); “Isolation forest” con valores de parámetros “max_samples”= 100, “random_state”= 0 y “contamination”= 0,1; “Local outlier factor” con parámetros “n_neighbors”= 30 y “contamination”= 0,1; “Elliptic envelope” con “contamination”= 0,1 y “One class support vector machines” con parámetros “nu”= 0,2, “kernel”= rbf y “gamma”= 0,001.

En orden descendente de frecuencia de aparición en los métodos utilizados, se obtuvieron como posibles datos atípicos los registros $40, 86, 106$ (en los nueve métodos aplicados) y $68, 66, 0, 1, 89$ (en la mayoría de los métodos con distinta cantidad de apariciones según la técnica aplicada).

Para realizar un estudio de los datos obtenidos como atípicos por los métodos, en primer lugar se visualizaron las trayectorias indicadas como anómalas en búsqueda de algún indicio de la posible causa de la diferencia entre estos registros y las demás observaciones. A partir de este análisis visual de las trayectorias señaladas como atípicas, en ellas se detectó una predominancia del movimiento ocular durante la resolución del test hacia un sector de la pantalla, es decir, el movimiento ocular de las observaciones anómalas no parece estar distribuido uniformemente en toda la pantalla, cosa contraria a lo esperable dada la disposición de los números en la imagen del test. Además, la exploración inicial de los datos realizada en la sección 3 mostró que las medias y medianas de las variables espaciales x e y son cercanas al valor 0,5 (lo que se condice con la forma del TMT-A), por lo que estas trayectorias atípicas no parecen estar dentro de esos valores esperables. Considerando esta situación, se calculó el porcentaje de datos registrados por cada individuo en el semiplano superior y en el semiplano derecho de la pantalla, es decir, el porcentaje de datos tales que $x > 0.5$ y $y > 0.5$. Estos resultados mostraron que efectivamente el registro 68 es el que tiene el mayor porcentaje de datos en la parte superior de la pantalla (un 72 %, muy lejano al siguiente registro con solamente el 64 %) y el registro 106 es el que tiene el mayor porcentaje de datos en el cuarto cuadrante (83 % en el semiplano derecho y 93 % en el semiplano inferior, seguido por un registro con 70 % a derecha y 81 % debajo de 0,5).

Por otro lado, a partir de la inspección de las trayectorias en relación al test se notó que dos de las observaciones reconocidas como atípicas no habían recorrido todos los números de la pantalla del test (lo que indica una resolución incorrecta del TMT-A). Dos de ellas tenían una trayectoria especialmente densa en la pantalla y otras dos tenían un aspecto muy desordenado, incluso indicadas como “de las resoluciones más desordenadas” por los especialistas en el estudio del seguimiento ocular durante tests TMT. De todos modos, este comportamiento observado solo a la luz de la representación gráfica de las trayectorias, no es suficiente para explicar su atipicidad en el conjunto de registros, dado que no podemos afirmar que son los únicos datos con esas características.

Sin embargo, el grupo de investigadores que realizó las mediciones ha llevado a cabo varios estudios de modelado y caracterización de datos de seguimiento ocular que los ha provisto del conocimiento de algunos parámetros que permiten detectar datos mal registrados, por ejemplo, registros que no cumplen con un muestreo temporal menor a 40 milisegundos (lo esperable es que los datos se registren cada 11

milisegundos) u observaciones que no presentan variación de posición de un registro a otro (hecho que es imposible considerando la mecánica del movimiento ocular). De esta manera, en un estudio conjunto con los expertos del tema, analizamos estos parámetros de datos mal registrados en los registros que resultaron atípicos luego de la aplicación de los métodos de detección, evaluando el porcentaje de datos mal registrados en cada trayectoria indicada como anómala. Este análisis muestra que tres observaciones de las detectadas como atípicas (registros $40, 86, 66$) son los tres registros con mayor porcentaje de error de frecuencia de muestreo y que también tienen un alto porcentaje de error en el registro de los desplazamientos espaciales dx y dy. El caso del registro 89, marcado como atípico por los métodos aplicados, también fue observado como un dato con alto error de frecuencia temporal de muestreo (el quinto mayor valor de error de frecuencia) y es uno de los individuos de más bajo tiempo de resolución (unos 27 segundos respecto a una media de más de 37 segundos).

En la Tabla 4.8 se muestran los valores de los parámetros analizados por los especialistas y los siete registros con mayor error en la frecuencia de muestreo, junto con su error en las mediciones de desplazamiento en las variables x e y, donde pueden observarse los cuatro registros señalados como atípicos por los métodos y el valor de sus parámetros indicadores de "malos registros".

Registro	**% error de frecuencia**	**% error desplazamiento en x**	**% error desplazamiento en y**
61	1,10041	0	0
89	1,13941	0	0
85	1,21894	0.0937647	0.0937647
88	1,49254	0	0
66	1,62602	0.0774293	0.0774293
86	1,66503	0.130591	0.130591
40	5,57873	3.44108	3.44108

Tabla 4.8: Tabla de porcentajes de errores en los registros con mayor número de errores según la frecuencia de muestreo y las mediciones en cada coordenada espacial.

Como último análisis en relación a los datos atípicos, se ordenaron todos los registros de acuerdo a la variable "Tiempo total" de resolución del test y se comprobó que las trayectorias $0, 1, 68, 106$ obtenidas como atípicas son resoluciones del test correspondientes a los registros de quienes tardaron tiempos extremos de resolución entre los 108 sujetos, siendo las observaciones 0 y 1 las de mayores tiempos de resolución y los registros 68 y 106 los de menores tiempos. De hecho, las trayectorias temporalmente más largas tienen una amplia diferencia respecto del resto de los datos, triplicando la media de tiempos de resolución en el caso del segundo registro (número 1) y superando el 250 % del valor del promedio de tiempo en el caso del primer registro (indicado como 0).

Por lo expuesto, estamos en condiciones de afirmar que los métodos de detección de anomalías aplicados en este trabajo han detectado como registros atípicos a los datos correspondientes a observaciones con errores de medición, escaneos visuales sesgados hacia alguna zona particular de la pantalla (números del TMT-A sin explorar) y trayectorias con tiempos de resolución atípicamente largos o cortos.

De esta manera, cabe mencionar que el proceso de detección de datos atípicos que se ha llevado a cabo tiene una clara interpretación en el marco del test TMT-A, más allá de la posible detección de observaciones mal registradas, dado que da indicios de resoluciones incorrectas del test, sesgos espaciales en la observación de la pantalla durante la prueba y tiempos de resolución notoriamente alejados de la media. Así, el proceso de localización de anomalías realizado resulta significativo para el problema en estudio y puede ser útil como método de detección de comportamiento atípico de personas a la hora de resolver el TMT-A o sujetos que sistemáticamente no realizan el test en forma correcta.

4.8. Determinación del número de grupos

Una de las dificultades con que nos encontramos al aplicar alguno de los métodos de agrupamiento disponibles es la elección del número de grupos, ya que algunas técnicas usan este número como parámetro para realizar el agrupamiento y debe ser introducido por el usuario. En general, y en nuestro

problema en particular, no es posible visualizar a priori la cantidad de grupos en que se pueden agrupar los datos y la elección de esa cantidad es un tema de suma importancia dentro del proceso de clustering. No existe un criterio objetivo ni ampliamente válido para la elección de un número óptimo de clusters en un problema dado pero una mala elección de ese número puede dar lugar a la realización de agrupaciones de datos muy heterogéneos (demasiados pocos clusters); o una agrupación en clusters diferentes de datos muy similares (demasiados grupos). El número óptimo de grupos es, de alguna manera, una elección muy subjetiva y depende del método considerado para medir las similaridades y los parámetros utilizados para el agrupamiento. De todos modos, a pesar de que no existe un criterio objetivo para la selección del número de grupos, se han implementado diferentes métodos que nos ayudan a elegir un número apropiado de clusters para agrupar un cierto conjunto de datos y en esta sección haremos uso de algunos de ellos para determinar el número "óptimo" de grupos en nuestro dataset.

En el apéndice C incluímos una breve descripción técnica de los métodos que utilizaremos con el fin de elegir el número de grupos para el que realizaremos el análisis de clustering de nuestros datos de seguimiento ocular.

Concretamente, para continuar el análisis de agrupamiento de los datos de estudio con un número "apropiado" de grupos a formar en mente, aplicamos a nuestro dataset los métodos siguientes:

- K-Medias optimizado por la aplicación de las técnicas conocidas como "Elbow method" y "Gap statistic" con parámetros $nrefs = 5$ y $maxClusters = 15$ (los gráficos pueden apreciarse en las Figuras 4.13 y 4.14). Para estos resultados, calculamos los índices de validación "Silhouette", "Davies Bouldin" y "Calinski-Harabasz".

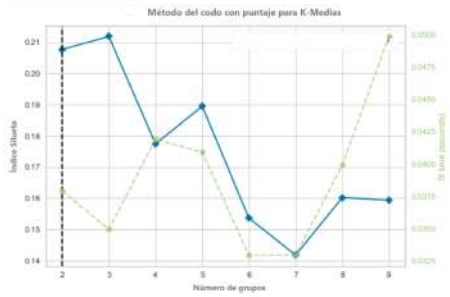


Figura 4.13: Codo de silueta para K-Means.

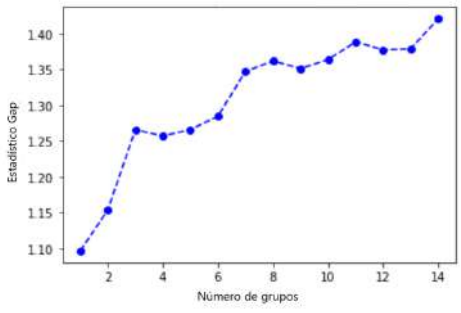


Figura 4.14: Gap para número de grupos.

- Agrupamiento jerárquico visualizando los dendogramas, por ejemplo el de la Figura 4.15, correspondientes con cinco distintos enlaces ("linkages"): "ward", "single", completo, promedio, centroide y mediana.

- Mezcla Gaussiana con tipos de covarianza esférica, "tied", diagonal y "full", variando el número de posibles grupos entre 1 y 14, y analizando los índices "BIC" de los agrupamientos obtenidos en cada caso.

Los métodos evaluados sugieren distintos posibles números de grupos:

K-Means: Gap Statistic y Elbow (3 u 8 grupos), Calinski-Harabasz (3 grupos), Silhouette (3 grupos), Davies-Bouldin (5 o 9 grupos).

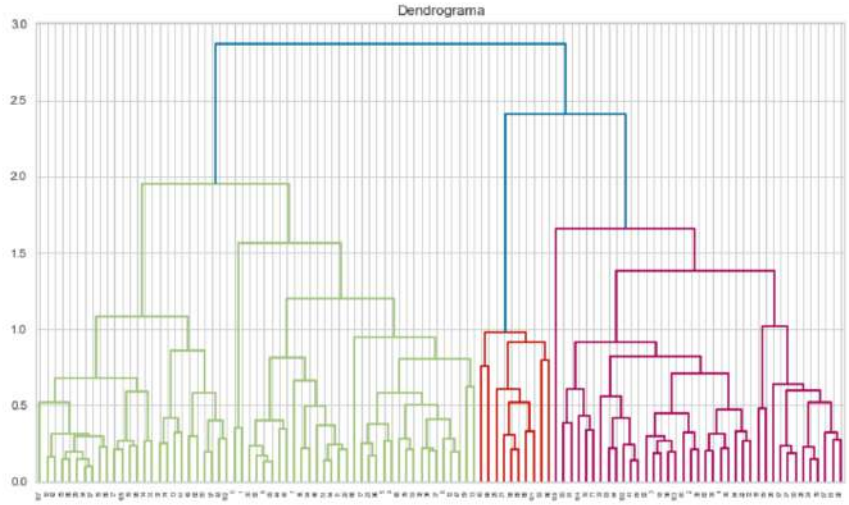


Figura 4.15: Resultado de uno de los agrupamientos jerárquicos.

Dendrogramas: Ward (3 grupos), Single linkage (2 grupos), Complete linkage (3 grupos), Average linkage (2 o 4 grupos), Centroid (3 grupos), Median (2 grupos).

Gaussian Mixture Models – BIC: 3 y 5 grupos.

En estos resultados se observa un consenso mayoritario entre los métodos en la sugerencia de tres grupos como número óptimo. También se identifica cierta coincidencia en torno a la posibilidad de dos y ocho grupos. En particular, el método de Silhouette —que evalúa simultáneamente la cohesión interna y la separación entre grupos— señala que tres grupos ofrecen la estructura más consistente. Esta misma conclusión es respaldada por varios de los métodos jerárquicos.

En esta etapa, se realizó una consulta a especialistas en registros de seguimiento ocular con el objetivo de conocer su opinión sobre cuál sería el número de grupos más adecuado para los datos analizados, particularmente entre las alternativas de 3 y 8 clusters.

Los expertos consultados pertenecen al grupo de investigación NEUFISUR, conformado por profesionales de diversas disciplinas: física, matemática, ingeniería electrónica, psicología, psicopedagogía y musicoterapia. Este grupo colabora, además, con el equipo interdisciplinario del Centro Integral de Neurociencias Aplicadas (www.cinabb.org), que abarca prácticamente todas las áreas de la salud mental.

Ante la consulta, los especialistas señalaron que, a priori, una solución con un número reducido de clusters resulta más natural e interpretable en este contexto, por lo que tres grupos fue considerado como una opción más razonable y coherente que ocho. Esta opinión coincide con los resultados obtenidos por la mayoría de los métodos cuantitativos aplicados para la estimación del número óptimo de grupos.

De esta manera, dado que el propósito del análisis es ofrecer una descripción general de los patrones presentes en los datos y considerando la concordancia entre la evidencia empírica y el criterio experto, se adoptó finalmente la partición en tres grupos como la solución más adecuada para continuar con el análisis.

4.9. Agrupamiento

En esta sección aplicaremos distintos métodos de clustering para obtener una separación en grupos o clústers de nuestro dataset de variables derivadas de los registros de posición ocular de las 104 personas (las 108 menos los 4 atípicos) de estudio durante la resolución del TMT-A.

No existe una definición universal de lo que es un clúster, depende del contexto, y diferentes algoritmos de agrupamiento capturarán diferentes tipos de grupos. Algunos algoritmos buscan instancias centradas en un punto particular llamado centroide, otros buscan regiones continuas densamente pobladas de instancias y estos grupos pueden adoptar cualquier forma, algunos otros algoritmos son jerárquicos y buscan grupos de grupos, y la lista continúa.

A lo largo del tiempo se han propuesto una gran variedad de algoritmos de agrupamiento, sin embargo, no existe un método de agrupamiento capaz de encontrar correctamente la estructura de todos los conjuntos de datos. La aplicación de un algoritmo particular de agrupamiento a un conjunto de datos objeto, implica una organización de los datos con el criterio interno propio del método, las características de la medida de proximidad (función de distancia o (dis)similitud) utilizada y del conjunto de datos *per se*. Por lo tanto, pueden obtenerse muy diferentes resultados para el agrupamiento de un mismo conjunto de datos al aplicar dos algoritmos de agrupamiento diferentes.

Dada la diversidad de métodos de agrupamiento que existen y la incerteza respecto a qué métodos pueden resultar óptimos para un cierto conjunto de datos u objetivo de clasificación particular, en este trabajo se exploraron los resultados de varias técnicas de clustering aplicadas a los datos de análisis pero solo se indicarán los procedimientos que arrojaron resultados relevantes e información de interés para el problema en discusión desde el enfoque considerado en cada caso.

Una breve descripción de los métodos de agrupamiento que se utilizaron en esta tesis pueden consultarse en el apéndice D y remitimos a las referencias allí mencionadas para obtener más información de las técnicas.

Cada uno de los métodos fue aplicado al dataset de estudio utilizando distintos valores de sus parámetros con el objetivo de encontrar la mejor combinación de valores paramétricos que arroje como resultado un "mejor", más claro o más significativo agrupamiento de los datos.

En lo que sigue, se detallan los métodos de clustering y los correspondientes valores de sus parámetros que fueron aplicados al conjunto de datos de variables derivadas de los registros de los 104 sujetos, y algunas conclusiones de su análisis comparativo.

El método K-Medias fue estudiado en relación al número de grupos (en la sección 4.8) calculando el índice Silueta y el "homogenity score", y se concluyó que el número óptimo de clusters en el conjunto de datos bajo estudio es tres. Se analizó la variación de calidad del método K-Means en relación al *random states* experimentando con los valores $0, 1, 2, 3, 19, 25, 42, 2595$ y se comprobó que es prácticamente nula en nuestro caso de estudio.

Se aplicó la técnica DBSCAN con valores de parámetro eps = 0,1, 0,3 y 0,5; y un mínimo de muestras por cluster igual a $2, 3, 5$ y 8.

También se analizó HDBSCAN con los tres índices para variaciones en el algoritmo, con opciones "best" y "generic", valores del parámetro "alpha" de 1 y 1,5, métrica euclideana y Manhattan, y un mínimo tamaño de grupo de 2 y 5 observaciones. Las mejores combinaciones de parámetros fueron $algorithm = generic, alpha = 1, metric = euclidean$ y $minclustersize = 2$ y, por otro lado, $algorithm = best, alpha = 1, metric = manhattanyminclustersize = 2$. En ambos casos, aunque los resultantes no son agrupamientos extremadamente buenos, el método agrupa los datos en 3 clusters y da indicios de posibles datos atípicos.

El método jerárquico aglomerativo se aplicó con métricas euclideana, "manhattan" y coseno; y con enlaces (o "linkages") complete, "ward", promedio y "single", en todas las combinaciones posibles. Analizando los dendogramas obtenidos de cada agrupamiento, los mejores resultados se obtuvieron con las combinaciones de métrica euclídea con linkage ward (mejor resultado) y complete, aunque con no tan buenos índices de calidad del clustering.

En el caso de la aplicación del método espectral, se realizaron variaciones del parámetro "affinity" y de "nearest neighbors" y "rbf"; también se experimentó con distintos valores del parámetro "assign labels" entre K-Medias y discreto. Se realizaron las 4 combinaciones y se compararon los resultados respecto a los índices Silueta, Calinski-Harabasz y Davies-Bouldin. En el primer análisis, el resultado más deseable es el caso "affinity=rbf" y "assign labels=discretize"; y le sigue la combinación "affinity=rbf" y "assign labels=kmeans". Variando el número de clusters entre 2 y 7, este método origina el mejor resultado con dos grupos y, como segunda mejor opción en relación a los índices Silueta y "homogenity score", el caso de tres grupos.

El algoritmo de mezcla gaussiana fue aplicado variando el parámetro "covariance type" entre las opciones "full", "tied", diagonal y esférica, y se analizaron los resultados a través de los índices silueta, Davies-Bouldin y Calinski-Harabasz para el caso de 3 grupos. No se observaron grandes diferencias entre los resultados de los distintos métodos, excepto que hay dos casos con un par de índices con

mejores valores que el restante. Considerando más deseables los casos que tienen dos de los índices con mejores valores, nos inclinamos por los tipos de covarianza "full" y esférico como resultados aceptables del método.

El caso del método Mezcla Gaussiana Bayesiana fue analizado con los mismos cuatro valores del parámetro "covariance type" y con "init" para valores random y kmeans. Calculados los mismos tres índices, los agrupamiento de mejores resultados son las combinaciones de tipo de covarianza "tied" y "full" con el método kmeans.

El método BIRCH se analizó variando el umbral y el parámetro "branching factor". Los casos estudiados fueron: $branchingfactor = 50$, para 3 clusters con un umbral de valor 0,1; $branchingfactor = 50$, para 3 clusters con un umbral de valor 0,05; $branchingfactor = 35$, para 3 clusters con un umbral de valor 0,1; $branchingfactor = 35$, para 3 clusters con un umbral de valor 0,05; $branchingfactor = 50$, para 3 clusters con un umbral de valor 0,45; $branchingfactor = 25$, para 3 clusters con un umbral de valor 0,26. Todos estos casos se compararon calculando los tres índices Silueta, Davies-Bouldin y Calinski-Harabasz, y se comprobó que no se producen grandes cambios a partir de la variaciones analizadas de los parámetros, salvo al variar el umbral, y el mejor agrupamiento indicado por los índices se obtuvo con los valores $branchingfactor = 50$ y umbral de 0,45, siempre para tres grupos.

Mini Batch K-Means se aplicó a los datos analizando cambios de los hiperparámetros "init" (con valores "random" y "kmeans"), "batch size" (para valores 50 y 100), "reassignment ratio" (valores 0,001 y 0,005) y $randomstate = 1000$. Luego del análisis con los índices mencionados, los mejores agrupamientos resultaron ser los obtenidos con "init" de valor "kmeans" sin diferencia variando los demás parámetros. Consideramos el caso $batch = 100$ y $ratio = 0.005$.

La técnica conocida como OPTICS fue aplicada con métricas Euclídea y de Minkowski, casi con iguales resultados, y con un mínimo de muestras por grupo de 2,3 y 5 en cuya variación sí se observa un cambio en el agrupamiento resultante. Analizados los tres índices, se decidió continuar el estudio con el resultado del caso de un mínimo número de 5 muestras por grupo con la métrica euclídea.

4.9.1. Obtención de grupos

En base a los resultados obtenidos a partir de la variación de parámetros de las distintas técnicas comentadas en la sección anterior, específicamente se sometió al conjunto de datos de las variables derivadas a los siguientes métodos y valores de parámetros: affinity propagation ("damping" entre 0,5 y 1) que solo arroja agrupamiento para 14 clusters; K-Medias para 3 grupos con parámetros por defecto (*algorithm=auto, copy x=True, init=K-Means++, max iter=300, n clusters=3, n init=10, n jobs=None, precompute distances='auto', random state=None, tol=0.0001, verbose=0*); aglomerativo con métrica Euclídea y "linkage" ward; espectral para 3 grupos (*affinity=rbf* y *assign labels=discretize*); mezcla Gaussiana ("covariance type" de valores "full" y "spherical") también para 3 clusters; BIRCH ($branchingfactor = 50, nclusters = 3, threshold = 0{,}45$); "Mini Batch K-Means" ("init=kmeans++, batch size=100", "reassignment ratio=0,005, random state=1000"); DBSCAN ($eps = 0,38, minmuestras = 2$) que arroja 2 grupos como resultado; mezcla Gaussiana Bayesiana (con parámetros "full" y "kmeans"); OPTICS con la métrica euclídea y con un mínimo de muestras por grupo de 5 observaciones, que no arroja resultado de agrupamiento; HDBSCAN para algoritmo genérico $alpha = 1$, métrica euclideana y mínimo tañamo de grupo de 2 muestras, y también para el algoritmo "best" mismo valor de *alpha*, la métrica conocida como "Manhattan" y un tamaño mínimo de grupo de 2 muestras; "MeanShift", sin obtener agrupamiento alguno; y se aplicaron los métodos de consenso de tres aplicaciones de K-Medias para 3 clusters utilizando el método "Simple Consensus Clustering" (con parámetros $nclusters = nc, nclustersbase = 10, ncomponents = 30, ncbrand = False$); y un combo de 3 métodos distintos con Kmeans, "MiniBatch" y Aglomerativo para 3 grupos usando "evidence accumulation".

Todos estos métodos fueron aplicados al dataset de variables derivadas correspondientes a los 108 sujetos bajo estudio luego de proceder a la extracción de los registros atípicos detectados con errores de frecuencia de muestreo, es decir, extrayendo los registros 40, 86, 66 y 89.

Además, considerando que nuestro objetivo es detectar patrones de resolución y no estamos intere-

sados especialmente en analizar las observaciones en forma particular, descartamos los agrupamientos que resultaron de la aplicación de esos métodos y tenían grupos de menos de 3 elementos o dejaban "demasiados" registros sin asignar a un grupo.

Una vez creados los clústeres de cada método, el modelo genera una etiqueta para cada observación (fila), que representa el grupo al que pertenece ese registro. La Figura 4.16 presenta, únicamente a modo ilustrativo, el encabezado del dataset que se utilizará en los análisis posteriores. Este conjunto de datos incluye todos los registros sin errores de frecuencia y las etiquetas asignadas por cada uno de los métodos de clustering previamente descritos, considerando particiones en tres grupos de al menos dos elementos cada uno.

Cabe aclarar que las etiquetas de grupo no poseen un significado intrínseco; su inclusión tiene únicamente fines organizativos. El objetivo de la figura es mostrar la estructura del dataset consolidado que servirá como base para las siguientes etapas del análisis, tales como la generación de ensambles de agrupamientos.

	KMeans	AgglomEW	Spectral	GaussianFull	GaussianSpherical	BIRCH	MiniBatch	ComboDe3	KMCONS
0	0	1	1	2	2	0	2	1	0
1	0	1	1	2	2	0	2	1	0
2	2	0	0	0	0	2	1	2	1
3	1	0	0	0	1	1	2	2	1
4	2	0	2	0	0	2	1	2	1
5	0	1	1	1	1	0	0	1	0
6	0	1	1	1	1	0	0	1	0
7	0	1	1	1	2	0	0	1	0
8	0	1	1	1	1	0	0	1	0
9	0	1	1	1	1	0	0	1	0

Figura 4.16: Resultados de los agrupamientos considerados

4.9.2. Análisis de los agrupamientos

Más allá de la dificultad inherente a un problema de aprendizaje no supervisado, el problema de revelar un agrupamiento significativo de los datos se deriva, por un lado, del amplio espectro de estructuras de conglomerados que pueden obtenerse, abarcando grupos de forma arbitraria, tamaños de conglomerados altamente desequilibrados, densidades de agrupación variables, y posibles superposiciones de grupos y, por otro lado, de la idoneidad de la representación de datos seleccionada y los criterios de agrupamiento.

A la vista de los distintos resultados obtenidos mediante la aplicación a nuestro dataset de los muy variados algoritmos de agrupamiento, a priori, podemos afirmar que todas las soluciones pueden ser igualmente plausibles, incluso más allá de la evaluación que realizamos y de alguna otra que podemos realizar con métodos generales, respecto a la "calidad" del agrupamiento.

Es así que, concentrándonos en nuestro objetivo de obtener posibles patrones de resolución del test TMT-A por parte de los individuos analizados, fijamos nuestra prioridad en hallar una clusterización de los datos que arroje algo de luz sobre las diferencias en las trayectorias espacio-temporales de resolución de los sujetos, más allá de la validación del método de agrupamiento que la origina. En otras palabras, resulta más relevante en este punto el análisis de qué datos se agrupan con cada método, que si el clustering tiene una performance de excelente calidad. Por tanto, luego de obtener los resultados del proceso de clustering, en lo que sigue nos ocuparemos de interpretar las agrupaciones logradas.

Esta tarea de comprender o interpretar el resultado de un agrupamiento es quizás más importante que realizar el clustering. El proceso de creación de grupos tiene una orientación matemática, sin embargo, la interpretación del significado de los grupos es una combinación de matemática e intuición, además de conocimiento del problema entre manos. No hay una manera fácil de entender las características de cada clúster. De todos modos, existen algunas técnicas para intentar comprender el resultado de un agrupamiento, por ejemplo análisis estadísticos, visualizaciones de comparación de grupos, estudios de relevancia de las variables en la caracterización del agrupamiento (impacto de las variables en las

etiquetas) y algunas otras técnicas. Pueden obtenerse algunos detalles y características de los métodos utilizados en este trabajo en el apéndice E.

Nos abocamos ahora a la tarea de analizar e interpretar los resultados de los agrupamientos que obtuvimos para nuestro dataset. Comenzaremos el análisis de los agrupamientos considerando la caracterización univariada de los grupos mediante la visualización de los gráficos de boxplots de los nueve agrupamientos con que estamos trabajando cuyas técnicas y primeras etiquetas pueden verse en la Figura 4.16.

A partir de la observación de los gráficos de boxplots de cada uno de los agrupamientos obtenidos por los distintos métodos, podemos realizar los siguientes comentarios.

El método aglomerativo con distancia euclídea y enlace "ward" no muestra diferencias claras entre los tres grupos en los valores de las distintas variables, solo diferencia un grupo de los otros dos en los atributos de valores medios en la coordenada y, y en la media de velocidades medias puntuales en y y de velocidades medias puntuales bidimensionales. Sí se observa una diferencia notoria en la variable que corresponde a la velocidad de la trayectoria de resolución del test entre dos de los tres grupos.

El agrupamiento BIRCH distingue un grupo de los otros dos en las variables media en y, media de velocidades medias puntuales en y y media de velocidades medias puntuales bidimensionales.

El caso del clustering obtenido de la combinación de tres métodos distintos, llamado Combo de 3 en la tabla de la Figura 4.16, divide los datos en dos grupos con diferencias en los valores medios de la coordenada y, de la media de velocidades medias puntuales en y y de la media de velocidades medias puntuales bidimensionales.

Un grupo se distingue de los otros dos en los clústers generados por el algoritmo Mezcla Gaussiana con parámetro "Full" en la variable media en y y media de velocidades medias puntuales en la coordenada y. Con el parámetro "Spherical" en lugar de "Full", la Mezcla Gaussiana distingue un grupo de dos en las variables "media de velocidades medias puntuales en y", "media de velocidades medias puntuales bidimensionales" y "velocidad total".

El método de K-Medias para tres grupos distingue un cúster de los otros dos en tres variables que son: media de velocidades medias puntuales en y, media de velocidades medias puntuales bidimensionales y velocidad media de la trayectoria de resolución (ver definición en página 17).

El algoritmo de agrupamiento que consiste en la combinación de tres métodos K-Medias con distintos parámetros indica una diferencia en las variables correspondientes al valor medio en la coordenada y, en el valor medio de velocidades medias puntuales en la ordenada y y en el valor medio de velocidades medias puntuales bidimensionales.

Solo en los atributos media de velocidades medias puntuales en y y media de velocidades medias puntuales bidimensionales se observan diferencias en los valores de los grupos generados por el método Mini Batch K-Means, y solamente en un grupo respecto de los otros dos.

El método espectral muestra alguna diferencia en los valores de las variables media de velocidades medias puntuales en y, media de velocidades medias puntuales en xy y velocidad total, solamente de uno de los grupos en relación a los otros dos.

Si bien la información que podemos extraer del análisis univariado de las características de cada grupo mediante la visualización de los gráficos de boxplots es escasa, indica que el agrupamiento originado con el algoritmo aglomerativo utilizando la distancia euclídea y el enlace "ward", parece realizar la "mejor" separación en tres grupos de acuerdo a los valores de las variables consideradas y las técnicas de clustering aplicadas, aunque la diferenciación se observa solo para un grupo respecto a los otros dos.

En la Figura 4.17 se muestran los diagramas de boxplot de las ocho variables utilizadas (ver 4.6) para realizar el agrupamiento aglomerativo mencionado, donde pueden apreciarse las diferencias en los valores de las variables consideradas en los distintos grupos.

Mencionaremos en este punto que, ante la situación de que las diferencias en valores de atributos al agrupar en tres clústers se observan principalmente en un grupo respecto de otros dos, se repitió el proceso de agrupamiento con los mismos métodos que se aplicaron para clusterizar en tres grupos solicitando la formación de solo dos grupos. Al finalizar este procedimiento, comprobamos que las caracterizaciones univariadas mostradas por los boxplots en dos grupos no reflejan las diferencias significativas en los

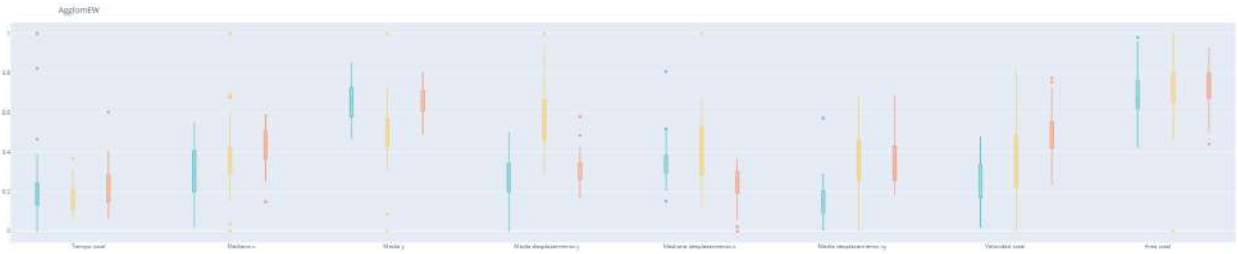


Figura 4.17: Boxplots comparativos de variables en el agrupamiento aglomerativo

valores de las variables esperadas a la luz de los mismos resultados en el agrupamiento de tres conglomerados. En otras palabras, el proceso de agrupamiento en dos clústers no arroja más información sobre la caracterización de los conglomerados en relación a las variables consideradas que el agrupamiento en tres grupos.

A continuación, realizamos un análisis de las variables que resultan más relevantes en el proceso del agrupamiento aglomerativo. La Figura 4.18 muestra las gráficas de la importancia relativa de las ocho variables consideradas en el agrupamiento, desde el punto de vista de dos técnicas diferentes. En el primero de los gráficos de barras de la figura puede observarse la importancia relativa de las variables en la distinción de los clusters en el agrupamiento que origina el método de análisis de importancia conocido como "Gradient Boosting Classifier", y el segundo de los gráficos muestra la comparación en la importancia de las variables que ve el método "Random Forest Classifier".

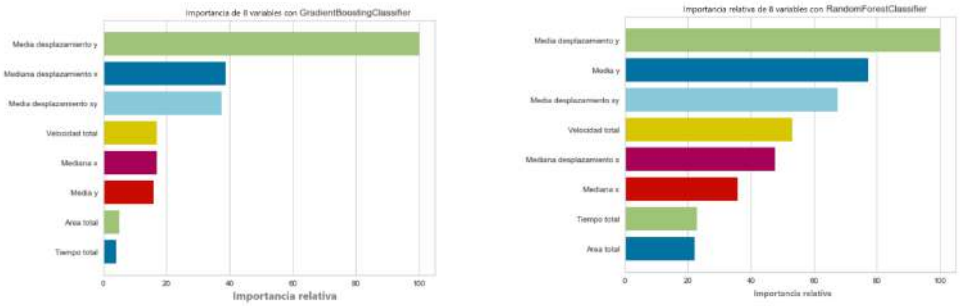


Figura 4.18: Importancia relativa de variables en el agrupamiento aglomerativo

A la vista de los gráficos de barras podemos comentar que, si bien la importancia que otorga a cada variable cada uno de los dos métodos es muy distinta, en valores absolutos y en valores relativos en relación a las demás variables, en ambos gráficos observamos que los cuatro atributos más importantes en la conformación del agrupamiento aglomerativo son los mismos cuatro que observamos como caracterizadores de los grupos en al análisis univariado que realizamos a partir de los boxplots. Es claro que desde el punto de vista teórico este comentario puede ser una obviedad, pero el hecho corrobora los resultados que obtuvimos aplicando distintos métodos a nuestro dataset y nos indica, además de que nuestros resultados se condicen entre sí, que el agrupamiento aglomerativo está dando cierta información relevante en relación a posibles grupos entre los desempeños durante la resolución del TMT-A, caracterizados por los valores de algunas variables derivadas de los registros del dispositivo de seguimiento ocular.

A título ilustrativo, en la Figura 4.19 se muestra una visualización de las proyecciones de las observaciones, que están en el espacio de dimensión ocho de variables derivadas, al espacio de las tres variables más relevantes del agrupamiento aglomerativo, que son la media de velocidades puntuales en la ordenada y, la media de velocidades medias puntuales bidimensionales y la velocidad total, coloreadas según el grupo de pertenencia en el clustering.

Para poder hacernos una idea más acabada de qué valores centrales tienen cada una de las variables

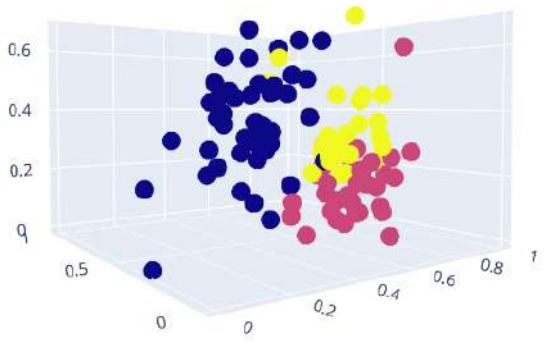


Figura 4.19: Proyecciones tridimensionales del agrupamiento aglomerativo

de análisis en cada uno de los grupos que hemos obtenido con el método aglomerativo, realizamos una visualización mediante gráficos de línea de los valores medios y medianas de cada una de las variables en cada grupo. La Figura 4.20 muestra, en primer lugar, las diferencias en los valores medios de cada variable dentro de cada clúster generado por el método aglomerativo y, en el segundo gráfico, muestra los valores de las medianas de cada variable por grupo formado.

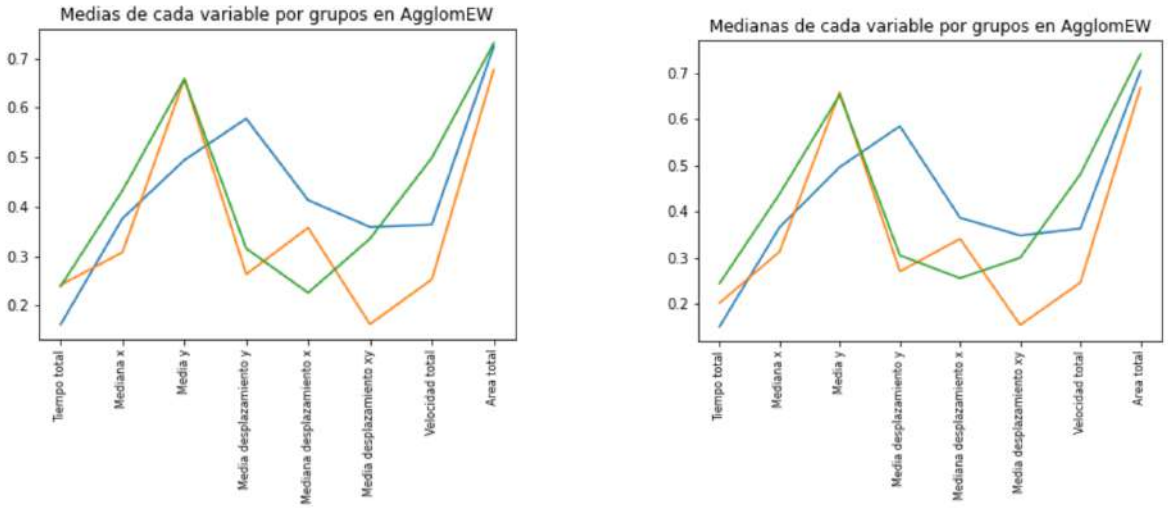


Figura 4.20: Valores centrales de las variables en cada grupo

En las gráficas de la Figura 4.20 puede apreciarse que en el agrupamiento aglomerativo diferencia a los grupos en relación a sus valores en las variables "media de velocidades medias puntuales en y", "mediana de velocidades medias puntuales en y", "media de velocidades medias puntuales en xy" y "velocidad total", aunque en estos gráficos de línea no se observan las dispersiones de las variables que permiten inferir diferencias en los valores de la variable de todo un grupo respecto de otro, que sí pueden observarse en el gráfico de boxplots. Más específicamente, los gráficos de los valores centrales de las variables en cada grupo indican que el agrupamiento aglomerativo obtiene tres grupos que se diferencian especialmente en los valores medios y medianas de las variables "medianas de velocidades medias puntuales en x" y "velocidad total" del recorrido en pantalla. En las otras variables las diferencias entre los valores son parciales, distinguiendo dos grupos del tercero.

A continuación, habiendo formado idea sobre qué variables tienen relevancia en los grupos del clustering llevado a cabo mediante el método aglomerativo, nos abocaremos a la tarea de obtener información sobre las reglas que determinan el agrupamiento, es decir, sobre las condiciones debe cumplir una observación para ser asignada por el algoritmo a un grupo u otro. Como comentamos al inicio de esta sección, una forma sencilla de generar automáticamente las reglas del clustering es entrenando un modelo de

árbol de decisión usando las variables del dataset y el resultado del agrupamiento como etiqueta. De esta manera, generamos el árbol de decisión del agrupamiento aglomerativo con distancia euclídea y enlace ward. La manera en que el proceso de clustering con el método aglomerativo determina la división de grupos puede visualizarse en el árbol de decisión que se muestra en la Figura 4.21.

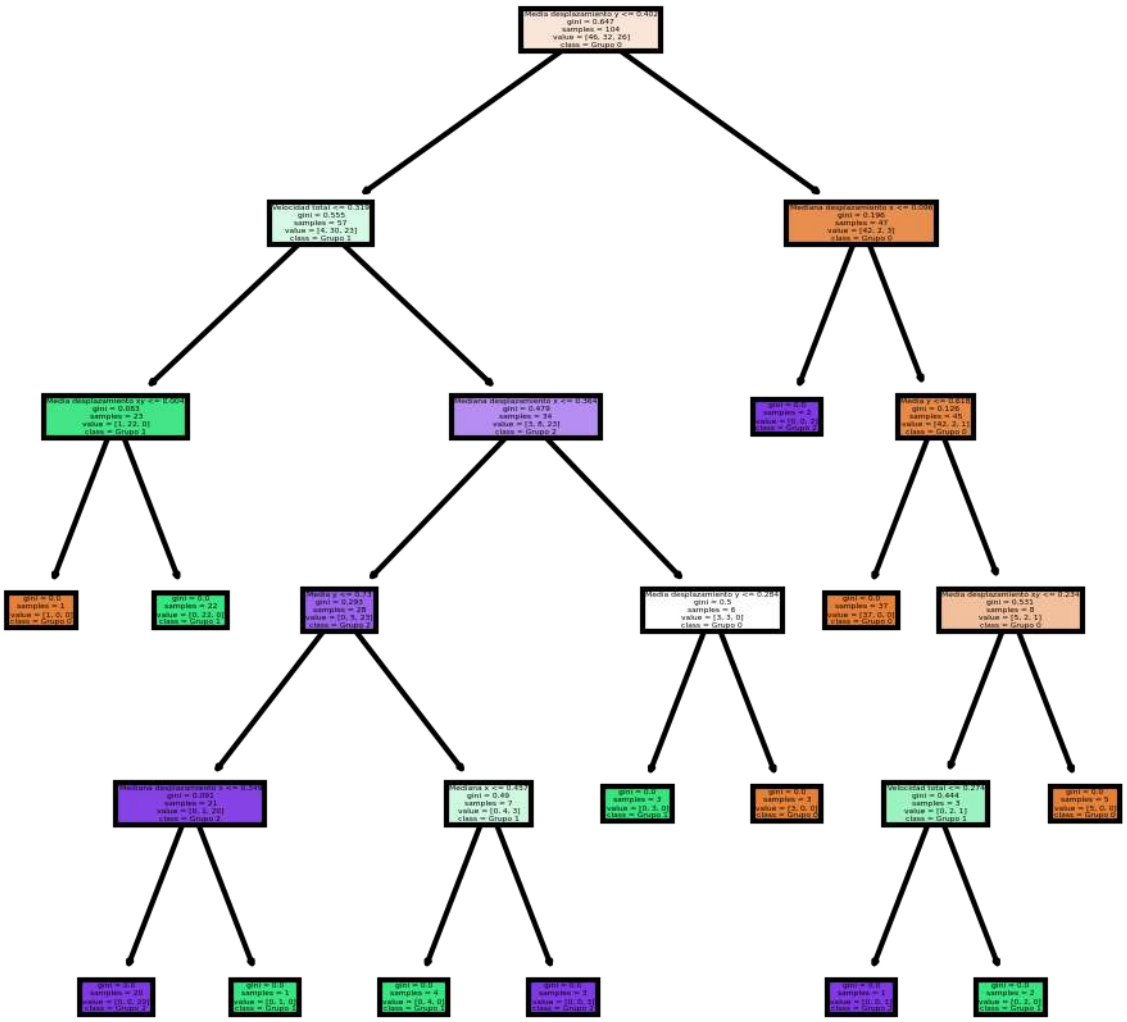


Figura 4.21: Árbol de decisión agrupamiento aglomerativo

El árbol de decisión obtenido nos permite visualizar que la primer variable considerada por el algoritmo aglomerativo para agrupar los registros es la variable "Media de velocidades medias puntuales en y" y el valor de corte es $0,402$. En segundo lugar, el agrupamiento queda determinado por los valores de la variable "velocidad total", al dividir en distintos grupos aquellas observaciones con valores menores y mayores a $0,319$, y luego considera las variables "media de velocidades medias puntuales bidimensionales" y "mediana de velocidades medias puntuales en x".

Cabe mencionar que, a partir de la información proporcionada por el árbol de decisión del agrupamiento aglomerativo, vemos que las variables reconocidas como más importantes a la hora de la distinción de los clústers obtenidos de los métodos que aplicamos en la Sección 4.9.2 y que pueden visualizarse en la Figura 4.18, son las variables que deciden las reglas de asignación a los grupos en el método aglomerativo, como es natural pero se ha explicitado en nuestro análisis.

Análisis según edad y sexo

Si bien siempre se contó con la información de la edad y el sexo biológico de algunos individuo cuyos datos de seguimiento ocular son objeto de este estudio, se decidió no incluir esas variables al llevar a cabo el proceso de agrupamiento porque, además de que la información disponible es parcial, los clústers estaban extremadamente condicionados por esos dos atributos, ya que una de esas variables es binaria y la otra tiene un orden de magnitud muy superior a los demás atributos. Sin embargo, en este momento en que ya fueron obtenidos los agrupamientos, se analizó si el sexo y la edad de los sujetos tiene relación con los grupos que han sido formados.

El análisis de la distribución por edad en los distintos grupos se realizó mediante la visualización de diagramas de boxplots. En la Figura 4.22 se observa la diferencia de edades de uno de los grupos respecto a las de los otros dos clústers en el agrupamiento aglomerativo con distancia euclídea y enlace ward. Los individuos de menor edad parecen estar agrupados mayoritariamente en uno de los grupos.

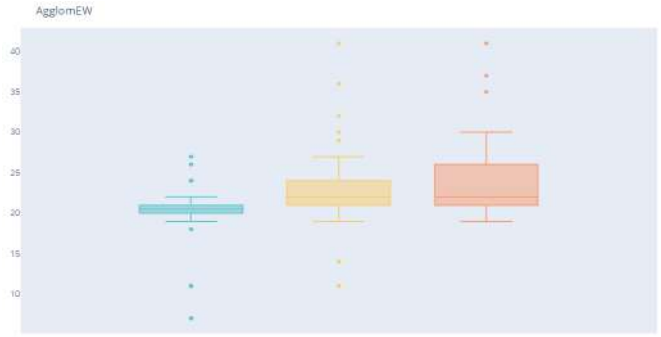


Figura 4.22: Distribución por edad en los grupos del método aglomerativo

También se realizó el analisis de la distribución etaria en los clústers de los otros agrupamientos llevados a cabo en la Sección 4.9.1 pero en ellos no se observan diferencias notorias en las edades de los sujetos en los distintos grupos. Este hecho es otro indicio de que el agrupamiento aglomerativo aprehende una importante cantidad de información de los registros de los distintos grupos.

Respecto al estudio de una posible relación entre el agrupamiento y el sexo de los individuos, en este caso se procedió a analizar un gráfico de barras de cada uno de los grupos del agrupamiento aglomerativo distinguiendo con colores cada sexo, ese gráfico puede verse en la Figura 4.23.

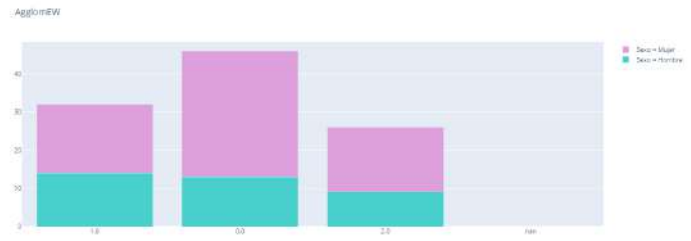


Figura 4.23: Cantidad de personas por sexo biológico en los grupos

A la luz de éste gráfico, observamos que en uno de los grupos la cantidad de personas de cada sexo es muy similar pero en los otros dos clústers hay predominancia de sexo femenino. Si bien es un hecho que el número de datos registrado de mujeres es mayor, el gráfico parece mostrar una leve tendencia del sexo femenino a agruparse en los dos últimos clústers que se presentan en el gráfico de barras, en particular, en el segundo de ellos. Sobre las resoluciones de personas de sexo masculino, podemos comentar que se agrupan en cantidad similar en los dos primeros grupos y en menor cantidad en el tercer grupo, aunque la distribución de los hombres prácticamente es de un tercio del número de individuos en cada uno de los dos últimos grupos.

4.9.3. Ensambles

Como hemos mencionado con anterioridad, es sabido que no existe un método de agrupamiento que sea capaz de encontrar correctamente la estructura subyacente de cualquier conjunto de datos. De hecho, un dilema bien conocido en el agrupamiento es la multitud de modelos para el mismo conjunto de datos basados en diferentes algoritmos, diferentes medidas de distancia, diferentes características a elegir de los objetos, diferentes escalas, distinto número de grupos, etc. Dada la diversidad de modelos con características distintas que hemos aplicado y el hecho de que no existe un método de agrupamiento "correcto" o más efectivo en nuestro caso (ni en ninguno), si bien hemos considerado el agrupamiento aglomerativo como el de mejores resultados, somos conscientes de que cualquiera de los clusterings obtenidos en la Sección 4.9.1 puede dar información valiosa sobre el agrupamiento de nuestros datos de seguimiento ocular. Además, el estudio de este trabajo se concentra en hallar posibles agrupamientos de los datos desde cualquier punto de vista, por lo que en esta sección estudiaremos posibles combinaciones de los distintos agrupamientos obtenidos para nuestros datos de resolución del TMT-A.

Así, en lo que sigue consideraremos los diversos métodos de clustering que se han aplicado, y sus distintos criterios de agrupamiento, en un pie de igualdad y combinaremos sus resultados en un intento por estabilizar y robustecer los grupos obtenidos por cada uno de ellos de los datos de estudio, combinando las fortalezas de los algoritmos de agrupamiento individuales.

La idea subyacente es combinar diferentes técnicas para aunar los resultados de distintos métodos como un enfoque alternativo para mejorar la calidad de los algoritmos de agrupamiento. Esta idea se conoce como *ensamble* de agrupamientos y combina varias particiones generadas por diferentes algoritmos de clustering en un solo agrupamiento solución del ensamble (ver, por ejemplo, [12, 14, 15, 17, 41–43]). La partición combinada se obtiene como resultado de otro algoritmo de agrupamiento cuyas entradas son las etiquetas de los clústers de las particiones contribuyentes.

El problema del diseño de consensos de clústeres es más difícil que diseñar consensos de clasificadores ya que las etiquetas de los agrupamientos con que el método de clustering identifica al grupo, aunque sean numéricas, son de carácter simbólico y, por lo tanto, también hay que resolver un problema de correspondencia. Por ello, el consenso de agrupamientos de datos es un problema desafiante debido a la falta de etiquetas de clúster definidas globalmente.

Los métodos de ensamble de agrupamientos basados en el enfoque de la votación intentan establecer paralelos directos con las técnicas de consenso para múltiples clasificadores, buscando un reetiquetado de las particiones del conjunto [9, 11, 12, 44]. Estos métodos lidian con el problema de correspondencia de etiquetas para que una votación simple pueda ser utilizada para asignar objetos a distintos grupos y determinar la partición final de consenso. Sin embargo, la correspondencia de etiquetas es exactamente lo que dificulta la combinación en el aprendizaje sin supervisión pues, a diferencia de la clasificación supervisada, no existe una correspondencia explícita entre las etiquetas entregadas por diferentes particiones. Así, los métodos de ensamble no supervisado se realizan en dos etapas. En primer lugar se lleva a cabo un consistente reetiquetado óptimo de una partición dada con respecto a una partición de referencia fija y luego se aplica el voto mayoritario para determinar la membresía de cada registro a un grupo en la partición de consenso.

Una posible manera de encontrar soluciones de agrupación en clústeres coincidentes es mediante el conocido como método de *Kuhn-Munkres o método húngaro*, que resuelve la asignación de dos soluciones de agrupamiento diferentes entre sí. El algoritmo de coincidencia húngaro es un algoritmo para hallar coincidencias de peso máximo en gráfos bipartitos, problema que a veces se denomina problema de asignación. Como resultado de la aplicación de éste método, se reorganizan las etiquetas de un agrupamiento con vistas al ensamble de diferentes clusterings.

Proceso de ensamblado

Concretamente, a partir de los agrupamientos que resultaron de los distintos métodos aplicados que pueden verse en la Figura 4.16, realizamos un proceso de ensamble en dos etapas. La primera etapa consiste en un procedimiento de reetiquetado de los grupos de los distintos métodos utilizando el algoritmo de Munkres para grafos bipartitos, con el fin de obtener un etiquetado común en todos los agrupamientos que los haga comparables. El algoritmo de Munkres reetiqueta los registros en función

de un etiquetado de referencia, por lo que se procedió a aplicar el algoritmo de reetiquetado respecto a las etiquetas de cada uno de los métodos.

Una vez reetiquetados los resultados de todos los métodos, se obtuvo una tabla de agrupamientos correspondientes a los distintos algoritmos con etiquetas "comparables" entre sí. Luego, se llevó a cabo la segunda etapa del ensamble que consistió en la aplicación de la técnica conocida como "majority voting" para aprendizaje supervisado, para etiquetar cada registro (que ahora tiene esa etiqueta "común" obtenida del algoritmo de Munkres) con la etiqueta de mayor frecuencia de aparición en el conjunto de los distintos métodos.

De esta manera, luego de finalizado el proceso de ensamble, se tiene un único agrupamiento que "combina" todos los métodos (en algún sentido) y se espera más robusto y estable que cada clusterización particular.

Los resultados de los agrupamientos por ensamble que obtuvimos, habiendo realizado el análisis de las variables relevantes en cada agrupación, son los siguientes.

- Ensamble respecto al etiquetado con el agrupamiento aglomerativo como referencia: separa el dataset en solo dos grupos diferenciados muy claramente respecto a sus medias de velocidades medias puntuales en la ordenada y, y notoriamente respecto a la media de velocidades medias puntuales bidimensionales y la media en y.

- Ensamble con el método BIRCH como referencia: separa en tres grupos y diferencia uno de ellos respecto de los otros dos en media de velocidades medias puntuales en y y en xy, y velocidad total.

- Ensamble respecto a la etiqueta del combo de 3 métodos distintos como referencia: arma dos grupos en el total de sujetos, que se distinguen por sus valores en las variables media en y y media de velocidades medias puntuales en y.

- Ensamble respecto etiqueta de Mezcla Gaussiana con parámetro "Full": separa en dos grupos y los distingue en media y, media de velocidades medias puntuales en y, y media de velocidades medias puntuales bidimensionales.

- Ensamble con etiquetas respecto a las de la Mezcla Gaussiana con "Spherical": forma dos grupos y los separa notoriamente según las mismas variables que con parámetro Full.

- Ensamble respecto al consenso de tres KMeans: divide en dos grupos y las variables cuyos valores se distinguen en los dos grupos son media y, media de velocidades medias puntuales bidimensionales muy notoriamente y media de velocidades medias puntuales en xy.

- Ensamble con reetiquetado respecto a la etiqueta de KMeans: separa en 3 grupos y se distinguen las variables media de velocidades medias puntuales en y, media de velocidades medias puntuales en xy y velocidad total, de un grupo respecto de los otros dos.

- Ensamble con referencia a Mini Batch K-MEans: construye dos grupos y se distinguen en ellos los valores de las variables media en y, media de velocidades medias puntuales en y y media de velocidades medias puntuales bidimensionales.

- Ensamble respecto del método espectral: arma 3 grupos en los que un grupo se distingue de los otros dos en los valores de media de velocidades medias puntuales en y, media de velocidades medias puntuales bidimensionales y velocidad total (el mismo grupo respecto de las últimas dos variables).

Análisis de los ensambles

A la vista de los resultados de los ensambles y a partir de la observación de los boxplots de estos agrupamientos, podemos afirmar que los ensambles que clusterizan con diferencias unidimensionales notorias de las variables en los grupos son aquellos llevados a cabo reetiquetando con referencia al método aglomerativo con distancia euclídea y enlace Ward, y con referencia al algoritmo de Mezcla Gaussiana (con ambos parámetros "full" y "spherical"). Estos métodos dividen al dataset de estudio en dos grupos con diferencias claras en sus valores de las variables medias de velocidades medias puntuales en y (siempre la de diferencias más notorias), media de velocidades medias puntuales bidimensionales y valores de la media en la ordenada y. A modo ilustrativo, en la Figura 4.24 incluímos la visualización del gráfico de boxplots comparativo de los valores de las variables en el agrupamiento resultante de la aplicación del método de ensamble respecto al etiquetado del algortimo aglomerativo con distancia euclídea y enlace Ward.

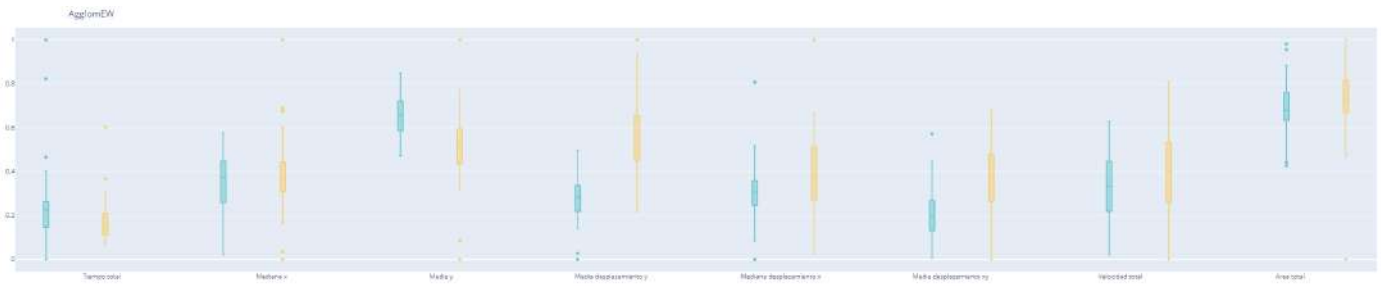


Figura 4.24: Distribución de valores de las variables en cada grupo del ensamble

Para tener una idea de los valores centrales de las ocho variables en las observaciones correspondientes a cada grupo de este ensamble respecto a las etiquetas del agrupamiento aglomerativo, puede observarase en la Figura 4.25 el gráfico de líneas de los valores medios y medianas de los dos clústers obtenidos en ese ensamble.

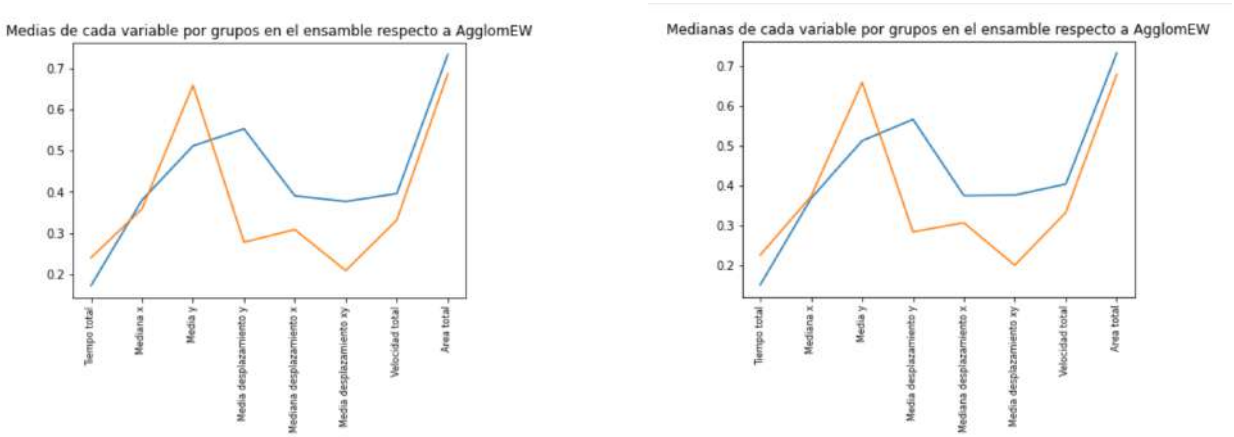


Figura 4.25: Valores centrales de las variables en cada grupo del ensamble

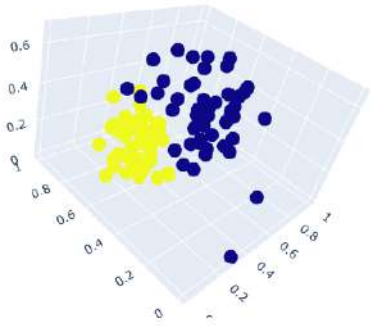


Figura 4.26: Proyección de grupos en variables relevantes del ensamble respecto al aglomerativo

Como visualización final del resultado del agrupamiento por ensamble en dos grupos, habiendo reetiquetado los agrupamientos con el clustering producido por el método aglomerativo, se muestra en la Figura 4.26 la proyección de las observaciones en el espacio de dimensión dos generado por las variables de mayor relevantes en el agrupamiento, donde puede observarse la división notoria de los valores de estas dos variables para los grupos.

Siguiendo nuestra línea de análisis del agrupamiento aglomerativo, en la Figura 4.27 reproducimos el árbol de decisión correspondiente al agrupamiento del ensamble con respecto al método aglomerativo donde se aprecian los valores que determinan la pertenencia de una observación a cada grupo.

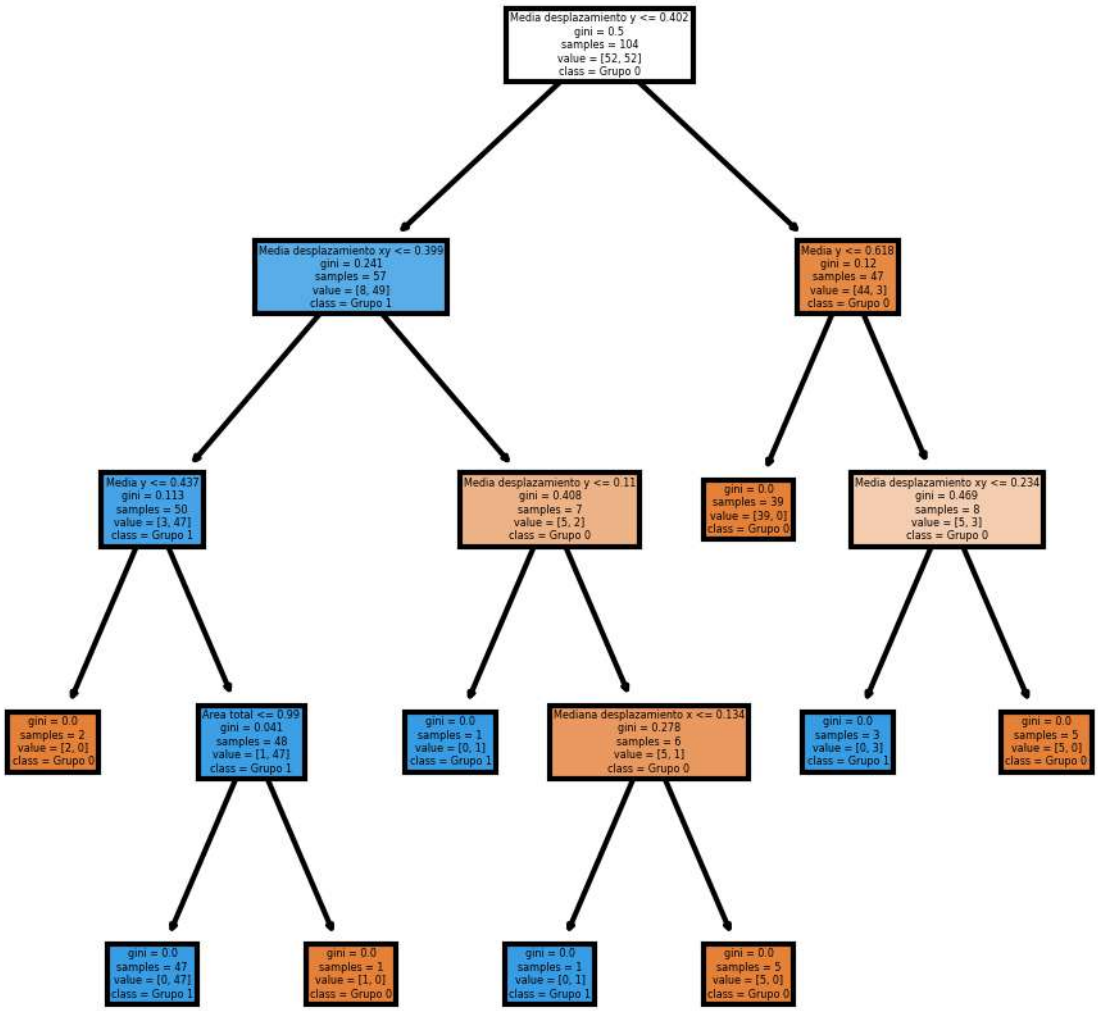


Figura 4.27: Árbol de decisión del agrupamiento por ensamble

Capítulo 5

Series temporales

Como ya hemos mencionado, este trabajo intenta comprender la estructura detrás de los datos de seguimiento ocular y su evolución a lo largo del tiempo. En particular, se trata de hallar agrupamientos de las observaciones que consideren toda la información recabada, tanto posicional como temporal de las mediciones de cada individuo. En otras palabras, deseamos desarrollar una herramienta que pueda discernir la estructura base de los datos a lo largo del tiempo, no solo su estructura en un momento particular. Así, desde el primer momento de este estudio de registros de posición en pantalla de la mirada de un centenar de sujetos durante la resolución del TMT-A, se entendió que considerar solamente la información de las mediciones de las variables x e y como atributos usuales de un dataset a clusterizar no resulta lo más deseable en este caso, puesto que de esta manera solo se consideraría la información espacial de la mirada y se perdería la crucial información temporal de las mediciones que liga los valores de las variables x e y en cada registro. De hecho, el tiempo es la única variable que se considera cuando se solicita la resolución del test con propósitos clínicos.

En esta línea de trabajo, en el capítulo anterior se llevó a cabo un estudio de agrupamiento de los datos a partir de variables derivadas de los registros surgidos del dispositivo de seguimiento ocular y que incluyen combinaciones razonables de la información espacial y temporal de las observaciones en el contexto del problema.

Sin perder de vista la importancia de identificar patrones de resolución del test como datos con evolución en el tiempo y agruparlos con esa idea, abordaremos en este capítulo otro enfoque de pensamiento de nuestros datos. Este diferente enfoque de análisis consiste en estudiar el agrupamiento de las trayectorias visuales pensando en la evolución temporal de cada coordenada espacial como una serie de tiempo, y agrupándolas en función de la similitud de patrones presentes en estas series temporales.

La determinación de grupos de series de tiempo es un problema extremadamente desafiante debido a la dificultad en la definición de similitud a través de diferentes series de tiempo que se pueden escalar y traducir de manera diferente tanto en la dimensión temporal como en la conductual. De hecho, es sabido que los métodos generales de agrupamiento, por ejemplo, varios de los que hemos utilizado en el capítulo anterior como K-Medias, no están diseñados para datos de series de tiempo y, por lo tanto, pueden no funcionar bien en este caso. Sin embargo, sí sabemos que el agrupamiento de datos se basa en medidas de distancia para determinar qué tan cerca están los datos entre sí y poder, a partir de esta noción de distancia, agruparlos en clústers distintos y homogéneos. La agrupación de datos de series de tiempo funciona de manera similar, pero necesitamos una medida de distancia que sea invariante de escala y desplazamiento, de modo que los datos de series temporales similares se agrupen independientemente de las diferencias triviales en amplitud, período, cambio de fase, etc.

Otra diferencia significativa entre la agrupación de datos de series temporales y la agrupación de datos en un espacio euclídeo es que la series de tiempo que se agruparán pueden no tener la misma longitud. En el caso en que todas las series de tiempo tengan la misma longitud, se pueden aplicar las técnicas de agrupamiento estándar representando cada serie de tiempo como un vector y usando una distancia norma Lp tradicional. Aunque, con tal enfoque, solo se puede explotar la similitud en el tiempo, mientras que la similitud en la forma y la similitud en el cambio temporal son ignorados. Justamente, nosotros estamos interesados en trabajar en la agrupación de datos de series de tiempo de

distinta longitud y basada en la similitud de patrón.

Existen en la literatura algunos trabajos de datos de seguimiento ocular estudiados con el paradigma de las series temporales. En particular, el grupo Neufisur que proveyó de los datos para este trabajo y con el que se discutió esta tesis, ha estudiado distintos aspectos de las señales resultantes de registros oculares de sujetos realizando distintas tareas cognitivas con herramientas de análisis de series temporales. En el artículo [5] se ha propuesto la representación y descomposición de señales de eye tracker en términos de wavelets construídas específicamente para abordar el modelado de movimientos oculares.

5.1. Series de coordenadas de posición

En este capítulo, comenzaremos el análisis de las series temporales con una pequeña exploración inicial de los datos para luego estudiar el agrupamiento de las series de coordenadas, con métodos básicos de clustering y combinados con las técnicas de ensamble, de manera similar al proceso utilizado en el estudio del agrupamiento de los datos de seguimiento ocular en función de las variables derivadas. Más precisamente, aplicaremos métodos de procesamiento y agrupamiento de datos a las señales correspondientes a las distintas coordenadas, abscisas y ordenadas, producidas por el registro de seguimiento ocular en cada tiempo de medición. Estas coordenadas se presentan como datos posicionales ordenados por la variable temporal de muestreo, lo que las concibe como series de tiempo por sí mismas. Analizaremos separadamente la serie temporal de las abscisas $x(t)$, la serie temporal de las ordenadas $y(t)$ y la serie temporal bidimensional de las coordenadas en el plano de la pantalla $(x(t), y(t))$.

5.1.1. Análisis exploratorio

En primer lugar, y para tener un panorama visual general de las características globales de las series temporales que se analizan, se graficaron los valores de las variables x e y como función del tiempo de medición asociado t para obtener una visualización de la evolución temporal de las coordenadas unidimensionales de las trayectorias de resolución de cada sujeto. La Figura 5.1 ilustra la idea mediante las representaciones gráficas de la variable $x(t)$ en el caso de cuatro individuos bajo estudio.

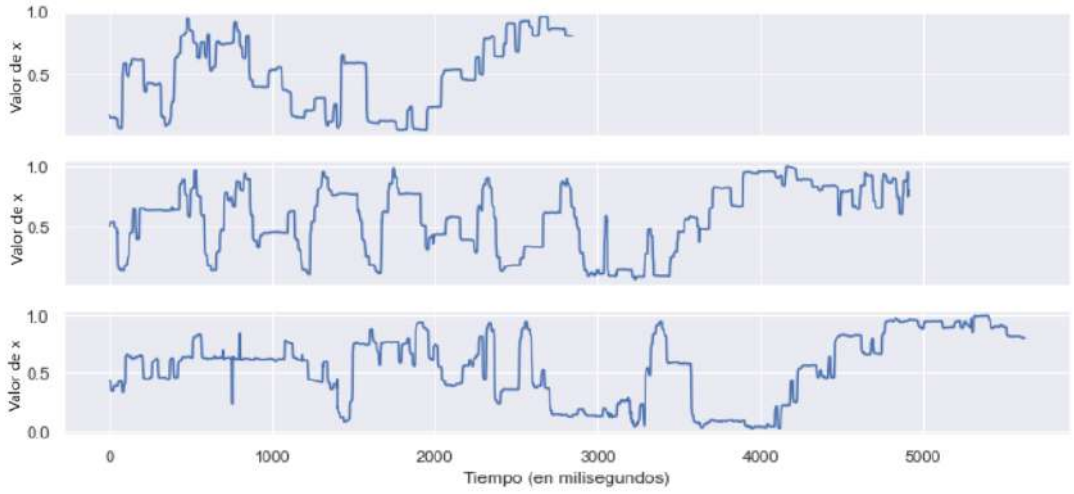


Figura 5.1: Representación gráfica de cuatro series temporales de abscisas

Aunque las series temporales de la Figura 5.1 corresponden solo a cuatro personas y a la variable de la abscisa, su representación nos permite elaborar una idea de cuán diferente es el comportamiento de las abscisas en función del tiempo para cada sujeto. En particular, en los gráficos de la figura se destaca la diferencia de tiempos de resolución del test por parte de los cuatro individuos, reflejada en la figura por la diferente longitud temporal de los dos primeros casos graficados respecto de los dos últimos.

Dadas estas diferencias de tiempos de resolución del test por parte de cada sujeto, cuyo detalle estadístico ya realizamos en la Sección 4.1, la cantidad de registros (x, y) de cada persona es muy

distinto (cada uno tiene un número de registros en relación a los milisegundos que haya tardado en la realización del TMT-A), razón por la que es difícil pensar en realizar una comparación directa entre individuos de los valores de las coordenadas espaciales y su variación en el tiempo.

Para obtener una representación visual de la serie temporal bidimensional (x, y) en función del tiempo, se realizó una visualización que muestra en formato de video el recorrido de la trayectoria completa en pantalla en función del tiempo durante la resolución del test por parte de cada individuo, mediante el movimiento de una pequeña bolita que transita cada posición $(x(t), y(t))$ de la trayectoria en pantalla en su tiempo de medición correspondiente, es decir, la bolita recorre la trayectoria de resolución a la velocidad de la mirada del sujeto. En la Figura 5.2 puede verse la visualización para un sujeto particular y un tiempo específico durante la resolución.

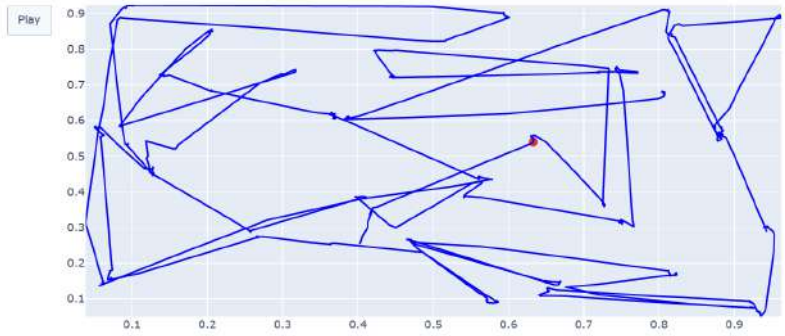


Figura 5.2: Foto del video de resolución en tiempo real de un sujeto

Cabe mencionar que en el presente enfoque de series temporales, la mayoría de las técnicas utilizadas en el preprocesamiento y análisis previo al agrupamiento que se llevó a cabo con las variables derivadas de los datos en el capítulo anterior, dejan de tener sentido. En particular, la reducción de dimensión es fútil en este contexto dado que las series temporales de las coordenadas espaciales que analizamos ahora tienen una dimensión, o a lo sumo dos dimensiones si consideramos la serie de posición en pantalla. Otra idea que es muy distinta en el caso de series de tiempo respecto de datos estándar, es el análisis de anomalías. Los datos atípicos en el marco de las series de tiempo son puntos anómalos dentro de la serie, y en este trabajo, no solo que no estamos en condiciones de decidir que un punto en particular de la mirada de una persona pueda llegar a ser un dato atípico como punto de la serie (excepto por el caso inexistente de un dato x o y que fuese mayor a 1), si no que también consideramos que para el análisis global de la caracterización de los individuos por grupos, unos pocos puntos atípicos en una serie temporal no harán una gran diferencia en el agrupamiento de los individuos respecto a su comportamiento general de búsqueda durante todo el test.

5.1.2. Agrupamiento

En una primera etapa, comenzamos realizando un proceso de clustering de las series temporales directamente surgidas del dispositivo de seguimiento ocular, es decir, de las series $x(t)$ e $y(t)$. Aquí nos encontramos con el primer problema en el agrupamiento que es el hecho de que las series de tiempo, ya sea el conjunto de series de abscisas $x(t)$ o el conjunto de todas las ordenadas $y(t)$, no tienen la misma longitud para todas las personas, su longitud en cada caso depende del tiempo de resolución de cada individuo. Lo primero que se hizo para poder obtener grupos en el conjunto de series coordenadas, en particular para aplicar el método K-Medias con la distancia euclídea, fue completar la serie con valor cero (proceso conocido como "zero padding") hasta el tiempo correspondiente al mayor tiempo insumido por algún sujeto para la resolución del test, es decir, se llevaron todas las longitudes de las series hasta la mayor longitud del conjunto de individuos. Si bien entendemos que este completamiento con valores nulos distorsiona bastante el comportamiento original de las series temporales de resolución, puede ocasionar que series de longitudes similares tengan buena chance de agruparse juntas dado que tendrán una cola de ceros de longitud similar.

Así, con este enfoque, se procedió a aplicar varias técnicas de clustering a las series temporales de abscisas y de ordenadas separadamente. El número de grupos que consideramos es 3, dado que es el número que conglomerados en que se agruparon los datos en el capítulo anterior y se tiene al intención de comparar los resultados obtenidos en los distintos enfoques.

Específicamente, los métodos de agrupamiento que se aplicaron a las series temporales son los siguientes: K-Means con métrica euclídea, KShape, jerárquico aglomerativo, BIRCH, "MeanShift" (que solo arma 2 grupos), "Affinity propagation" (construye más de 3 grupos), OPTICS y SoftDTW. Se experimentó también con la aplicación de los métodos DBSCAN y "Kernel K-Means" pero ninguno de los dos arrojó agrupamiento para los series de abscisas y ordenadas. Por otro lado, se utilizó con métrica euclídea el método de "Dynamic Time Warping (DTW)" pero no se logró llegar a una clusterización de los datos por esta vía.

Por completitud y a título ilustrativo, se muestra en la Figura 5.3 el encabezado del dataset de los ocho agrupamientos que arrojaron resultados para las series temporales de las coordenadas x de la posición de la mirada.

	KShape	KMeans	Agglomerative	BIRCH	MeanShift	Affinity	OPTICS	SoftDTW
0	0	1	0	1	2	0	-1	2
1	0	1	0	1	1	0	-1	2
2	0	1	0	1	1	1	-1	2
3	0	1	0	1	1	1	-1	2
4	0	1	0	1	1	1	-1	2

Figura 5.3: Encabezado de la tabla con los resultados de los agrupamientos de abscisas

Visualizaciones de los agrupamientos

Para poder realizar una inspección ocular de las características de los grupos que se formaron con los distintos algoritmos de clustering aplicados a las abscisas y a las ordenadas del movimiento ocular en pantalla durante la resolución del test, se confeccionaron los gráficos de las series temporales de coordenadas x e y de cada grupo conjuntamente con la correspondiente "serie mediana" de cada grupo, definida como la serie que en cada tiempo de medición tiene la posición del valor mediana de todas las series del grupo. Dos de estos gráficos con los tres grupos diferenciados que se obtuvieron en el caso de las abscisas, el primero para el método de K-Medias con distancia euclídea y el segundo asociado al método KShape, son los que muestran las Figuras 5.4 y 5.5.

Dado que en el agrupamiento de series de tiempo no se cuenta con variables de referencia que puedan indicarnos caracterizaciones de los grupos en relación a ellas, lo que hicimos en este enfoque para intentar interpretar el significado del clustering fue analizar los boxplots de los agrupamientos como series de tiempo obtenidos de las abscisas respecto de las catorce variables derivadas con las que se hizo el estudio en el Capítulo 4. El resultado de este análisis arrojó que, en general, no se observan diferencias respecto de los valores de estas variables entre los grupos de series temporales de coordenadas horizontales. El único agrupamiento en el que se observa alguna pequeña diferencia corresponde al método "Mean Shift" que distingue un grupo respecto de los otros dos en los valores de la variable "media x" y dos grupos respecto del tercero en la variable "mediana x".

Al analizar los grupos de las series temporales de ordenadas en relación a las variables derivadas, tampoco se observan diferencias en los distintos clusters en relación a estas variables. La única mención que puede hacerse es respecto al agrupamiento por ensamble con etiqueta de referencia correspondiente al método "Affinity propagation". Este método construye cuatro grupos y diferencia uno de ellos respecto de los otros tres en los valores de la mediana de velocidades medias puntuales bidimensionales. Por otro lado, el método OPTICS muestra una ligera diferencia entre dos grupos respecto del tercero de las variables "tiempo total" y "media x".

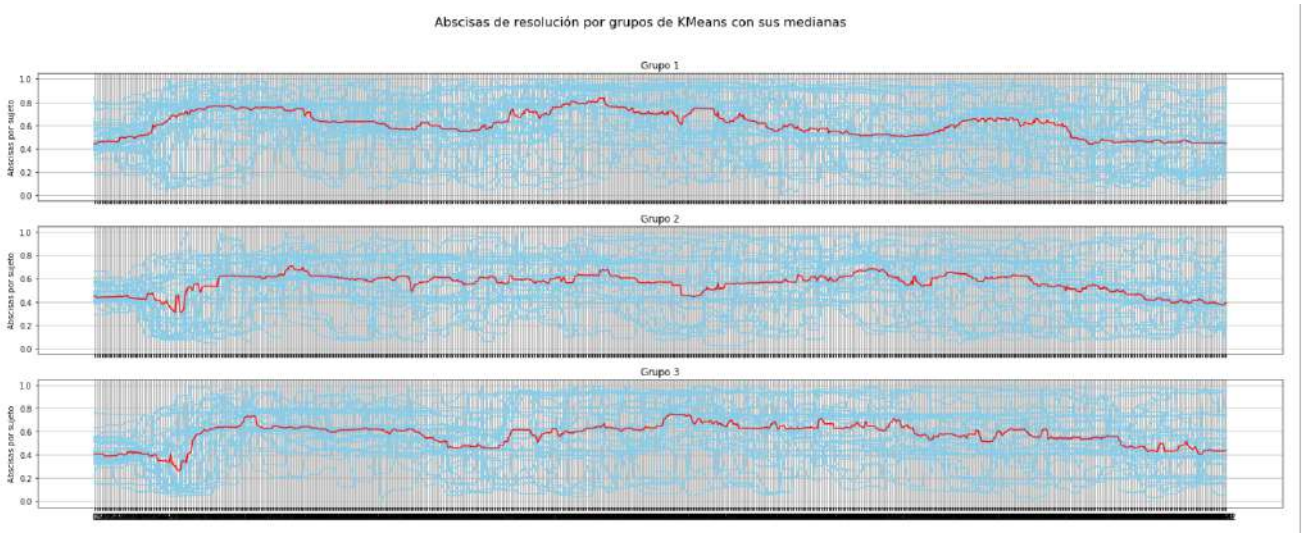


Figura 5.4: Series y centroides en cada grupo de abscisas

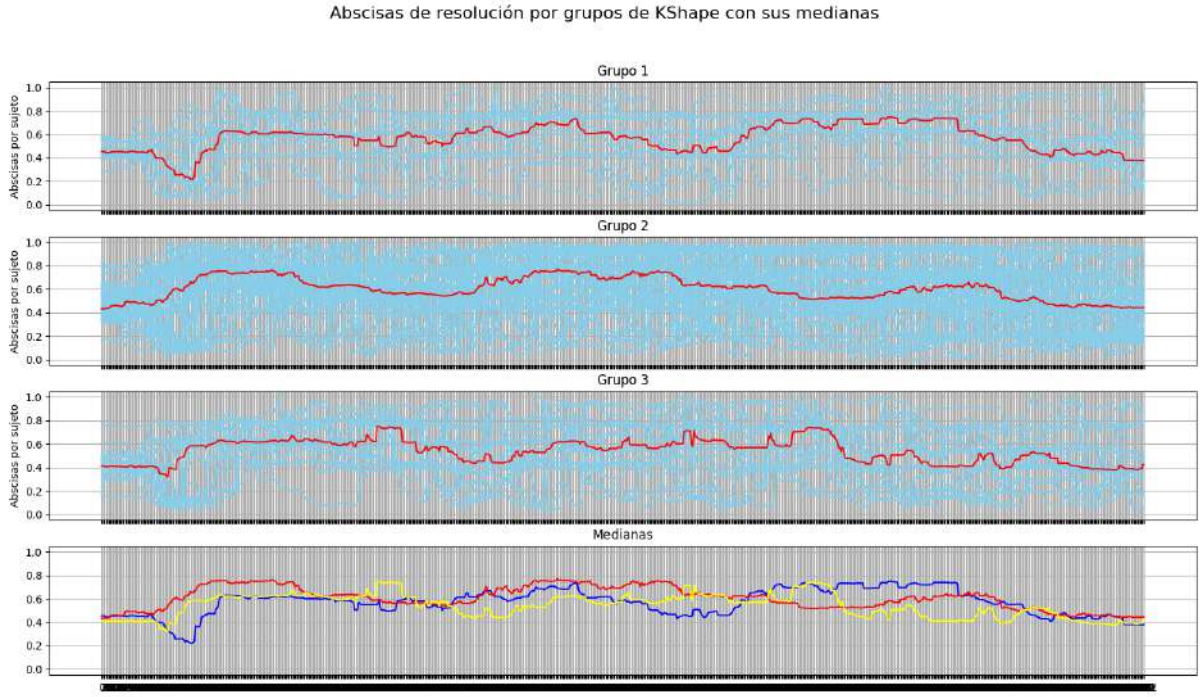


Figura 5.5: Series y medianas de cada grupo de abscisas con KShape

Es resumen, los grupos formados según valores de las series temporales de abscisas y ordenadas, no se distinguen claramente por sus valores respecto de las variables derivadas analizadas en el Capítulo 4. Lo mismo sucede al realizar el proceso de ensamble análogo al realizado en el capítulo anterior. Luego de proceder a ensamblar los resultados de los agrupamientos en función de las etiquetas de cada uno de ellos, los pocos ensambles que arrojaron resultados (clusterizaron en más de un grupo) no muestran diferencias notorias en sus grupos respecto a los valores de las catorce variables consideradas. Tampoco se ven diferencias de valores de esas variables derivadas al ensamblar los agrupamientos de abscisas y ordenadas en forma conjunta, que realizamos en un intento por combinar la información de las abscisas y de las ordenadas pensando, en algún sentido, en un agrupamiento bidimensional que es esencialmente lo que nos interesa obtener.

Análisis de agrupamientos

Para observar una comparación gráfica de los valores de las medianas de las abscisas de cada cluster, en la Figura 5.6 se muestran las medianas de las coordenadas $x(t)$ de cada uno de los grupos obtenidos por el método KShape.

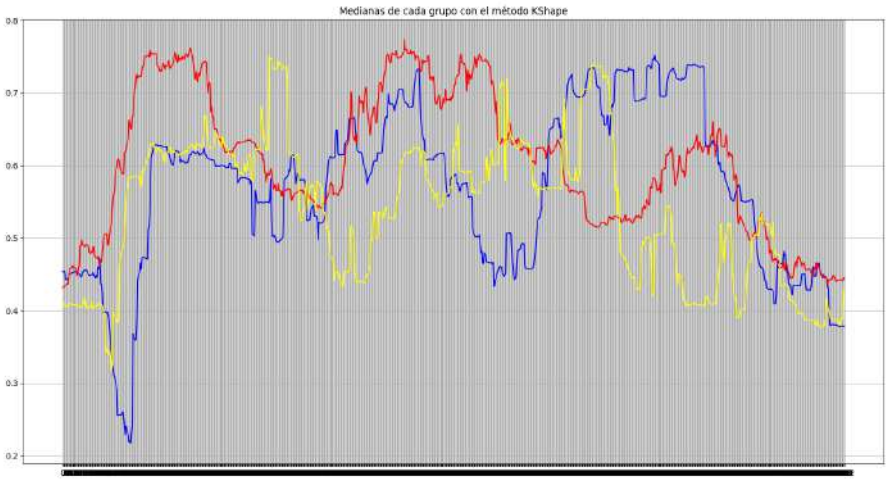


Figura 5.6: Series medianas de cada grupo de abscisas

Sobre agrupamiento de series de fijaciones

Dados los datos del seguimiento ocular, la trayectoria visual durante el proceso de hallar los números del TMT-A puede describirse naturalmente como una secuencia de fijaciones de la vista en la pantalla o incluso de fijaciones en los números del test. Esta secuencia puede construirse como sigue. Al comienzo de la prueba se le indica al individuo la ubicación del número 1 en la pantalla, por lo que su mirada se dirige a la posición de ese número. En esa posición permanece la mirada del sujeto unos milisegundos mientras toma la decisión de dirigir su mirada a otro de los números del test aleatoriamente elegido. En ese momento, comienza el proceso de búsqueda que consiste en la realización de movimientos sacádicos desde la posición de un número a la de otro número en pantalla hasta terminar la secuencia de hallar los 25 números y, por consiguiente, finalizar el test.

Teniendo este proceso en mente, y a los efectos de detectar estrategias de resolución del test que es nuestro objetivo, aparece como interesante el estudio de las posiciones de las fijaciones de la vista durante la resolución del TMT-A. Por esta razón, se pensó en considerar para el análisis de agrupamiento las series temporales determinadas por las secuencias de las posiciones de las fijaciones en pantalla e incluso las series de fijaciones solo sobre números del TMT-A, pero esta idea de serie de tiempos ocasiona una problemática extra que es el hecho de que la división temporal de esta posible serie no es uniforme como la registrada por el dispositivo de seguimiento ocular, sino que está dada por los intervalos de tiempo entre fijaciones de cada sujeto. Por este motivo, la idea de considerar el análisis de agrupamiento de las series de fijaciones en este capítulo del trabajo de tesis fue desestimada.

5.2. Series adaptadas a la resolución del test

En el proceso de clustering de un conjunto de datos, para realizar el análisis de los distintos resultados y obtener un agrupamiento "exitoso" que pueda ser interpretado en función del problema, es crucial recurrir a la información del dominio de los datos. A partir de esa información que, en nuestro caso, es el contexto de seguimiento ocular durante el test TMT-A, en esta sección se construirán series temporales diseñadas específicamente para los datos de estudio y se procederá a su agrupamiento.

En la sección anterior se analizó la formación de grupos de las series de tiempos dadas por cada registro del dispositivo de seguimiento ocular con el tiempo de medición como variable independiente,

y las coordenadas espaciales de la posición de la mirada en la pantalla en ese instante como variable dependiente.

Con el convencimiento de que el análisis de agrupamiento de los datos de seguimiento ocular como series temporales es apropiado y conducente a obtener información valiosa, aún cuando los intentos por clusterizar las series de las coordenadas de la posición de la mirada no han sido todo lo fructíferos que se esperaba, se repensaron las definiciones de las series temporales construídas a partir de los datos utilizando fuertemente el hecho de que los datos de estudio surgen del seguimiento ocular de individuos realizando la prueba conocida como TMT-A y el conocimiento que se tiene de la forma en que una persona resuelve el test.

En esta sección, el enfoque sigue siendo de agrupamiento de los datos como series temporales pero construyendo a partir de los datos series en las que la variable independiente no es el tiempo (y es "discreta") y la variable dependiente no es una posición. Con el objeto de llevar a cabo un análisis de agrupamiento de la evolución durante la resolución del test, en primer lugar se estudiará el agrupamiento de las resoluciones de los sujetos de acuerdo a sus tiempos y velocidades en cada uno de los 25 trials del test (es decir, en cada uno los intervalos entre que se localiza un número del test y el siguiente) y, en segundo lugar, de acuerdo a su desempeño en ciertas zonas que se conocen distinguidas en la resolución del TMT-A y que se detallarán más adelante.

Cabe mencionar, que es razonable ocuparse paralelamente del análisis en detalle de cada registro dado por el dispositivo de seguimiento ocular y del análisis más global de la resolución del TMT-A de cada individuo, por ejemplo analizando tiempos o velocidades por trial, dado que son análisis en contextos diferentes y con requerimiento de recursos tecnológicos muy distintos. El análisis de todas las mediciones del dispositivo (en la línea de la sección anterior de series temporales) puede ser muy beneficioso en caso de que en un consultorio psiconeurológico se disponga de un eye tracker, pero será muy poco útil como herramienta clínica para un profesional sin acceso a un dispositivo de medición electrónica. En cambio, un análisis de los tiempos que demora un individuo en hallar cada número del test a partir del número anterior, no necesita de equipos de alta tecnología para la recolección de los datos.

5.2.1. Series de trials

Se construyeron las series de los tiempos que cada sujeto empleó en detectar cada uno de los números a partir del hallazgo del número anterior de la secuencia, es decir, la serie de tiempos por trial; y la serie de velocidades registradas en cada trial que se obtuvo como el inverso del tiempo de resolución del trial, considerando que la distancia entre números en el test es de una unidad de avance en la resolución y se la divide por el tiempo insumido en ese avance para obtener la velocidad. Así, las series temporales consideradas, tanto de tiempos como de velocidades, tienen una longitud de 25 registros para todos los sujetos, siendo el número de trial la variable "independiente" y el tiempo o velocidad en hallar ese número en el test, la variable "dependiente".

Considerando que una serie de tiempo técnicamente es una secuencia ordenada de valores de una variable a intervalos igualmente espaciados de tiempo, el enfoque propuesto en el planteo de este capítulo usando series temporales es considerar como "tiempo" discreto el dado por la secuencia de números a recorrer en el test y el correspondiente valor del tiempo insumido en hallar ese número desde el anterior como dato temporal de la serie.

Un motivo que refuerza la idea de considerar a estas secuencias de tiempos (y velocidades) por trial como series temporales es el hecho de que, en algún sentido, el tiempo insumido en cada trial tiene una dependencia del propio trial (ubicación de ese número en particular dentro del test) y del área de pantalla escaneada hasta ese momento (diseño del test y tiempo invertido de inspección de la pantalla hasta ese momento), es decir, una dependendencia de los trials anteriores de la secuencia.

Por otro lado, recordemos que las variables tiempo y velocidad de resolución del test son las únicas variables que consideran los profesionales de la salud que realizan el test TMT-A a sus pacientes.

A priori, las series de tiempos y las series de velocidades definidas como el inverso del tiempo parecen ser esencialmente equivalentes pero en este trabajo se realizan estudios de agrupamiento de ambas series

con una imputación distinta de valores faltantes en cada caso. Un objetivo secundario de este análisis es la detección de posibles diferencias en el análisis de clustering de estos datos de seguimiento ocular causadas por los efectos de los datos faltantes.

Otra decisión que se tomó para realizar el análisis de series por trials fue la de seleccionar un subconjunto de los heterogéneos 108 datasets con una mejor calidad de datos que el conjunto completo y mayor homogeneidad. De los 108 individuos se escogieron 49 que han sido estudiados por el grupo Neufisur con otras técnicas, todos mayores de edad con una media de edad de 23 años (estudiantes de nivel terciario o universitario), sin patologías visuales conocidas y sin datos faltantes de medición.

De esta manera, para el estudio de series de tiempo por trial se tiene un dataset de 49 series temporales, uno por cada sujeto cuya resolución del TMT-A está en estudio, con 25 registros unidimensionales cada una: los 25 tiempos insumidos para hallar el número $n + 1$ a partir del número n en la pantalla del test, para n con valores de 0 a 24. Es decir, se tienen 49 series temporales de 25 registros cada una, uno por cada trial.

Valores faltantes

Dado que hay errores en la resolución del test por parte de algunos individuos que no localizan correctamente la secuencia de números en la pantalla, algunos registros de la base de datos de tiempos en cada trial (determinados por la localización de los números del test) están incompletos. El conocimiento de cuáles son los trials que tienen un mayor número de valores faltantes y qué sujetos tienen esos registros faltantes es información valiosa sobre la resolución del test por parte de este conjunto de personas, por lo que realizamos un análisis de esos datos faltantes que detallamos en esta subsección.

En la Figura 5.7 puede apreciarse un gráfico de barras que muestra la cantidad de datos de tiempos de resolución de cada trial con que se cuenta. En ese gráfico podemos observar que los trials con registros temporales faltantes son los correspondientes a la búsqueda de los números 7 al 10, 13 al 19, y 23 en el test (trials 6 al 9, 12 al 18 y 22). El trial con más valores faltantes es el de los tiempos insumidos para hallar el número 14 (trial 13), con 9 registros faltantes de los 49 totales. Esta información indica que el hallazgo del número 14 en la pantalla del test podría tener una dificultad mayor que el del resto de los números para los individuos cuyos datos de seguimiento ocular están siendo estudiados. Una explicación posible de este hecho es que el número 15 se encuentra en el extremo superior izquierdo de la pantalla y "aislado" en la distribucion gráfica de los números en el test, por lo que posiblemente no se ha explorado esa zona con anterioridad en la búsqueda de números dada la baja densidad de dígitos en ese área El siguiente trial con más registros faltantes es el de la búsqueda del número 9 con 6 registros vacíos.

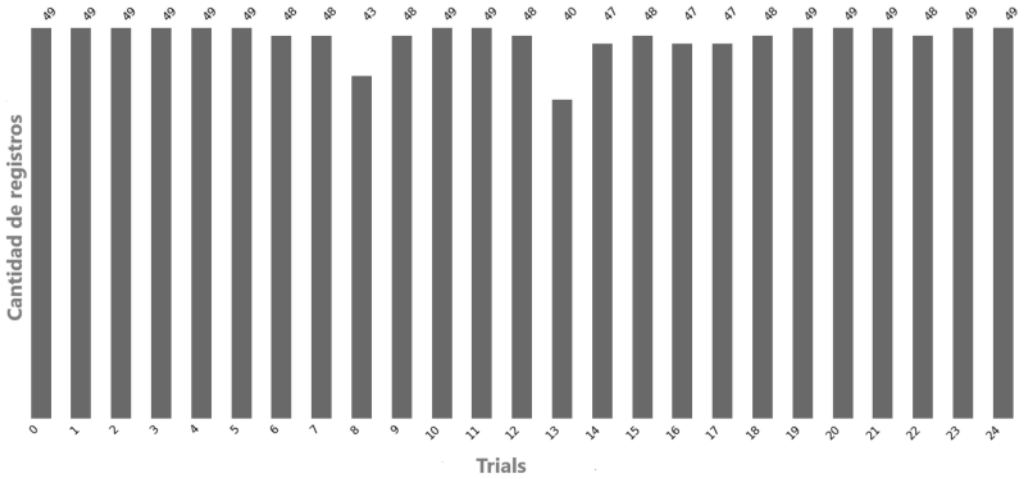


Figura 5.7: Número de registros por trial

Analizando ahora los datos faltantes desde el punto de vista de los individuos, vemos que existen sujetos que no completaron correctamente la resolución del TMT-A, es decir, que no lograron hallar

algunos de los 25 números. En este sentido, comprobamos que existen tres individuos que no lograron resolver tres de los trials de la prueba, estos son los sujetos 9, 11 y 32, y un único sujeto numerado con el 18 que no resolvió cuatro de ellos. Esta información puede visualizarse en el gráfico de la Figura 5.8, donde se representan en un mapa de calor los valores faltantes de los datos en cada individuo en función de los 25 trials.

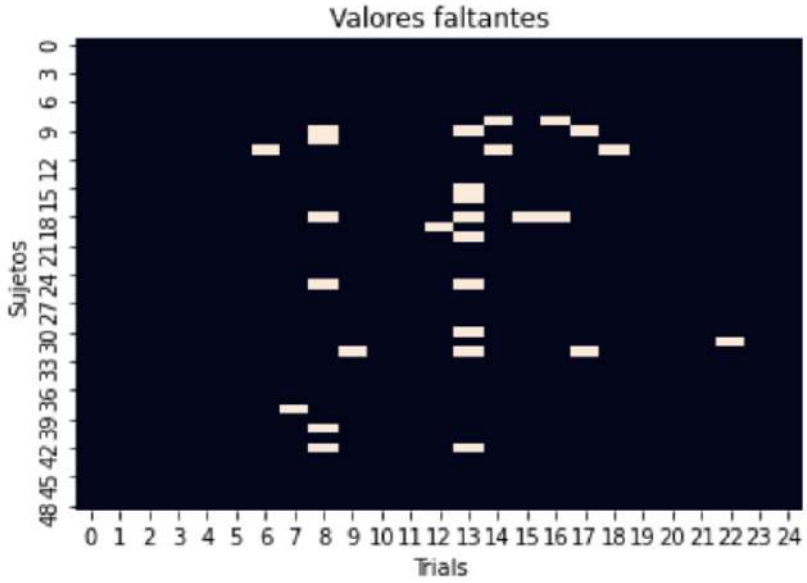


Figura 5.8: Valores faltantes por sujeto y trial

Para continuar con el análisis de agrupamiento de las series de tiempos por trials que tienen valores faltantes, se decide realizar una imputación de éstos valores con el de la mediana del tiempo insumido por todos los sujetos del dataset en el trial correspondiente al dato faltante.

En el caso de la serie de "velocidades" de resolución en cada trial, se imputaron los valores faltantes reflexionando sobre el significado de un valor faltante en un trial por parte de un individuo, que indica que el sujeto no encontró ese número en la pantalla del test, y la idea de velocidad, que asigna a cada trial la distancia discreta entre números consecutivos de la secuencia del test (de valor constante 1) sobre el tiempo insumido para encontrar el siguiente número del TMT-A. Así, se imputaron los valores faltantes de velocidades con el valor 0, reflejando de manera más explícita el hecho de que un valor faltante indica que ese número del test no fue hallado por el sujeto, es decir, que se tardó un tiempo "infinito" en resolver ese trial.

Exploración inicial

Luego de la imputación de valores faltantes con la mediana de todos los tiempos de resolución de ese trial y a modo de primer análisis exploratorio, se obtuvieron los gráficos de boxplot con los valores estadísticos de los tiempos de resolución de cada trial por parte de los sujetos de estudio, este gráfico se muestra en la Figura 5.9. Allí se observa que el tiempo medio insumido por los sujetos en la resolución de los trials es mayor en el tercer trial, donde también se aprecia una alta variabilidad en los tiempos de resolución de los sujetos analizados. Además, al observar en la Figura 5.9 la estadística de los valores de los tiempos de resolución de la muestra de sujetos, a rasgos generales, se aprecian los mayores tiempos de resolución en los seis primeros trials y luego una disminución de los tiempos insumidos para localizar los números 7 al 10 (trials 6 al 9 en la figura), seguidos por los tiempos asociados a la búsqueda de los números 8, 11, 12 y 13 (trials 7, y 10 al 12 en la Figura 5.9). También se ve una mayor dispersión en los valores de tiempos de resolución de los distintos sujetos en los seis primeros trials, es decir, el desempeño de los sujetos en los seis primeros trials del test es más variable que en el resto de los trials.

Figura 5.9: Estadística de tiempos insumidos por trial

De todos modos, las diferencias en la media de los valores temporales correspondientes a cada trial son bastante pequeñas, dado que el tiempo de resolución es de unas pocas decenas de segundos y se tienen 25 trials en el test. Sin embargo, se aprecian diferencias en la variación grupal de tiempos en cada trial por parte de los sujetos de estudio, con algunos trials de notable poca variación como el 13 y 21 de la figura, y otros de alta variabilidad como el segundo, tercero y sexto. Esto puede verse claramente en la gráfica de la figura 5.9, donde se visualiza la variabilidad al comienzo del test y la homogeneidad de la última parte.

Al analizar los valores de las medianas de los tiempos de resolución en cada trial en el gráfico de barras de la Figura 5.10, se puede distinguir una etapa de exploración inicial del test, que puede identificarse con los seis primeros trials y requiere los mayores tiempos de resolución. Si bien existen valores no tan altos de tiempos insumidos en la resolución del primer trial, que pueden parecer poco naturales comparados con el segundo y tercer trial, pueden corresponder al hecho de que a una parte de los sujetos del grupo analizado se les indicó la posición del primer número en la pantalla al iniciar el test.

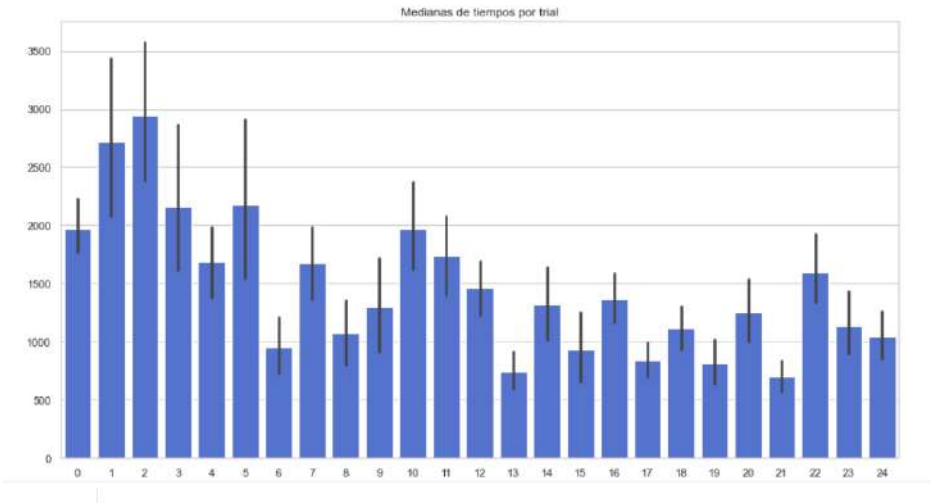


Figura 5.10: Mediana de los tiempos de resolución por trial

Más allá de la primera etapa del test con altos tiempos de resolución, el gráfico de barras de la Figura 5.10 sugiere una cierta tendencia natural a que los tiempos de resolución de los trials vayan

disminuyendo a medida que se va resolviendo el test, ya que al avanzar el tiempo de resolución se ha escaneado una mayor parte de la pantalla y consecuentemente se ha observado la ubicación de los números, tendencia que se ve alterada por el aumento del tiempo de resolución en los trials 11 a 13 y 23 (barras 10 a 12 y 22 en la Figura 5.10). Analizando esta situación, luego de transitada la etapa exploratoria inicial de altos tiempos, se observa un descenso en los valores de los tiempos insumidos para localizar los números 7 al 10 (trials 6 al 9 en la figura) que puede explicarse observando la posición de estos números en el test, dado que los números 6 al 8 están prácticamente en línea recta, muy cercanos entre sí, y los números 9 y 10 incluso más cercanos aún casi "de camino" entre el 7 y el 8. Después de estos trials, el espiral de recorrido del test cambia el sentido y los sujetos deben cambiar el sentido de la búsqueda que llevan, lo que implica una mayor dificultad de resolución que se observa en el aumento de los tiempos de resolución en los trials 10 a 12 en la figura (búsqueda de los números 11 al 13 en el test), y luego los tiempos vuelven a descender, incluso bruscamente en el trial 13 ya que el número 14 se encuentra muy cercano al número 13 en el test. A partir del trial 13, el test se vuelve bastante determinista dado que para este momento se ha explorado una gran parte de la pantalla, excepto por la resolución de algunos trials asociados al hallazgo de ciertos números que acarrean alguna dificultad por estar lejos del anterior o "tapados" por otros números en el recorrido visual, como lo son el número 17, el 21 y el 23.

Pasando a las series de velocidades, en la Figura 5.11 se exhibe el gráfico de boxplot de las velocidades por trial donde se detectan los tramos de mayor velocidad y se distinguen las búsquedas de resolución lenta en este conjunto de individuos.

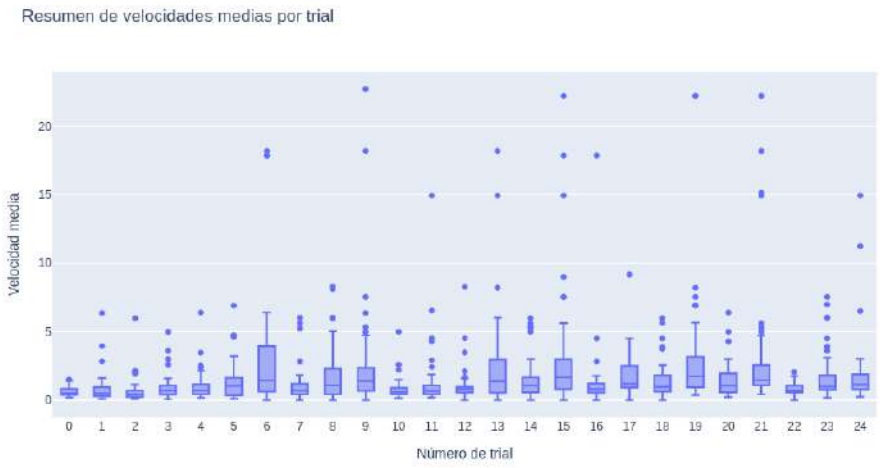


Figura 5.11: Estadística de velocidades de resolución por trial

A grandes rasgos, también se observa una mayor variabilidad entre sujetos en los trials de velocidades de resolución más altas, es decir, en las búsquedas que tienen mayores valores de mediana de velocidades.

Lógicamente, se detectan en general comportamientos inversos a los observados para los tiempos de resolución como consecuencia de la forma en que se relacionan estas variables, ya que se ha obtenido la velocidad de cada trial como el inverso del tiempo de resolución de ese trial. La única diferencia que podría apreciarse (y para eso el análisis de ambas variables) es el efecto de las distintas imputaciones de valores faltantes que se ha realizado en cada caso.

Correlación entre trials

Con el fin de detectar posibles relaciones entre los tiempos de resolución de los sujetos en pares o grupos de trials, por ejemplo, que un tiempo largo al resolver un cierto trial es coocurrente con otro tiempo largo (o corto) al resolver otro trial, se construyó la matriz de correlación entre los tiempos de los distintos trials de las 49 personas analizadas. La matriz obtenida puede verse en la Figura 5.12, donde no se observan valores de alta correlación entre los tiempos de los distintos trials. En algún sentido,

podría afirmarse que no existe autocorrelación significativa entre las series temporales de tiempos y, por ende, entre las de velocidades de resolución por trial para este conjunto de individuos.

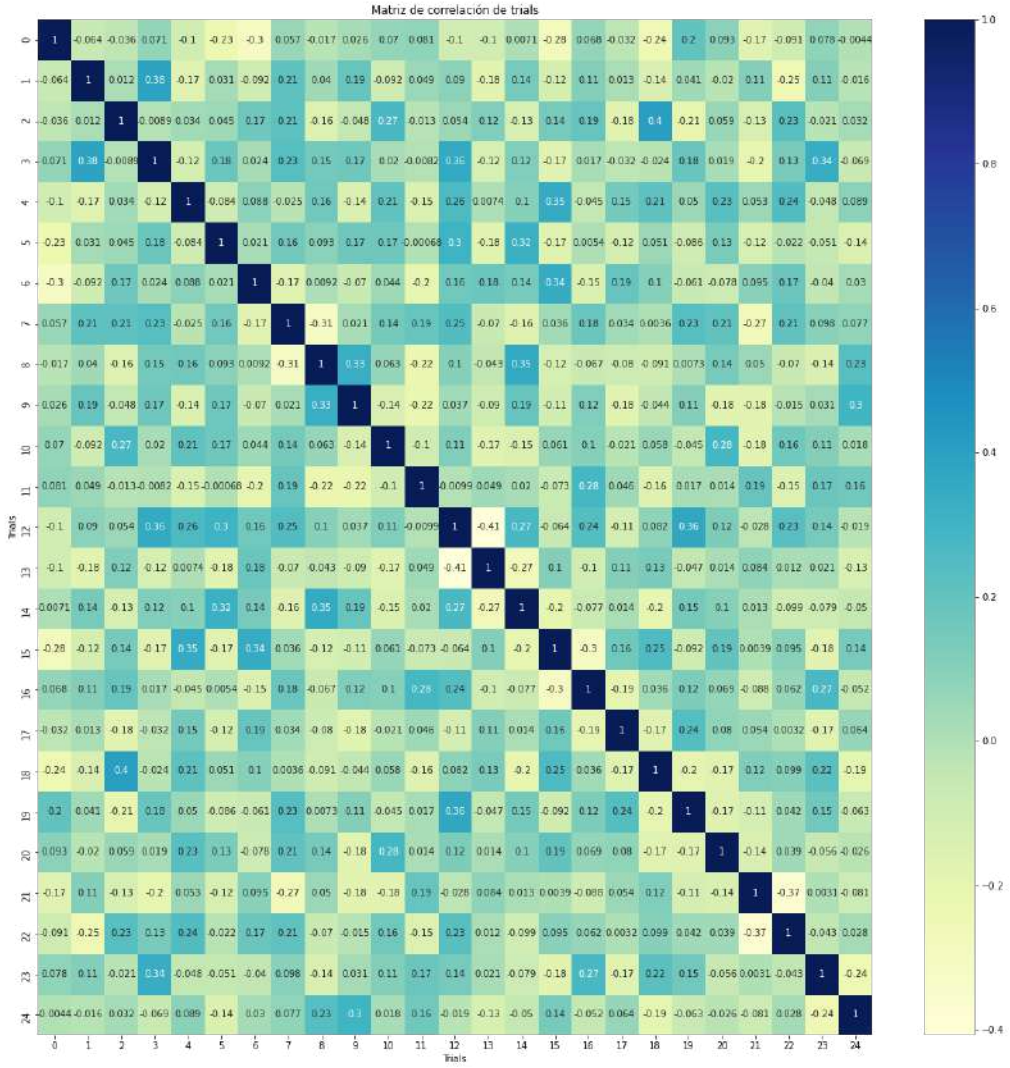


Figura 5.12: Matriz de correlación entre trials

Agrupamiento de tiempos por trial

Luego del análisis inicial realizado, se procede finalmente a llevar a cabo el estudio del agrupamiento de las series de tiempos de resolución por trial. En una primera etapa se utilizó el método de la inercia o "Elbow method" para obtener el número óptimo de grupos que pueden constituirse con el conjunto de series temporales que se tiene. El método aplicado sugiere que el número de grupos óptimo para congregar a las series de tiempos de resolución por trial del conjunto de sujetos bajo estudio es de 4 grupos.

Así, a las series de los tiempos de resolución por trial se le aplicó el método de agrupamiento K-Medias para cuatro conglomerados con las distancias euclídea y "Dynamic Time Warping"y, en el caso de la distancia DTW, ésta se aplicó también con el algoritmo conocido como SoftDTW. Obtenidos estos resultados, pueden observarse algunas características notables en el caso del agrupamiento con el método de K-Means considerando la distancia euclídea para las series de tiempos de resolución en cuatro grupos. En lo que sigue, se mencionan las particularidades que se detectan en este agrupamiento de los 49 individuos.

Figura 5.13: Boxplots de agrupamiento de tiempos de resolución por trial

En la Figura 5.13 puede verse el gráfico de boxplots del agrupamiento por K-Medias en 4 grupos con distancia euclídea de las series de tiempos de resolución por trial. De esta visualización, no puede concluirse que el método "Time Series K-Means" agrupe las resoluciones en cuatro grupos que separen claramente los valores de los tiempos insumidos en la resolución de ciertos trials.

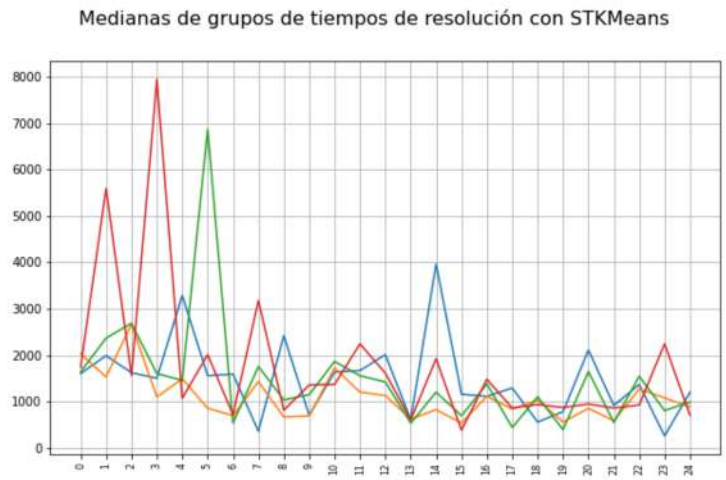


Figura 5.14: Centroides por grupo de las series de tiempos de resolución

Sin embargo, observando el gráfico de centroides de tiempos de los cuatro grupos en la Figura

5.14, se distinguen notorias diferencias entre los valores de las medianas de los tiempos de resolución, especialmente en algunos grupos en el caso de los trials segundo, cuarto, sexto y décimo quinto; lo que puede observarse más claramente en la Figura 5.15.

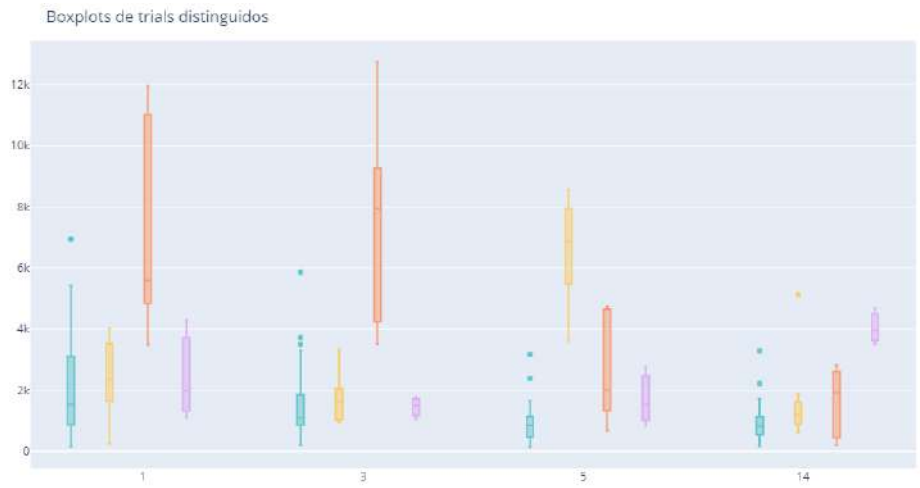


Figura 5.15: Boxplots del agrupamiento Time Series K-Means en los trials con grupos distinguibles

En la Figura 5.16 pueden visualizarse los valores de tiempos de resolución por trial de todos los sujetos en cada grupo y, a la vista de ese gráfico, los comentarios que podemos hacer son los siguientes.

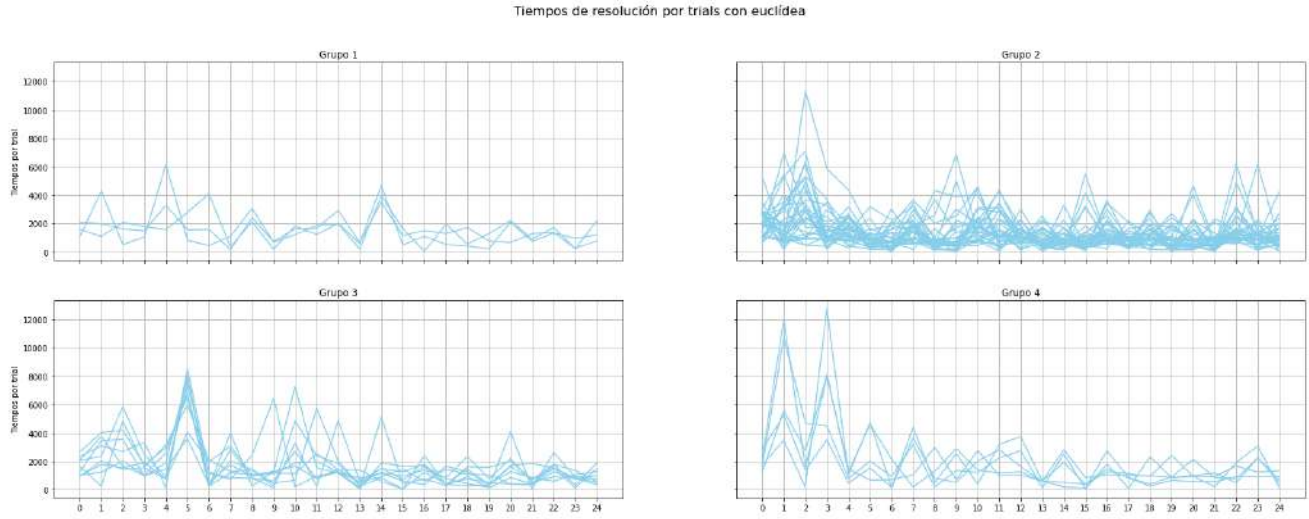


Figura 5.16: Visualización de todos los sujetos de cada grupo por tiempos de resolución

Bajo una mirada global del agrupamiento y en líneas generales, en él se reconoce un grupo (el Grupo 1 en la figura 5.16 identificado con el color violeta en el gráfico de boxplots 5.15) formado por 3 individuos caracterizados principalmente por haber empleado un tiempo considerablemente mayor en la resolución del trial décimo quinto respecto de los otros sujetos; un segundo grupo (el cuarto grupo de 5.16 con boxplots color naranja en la figura 5.15) constituído por 5 sujetos destacados por sus altos valores de tiempos de resolución en el segundo y cuarto trial; un tercer grupo "amarillo" formado por 9 sujetos caracterizados por haber dedicado mucho tiempo al sexto trial, es decir, tardaron más que la mayoría de los individuos buscando el número 6 en el test; y un cuarto grupo de 32 sujetos (el celeste en el gráfico de boxplots 5.15) cuyos valores de tiempos de resolución no destacan del conjunto de individuos en ninguno de los 25 trials.

Resulta de interés la conformación del tercer grupo, en color naranja en el gráfico de la figura 5.16.

Este grupo se caracteriza por muy altos valores de tiempo empleado para resolver el sexto trial, muy diferentes del resto de los individuos analizados. Esta diferenciación tan marcada en un trial específico nos sugiere una evidente diferencia de este grupo de sujetos respecto al resto en la búsqueda del número 6 en el test TMT-A. Este sexto trial presenta la particularidad de que para encontrar el número objetivo se debe inspeccionar el interior de la figura espiralada que se ha ido recorriendo para resolver los primeros cuatro trials (para hallar los primeros cinco números), incluso volviendo a observar una región que ya se ha explorado previamente. Los individuos para resolver el trial 5 deben "retroceder" en la dirección que llevan en el recorrido de sus miradas habiendo resuelto los primeros cuatro trials del test, como muestra la figura 5.17 del test.

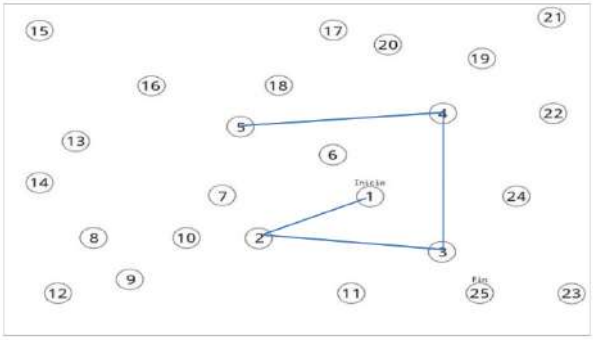


Figura 5.17: Imagen del test TMT-A hasta el cuarto trial

Examinando las trayectorias de las resoluciones del test por parte de los sujetos que conforman este tercer grupo en el panel diseñado para consulta visual de los datos (ver Figura 3.1), se observa que todos ellos, luego de encontrar el número 5 en el test, avanzan en la exploración de la pantalla hacia la izquierda (continuando la dirección de la espiral de hallazgo que traen) e incluso hacia arriba, lo que no les permite hallar rápidamente el número 6 localizado hacia la derecha pero hacia abajo del número 5. Por otro lado, podría conjeturarse que estos individuos que tardan en hallar el número 6, además de que siguen con la inercia de movimiento ocular siguiendo una espiral antihoraria que traen, no observan el interior de esa espiral que van recorriendo en la búsqueda de los 5 primeros números del test y ese lugar es justamente donde se encuentra el número 6 que sigue al último número de la espiral en sentido antihorario que vienen recorriendo (ver Figura 5.17).

Al analizar con detenimiento las trayectorias visuales de resolución del test de los individuos que integran el segundo grupo, en color amarillo en la Figura 5.16 con tiempos altos en los trials segundo y cuarto, se advierte que durante la búsqueda del número 2 en el segundo trial ellos vuelven a fijar la mirada en el número 1, y cuando se encuentran buscando el número 4 en el cuarto trial vuelven a hacer fijaciones sobre los tres primeros números, especialmente sobre el 1 y el 2. Estas observaciones podrían dar la pauta de que los sujetos del grupo "amarillo" no conservan en la memoria la ubicación de los números localizados con anterioridad o bien los "miran" sin "verlos", sin reconocerlos o sin registrarlos, en esos primeros trials.

Si observamos en general el agrupamiento obtenido con el método "Time Series K-Means", vemos que los grupos de sujetos no muestran grandes diferencias entre sí en los últimos trials, aproximadamente después del sexto trial. Sin embargo se aprecian notorias diferencias, entre los sujetos en general y en los grupos en particular, respecto a su desempeño temporal en los cinco primeros trials. De hecho, los grupos parecen estar conformados por sujetos con altos tiempos de resolución en los trials 1 y 3, los que emplearon mucho tiempo en los trials 2 y 5, los que destinaron bastante tiempo a la resolución de los cinco primeros trials y los que no emplearon demasiado tiempo en ninguno de los 25 trials (esos 3 individuos distinguidos en el primer grupo).

Una posible explicación a esta situación puede tener que ver con el hecho de que los primeros trials son los de mayor búsqueda y escaneo visual, dado que toda la pantalla del test está por explorar. En

esta primera etapa del test, la diferencia en tiempos de resolución individual por trial es mayor que en el resto de la prueba. El agrupamiento que obtuvimos parece distinguir a los sujetos en relación a el o los trials de entre los primeros cinco en los que encontraron más dificultad. Incluso el clustering parecería reflejar que quienes realizan una mayor exploración en el trial 1 para hallar el número 2, encuentran relativamente rápido el número 3 (podrían haberlo visto en la exploración del primer trial y recordarlo) pero demoran en hallar el número 4. La gran exploración previa al hallazgo del número 4 de estos sujetos podría inducirnos a suponer que probablemente hayan visto el 4 anteriormente en su exploración pero no recuerdan su ubicación. Estos son indicios que podrían relacionarse con conceptos clínicos como la memoria de trabajo y resultar de utilidad a los especialistas de la salud para analizar diferencias y similitudes de posibles comportamientos de resolución del test TMT-A por parte de los individuos.

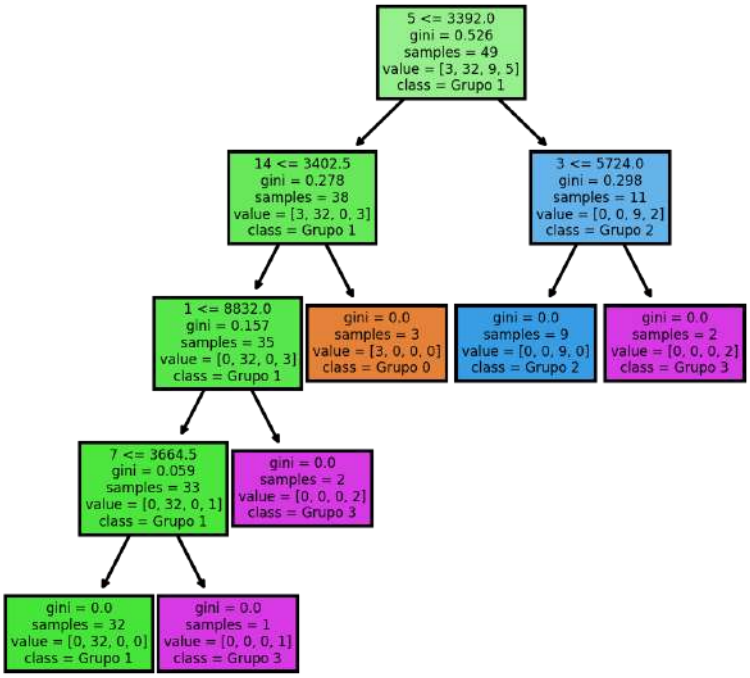


Figura 5.18: Árbol de decisión del agrupamiento de tiempos de resolución

Al obtener el árbol de decisión del agrupamiento de las 49 series temporales con el método de K-Medias y la distancia euclídea en 4 grupos que hemos estado comentando, árbol que se muestra en la Figura 5.18, observamos las reglas que determinan la partición del conjunto de individuos bajo estudio en los 4 grupos. En este árbol de decisión podemos apreciar que la variable más relevante en el agrupamiento de los sujetos por este método de clustering es el valor del tiempo empleado en la resolución del quinto trial, correspondiente a la búsqueda del número 6 en el TMT-A. Las siguientes variables que deciden los grupos son los tiempos en los trials tercero y décimo cuarto, seguidas por los tiempos en el primero y sèptimo trial. Según este árbol de decisión, el grupo mayoritario de 32 individuos está constituído por los sujetos cuyos valores de tiempos de resolución en los trials número 5, 14, 1 y 7 son los menores entre los datos de análisis; mientras que el grupo formado por 9 sujetos se ha conformado con quienes tienen mayores tiempos en el trial 5 y menores en el trial 3. El grupo minoritario del clustering, con tan solo 3 individuos, se caracteriza por incluir a quienes tienen pequeños valores de tiempos de resolución en los trials 5 y 14, y valores mayores de tiempos en el primer trial. El restante grupo está formado por 5 individuos que tienen tiempos altos en los trials 5, 3, 1 y 7.

Por otra parte, se realizó un análisis de la aplicación de los métodos de agrupamiento para menos de 4 grupos con la intención de obtener grupos con un mayor número de sujetos. De este análisis puede comentarse que el método SoftDTW para el caso del clustering en tres conglomerados, cuyos grupos pueden visualizarse en la Figura 5.19, arroja un grupo bastante uniforme con tiempos relativamente bajos en todos los trials y otros dos grupos a los que pertenecen los sujetos con tiempos atípicos en

algún trial: un grupo minoritario con los individuos con tiempos muy grandes en los primeros cinco trials y otro grupo con tiempos medios en varios trials.

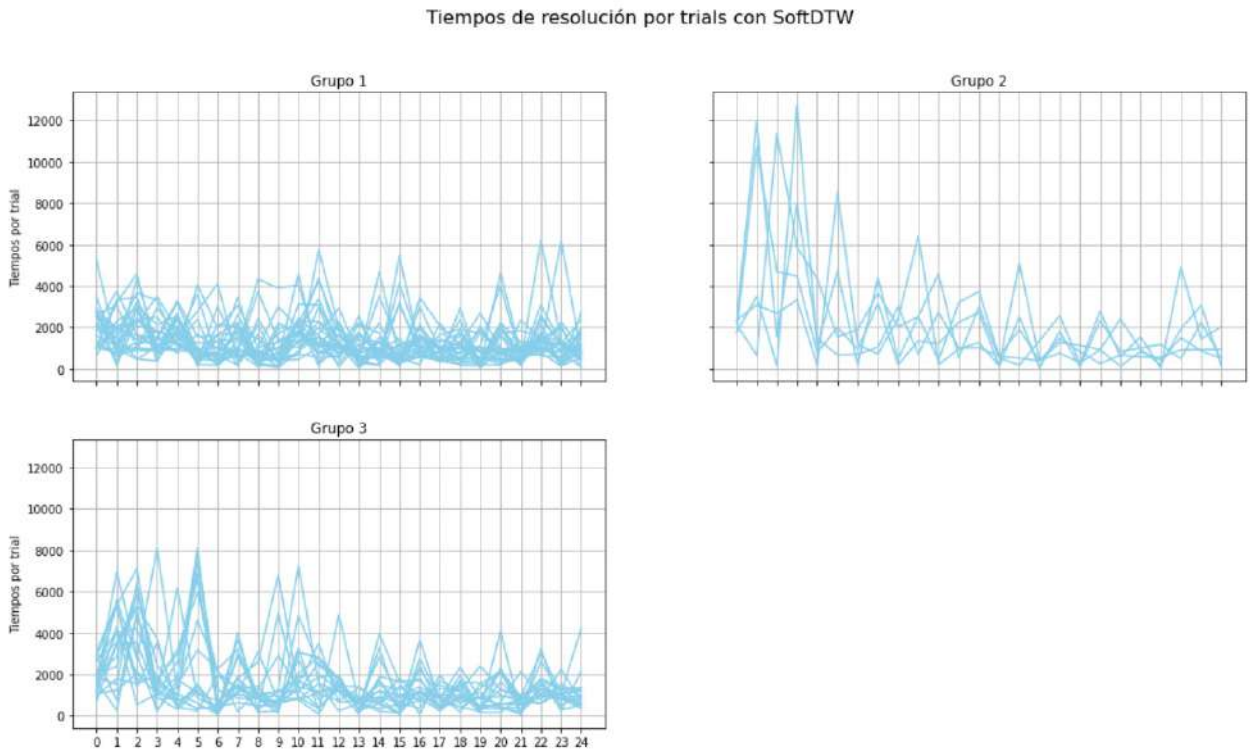


Figura 5.19: Agrupamiento de tiempos por trial

Agrupamiento de velocidades por trial

Como se mencionó anteriormente, se calcularon las velocidades por trial como el inverso del tiempo de resolución del trial, considerando el avance de una unidad de distancia sobre el tiempo empleado en ese avance. Así, por supuesto, las series de velocidades tienen también una longitud de 25 unidades.

Comenzando con el primer estudio exploratorio inicial de estas series, pueden verse los valores de las medianas de las velocidades de resolución en cada trial en la Tabla 5.1.

Trial	**Mediana**		**Trial**	**Mediana**		**Trial**	**Mediana**
0	0,617691		9	2,501761		17	1,751135
1	0,818449		10	0,818502		18	1,451641
2	0,663006		11	1,347505		19	3,103748
3	0,928685		12	1,082050		20	1,455110
4	0,979796		13	2,424890		21	3,198422
5	1,345374		14	1,454252		22	0,810162
6	2,953496		15	3,152859		23	1,688970
7	1,091454		16	1,301523		24	1,840650
8	1,927181		-	-	-	-	

Tabla 5.1: Tabla de medianas de velocidades de resolución de cada trial

Complementariamente, en la Figura 5.20 se muestran los valores estadísticos de las velocidades de resolución por trials de los sujetos de estudio, donde los valores faltantes se completaron con velocidad cero (idea equivalente a un tiempo de resolución infinito en el trial “no resuelto”) como se mencionó anteriormente que se haría.

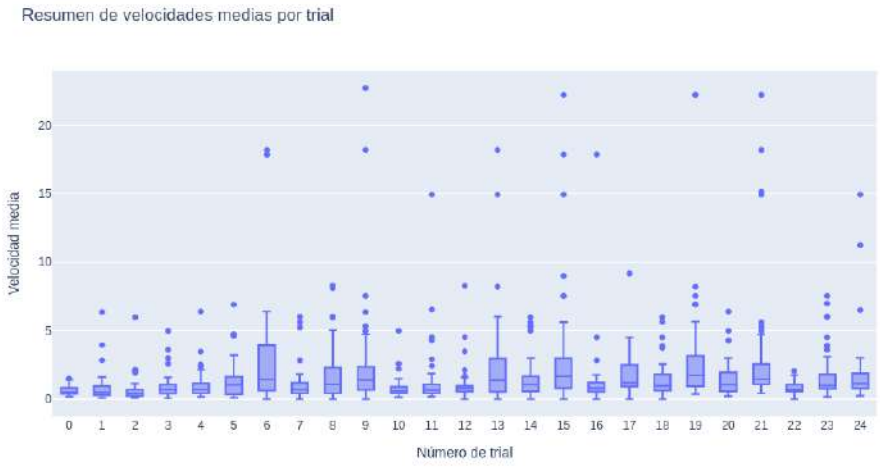


Figura 5.20: Estadística de los valores de velocidades de resolución por trial

Como era de esperarse y a grandes rasgos, analizando la tabla y los boxplots, se observan las más bajas velocidades de resolución en los primeros cinco trials y en la zona del trial 10 al 12. Otra observación que puede hacerse es que en estos boxplots de las velocidades puede observarse con mayor detalle la variabilidad entre sujetos en los trials de mayor velocidad, que no puede ser apreciada con detalle en los boxplots de tiempos de resolución dado que esas velocidades son las de menores tiempos.

Respecto al estudio de agrupamiento en el caso de estas series de velocidades por trial, también se utilizó el método de la inercia para obtener el número óptimo de grupos y este método indica que 5 es el número adecuado de grupos para el conjunto de los individuos bajo análisis respecto a las velocidades de resolución de cada uno de ellos en cada trial.

El hecho de que el número de grupos con este enfoque de grupos por velocidades sea distinto al número de grupos que se utilizó para agrupar a los mismos sujetos considerando sus tiempos de resolución por trial, da una variante más a la posible comparación de resultados del clustering que puede enriquecer el posterior análisis. Por esta razón, se procedió a agrupar los datos de estudio según las velocidades por trial en 5 grupos, sin unificar el número de grupos a obtener con el agrupamiento de series de tiempos y de velocidades por trial.

Así, a las series temporales de velocidades por trial se le aplicaron, para cinco grupos, los mismos métodos de agrupamiento que a las series de tiempos de resolución, es decir, el método K-Medias para series temporales con distancias euclídea y Dynamic Time Warping (DTW). En el caso de la distancia DTW se utilizó también con el algoritmo conocido como SoftDTW.

En este punto, resulta oportuno mencionar que en este trabajo no estamos especialmente interesados en obtener resultados de agrupamientos "óptimos" desde el punto de vista de la cohesión y demás propiedades deseables de los clusters. Si bien se realizaron testeos de los agrupamientos obtenidos por los distintos métodos utilizando los índices de calidad de un método de clustering, la extracción de conclusiones de los agrupamientos se focaliza en el análisis de los grupos obtenidos en los que se aprecian distinciones entre los sujetos que constituyen los grupos, respecto a algún criterio o característica que resulte útil para el problema de estudio.

Con esta idea, el método K-Means fue el que arrojó resultados más interesantes para el agrupamiento utilizando la distancia euclídea. En la Figura 5.21 podemos visualizar los valores de las medianas por grupo determinados por este método, donde se observa que los trials destacados respecto a posibles diferencias de los grupos en valores de velocidades son los trials 6, 9, 13, 15, 19 y 21.

Concretamente, a la vista del gráfico de la Figura 5.21 se distingue un grupo con altos valores de velocidad en el trial 6, un grupo con valores altos en los trials 9 y 19 (un grupo con ésta característica se observa en los tres métodos agrupamiento aplicados), un grupo con velocidad alta en el trial 15, uno con mediana alta de velocidad en el trial 21 y un restante grupo con velocidades relativamente bajas en todos los trials.

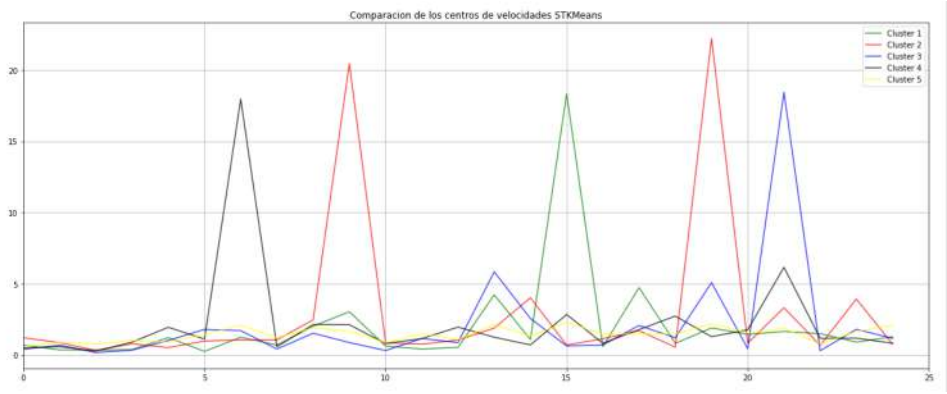


Figura 5.21: Medianas de los grupos de velocidades por trial

No obstante estas observaciones, debemos tener en cuenta que en el análisis de los valores de las medianas de una variable en un grupo tiene una gran influencia la presencia de valores atípicos en el grupo. Para morigerar el efecto visual de la diferenciación de medianas por la presencia de outliers en el grupo y ratificar o rectificar esta idea que nos da el gráfico de medianas, obtuvimos los gráficos de boxplots de cada uno de los 5 grupos en los 25 trials del test y estos gráficos de caja se muestran a continuación en la Figura 5.22.

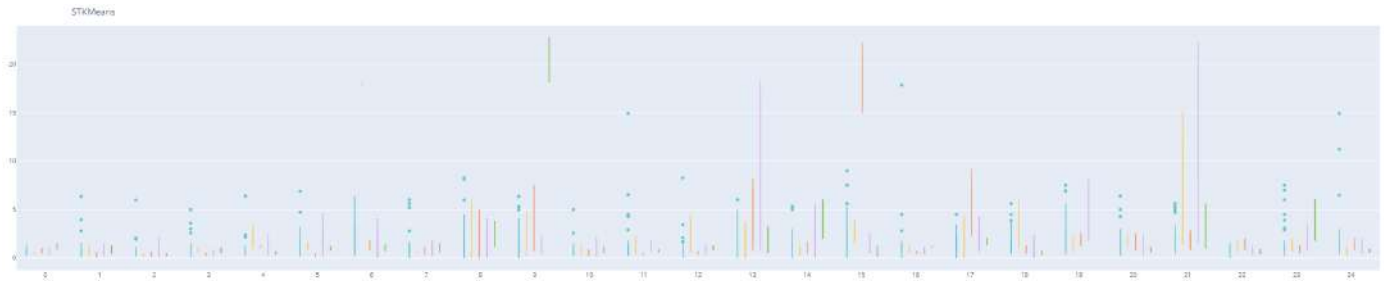


Figura 5.22: Gráfico de boxplots de los grupos de velocidades por trial

Justamente, en esta visualización no se observan trials donde se separen notoriamente los valores de las velocidades en algunos de los cinco grupos, rectificando la impresión que pudo habernos dado el gráfico de medianas.

El comentario de interés que puede hacerse a partir de estos gráficos de boxplots es que se distinguen claramente un grupo con muy alta velocidad en relación a los otros grupos en el trial 9; otro con igual situación, aunque en menor medida, en el trial 19; y un tercero con la misma característica de notoriamente grandes velocidades en el trial 15. Puede observarse esta situación en la Figura 5.22 y más claramente en la Figura 5.23.

Los gráficos mostrados en la Figura 5.24, que incluyen los valores de velocidades de todos los sujetos por grupo en todos los trials, corroboran este comportamiento grupal en las velocidades de ciertos trials en forma particular.

Otra observación que puede hacerse a partir de los gráficos de la Figura 5.24 es que el comportamiento de los individuos respecto a la velocidad con que resuelven los distintos trials del test es muy similar en los primeros cinco trials; en otras palabras, todos los sujetos son más o menos igual de lentos o de rápidos en el hallazgo de los 5 primeros números de la prueba. Las diferencias que distinguen los grupos se presentan en las velocidades de resolución de las últimas tres cuartas partes del test, lo que podría indicar diferencias entre los sujetos respecto a cuánto recuerdan en las etapas avanzadas del test de la exploración pasada, es decir, vuelve a aparecer en el ánalisis una posible conexión con la idea de memoria de trabajo.

De todas maneras, como puede verse en la Figura 5.24, la cantidad de sujetos por grupo es variada y este hecho debe tenerse en cuenta a la hora de concluir posibles características generales de los grupos.

Figura 5.23: Gráficos de caja de velocidades por grupo en trials distinguidos

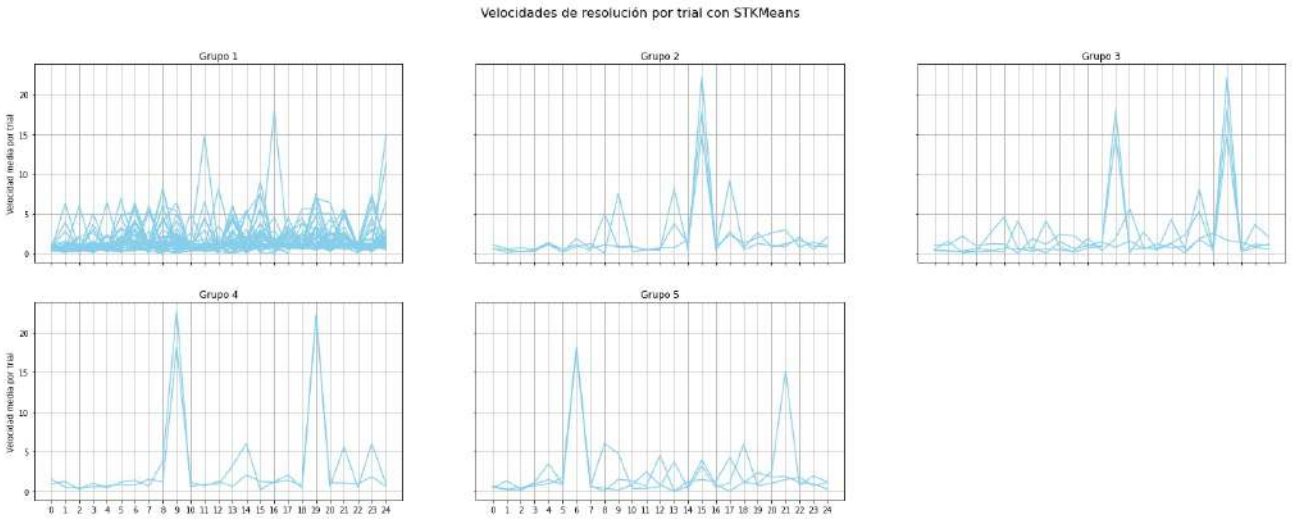


Figura 5.24: Agrupamientos de velocidades medias por trial

Un comentario general que puede hacerse es respecto a que los agrupamientos según el tiempo de resolución por trial asocian a los sujetos según su desempeño en los primeros trials, mientras que los agrupamientos por velocidades diferencian a los sujetos según su resolución en los últimos trials. Este hecho probablemente pueda explicarse recordando que se ha definido la velocidad de cada trial como la magnitud inversa del tiempo de resolución, por lo que la significancia de los valores de los tiempos y de las velocidades es inversa en cada sujeto en los dos enfoques de agrupamiento. Esta misma razón explica que los tiempos de resolución presentan globalmente más variabilidad en la primera mitad del test y las velocidades presentan una mayor varianza de valores en los trials de la segunda mitad del test, lo que puede apreciarse en las Figuras 5.9 y 5.20 de boxplots de valores de tiempos y velocidades.

El número óptimo de 5 grupos fue el resultado de la aplicación del método de la inercia pero se analizaron los agrupamientos en 2 y 3 conglomerados con los mismos métodos utilizados para el clustering en 5 grupos, con el objeto de observar posibles agrupamientos con un mayor número de sujetos en los grupos y alguna posible diferenciación más global de comportamiento entre los 49 sujetos.

Un comentario que puede realizarse de este análisis es que, para agrupamientos en 2 grupos, el método DTW agrupa a casi todos los sujetos con valores atípicos en algún trial en distinto grupo

respecto a aquellos individuos con valores más cercanos a la media de velocidades en todos los trials, aunque podríamos mencionar la excepción de dos sujetos con alta velocidad en el trial 25 pero con muy baja velocidad en el trial 14.

5.2.2. Series de zonas

A partir de estudios previamente realizados del Trail Making Test y la experiencia de los investigadores del grupo Neufisur en el estudio del seguimiento ocular en el TMT-A, se reconoce la existencia de regiones en el test que poseen una geometría distintiva y provocan cambios en las estrategias de búsqueda de los números durante la prueba. Estas diferencias geométricas en distintas regiones del test implican una dificultad de resolución dispar entre algunas zonas de trials del TMT-A.

Considerando esta situación, a continuación se desarrolla un estudio de agrupamiento de las resoluciones de los sujetos proponiendo una sectorización del test TMT-A que refleja estas posibles diferencias en distintos tramos de la resolución de la prueba. Esta sectorización en zonas de características particulares consiste en la definición de seis regiones dentro del test que parecen originar distintos comportamientos en la resolución por parte de los sujetos. Definidas estas seis zonas, y a partir de los datos registrados de movimiento ocular, se obtienen las series temporales de tiempos y velocidades de resolución en cada uno de estos sectores. Estas series se sometieron a los mismos métodos de agrupamiento que se les aplicaron a las series temporales de tiempos y velocidades por trials en las secciones anteriores, esta vez para obtener agrupamientos según la información de tiempo y velocidad zonal.

Concretamente, las nuevas series de tiempo tienen como variable dependiente los valores de tiempos y velocidades de resolución en cada zona y como variable independiente la zona correspondiente del test de acuerdo a los seis siguientes tramos distinguidos del TMT-A:

- Región 1: trials 1 al 5, inclusive.
- Región 2: trials 6 al 10, inclusive.
- Región 3: trials 11 y 12, inclusive.
- Región 4: trials 13 al 17, inclusive.
- Región 5: trials 18 al 21, inclusive.
- Región 6: trials 22 al 25, inclusive.

Tiempos de resolución por zonas

Al analizar las medidas estadísticas de los valores de tiempos de resolución por zonas del conjunto de individuos que se estudia, se obtienen los gráficos de boxplot que se muestran en la Figura 5.25.

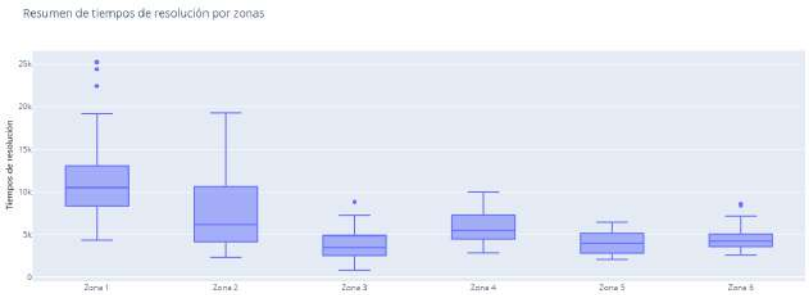


Figura 5.25: Estadística de los tiempos de resolución por zonas

En estos gráficos puede apreciarse una tendencia global decreciente de los tiempospor zona a medida que se avanza en la resolución del test, con un descenso bastante notable en el tiempo insumido por los sujetos bajo análisis en la zona 3 que muestra tiempos incluso menores que los de la zona 4, alterando la tendencia decreciente global. Es decir, se aprecia un descenso de los tiempos de resolución de los sujetos desde la zona 1 hasta la zona 3, con un leve ascenso en el tiempo medio de resolución en la zona 4 para volver a descender la media de tiempos levemente en la zona 5 y, con menor variabilidad, emplear tiempos con una media similar a la zona 5 en la zona 6.

A éstas series de tiempos de resolución por zonas se les aplicó el método de agrupamiento de K-Medias en 4 grupos, que es el número de grupos sugerido por el método de la inercia, utilizando la distancia euclídea y la medida Dynamic Time Warping con los dos algoritmos (estándar y soft).

En los gráficos de los valores de las medianas de los tiempos por zona obtenidas por grupo de acuerdo a los distintos métodos aplicados, que se muestran en la Figura 5.26, no se observan grandes diferencias entre los valores de las medianas de tiempos por zonas empleados por los individuos de cada grupo, excepto quizás en la primer zona. Además, tampoco se observan notorias diferencias en los valores de medianas por grupo en los resultados arrojados por los tres métodos de agrupamiento aplicados.

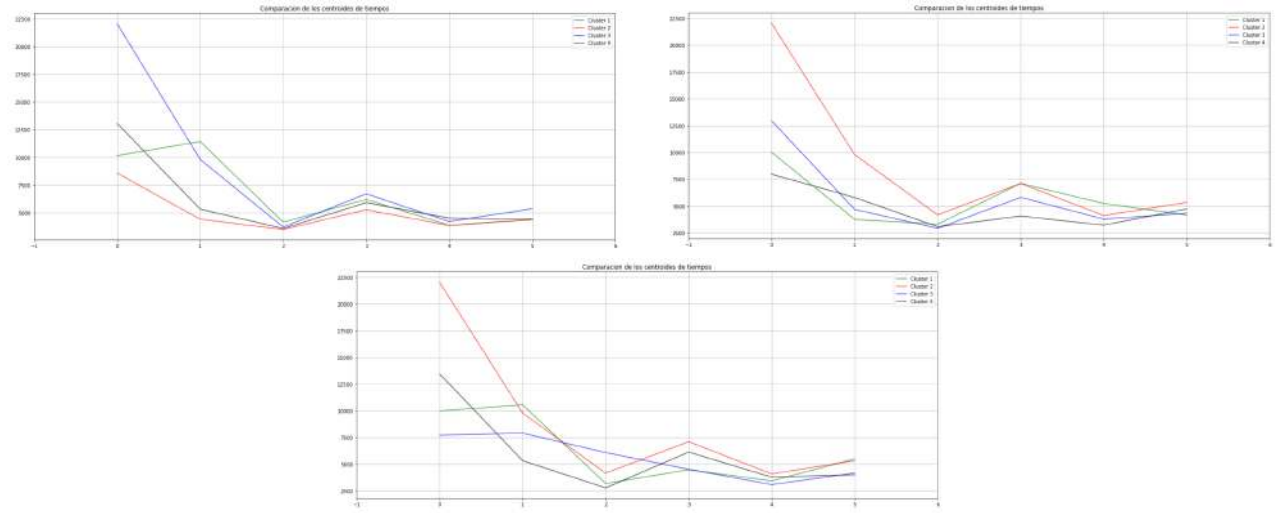


Figura 5.26: Medianas de tiempos de resolución por zonas

Al analizar los gráficos de boxplots de estos grupos conformados por cada método de clustering, se aprecia en todos los casos una diferencia en los valores de la mediana de tiempos de resolución en los distintos grupos únicamente en la primer zona. En particular, el agrupamiento obtenido utilizando el método de K-Means con la distancia Dynamic Time Warping y el algoritmo SoftDTW muestra una clara diferencia en los valores de los tiempos de resolución de la primer zona entre todos los grupos, es decir, este método construye los 4 grupos separando a los sujetos por el tiempo empleado en la resolución de la zona uno, correspondiente a los cinco primeros trials. En la Figura 5.27 se muestra este gráfico de boxplots de grupos para todas las zonas, y en la Figura 5.28 puede verse más claramente lo que sucede en la zona 1 del test en el caso de este mismo agrupamiento.

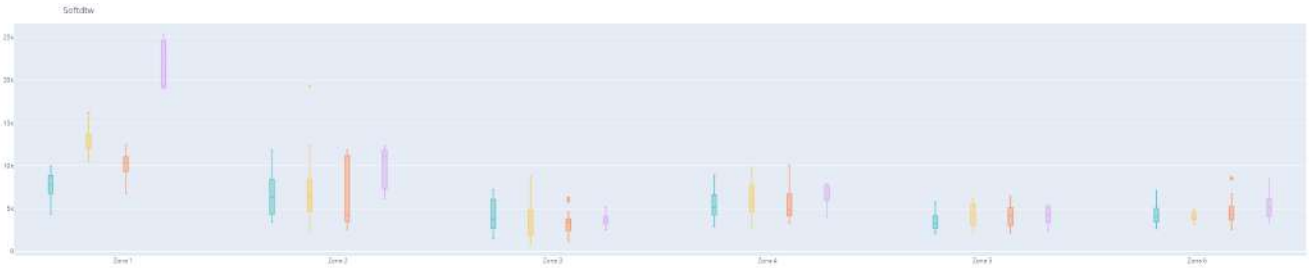


Figura 5.27: Comparación estadística de tiempos de resolución por grupos

Si se observa la distribución de todas las series de tiempos de resolución por zonas en cada grupo obtenida por cada método, que puede visualizarse en la Figura 5.29, puede mencionarse un dato interesante que es que los tres métodos de agrupamiento aplicados forman un grupo con los cinco individuos que tienen los más altos tiempos de resolución de la zona 1. Estos cinco individuos están agrupados en un mismo clúster y separados del resto de los sujetos en los tres casos. Por otro lado, si observamos con detenimiento los agrupamientos obtenidos, vemos que las diferencias más claras entre los valores de las

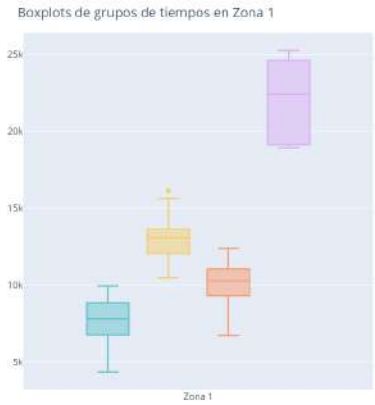


Figura 5.28: Gráfico de boxplots de tiempos de resolución en la Zona 1

medianas de tiempos por grupos están en las zonas 1 y 2. Este hecho ya fue observado en los agrupamientos obtenidos para la misma variable tiempo al considerar sus valores por trial y esta característica parece ser coocurrente con el hecho de que en los primeros trials hay más variabilidad en los valores de los tiempos empleados por cada sujeto para la resolución de esos trials.

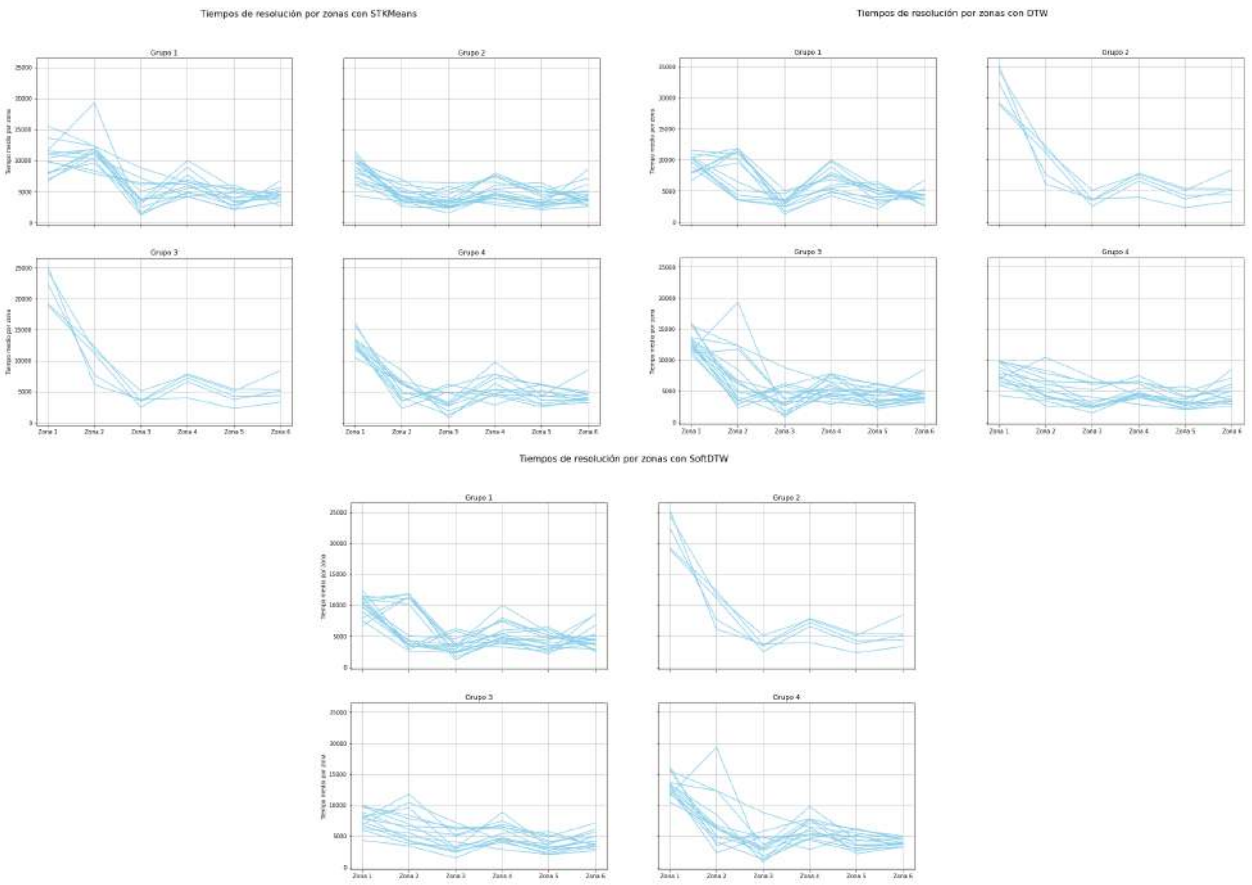


Figura 5.29: Visualizaciones de los agrupamientos de tiempos por zonas

En particular, en el agrupamiento que utiliza la distancia euclídea, tres de los grupos se diferencian por sus valores de tiempos de resolución en la Zona 1: uno de los grupos (el Grupo 3 en la Figura 5.29) está integrado por los 5 sujetos que tienen los valores más altos de tiempo, un segundo grupo con 12 sujetos cuyos valores de tiempo son intermedios y un tercer grupo formado por los 18 sujetos restantes con los menores valores de tiempos de resolución. Una observación que puede hacerse es que los cinco

individuos que más demoraron en la resolución de la Zona 1 tienen valores intermedios de tiempos en la Zona 2 (Grupo 3 de la Figura 5.29), mientras que quienes tienen los valores más altos de tiempos en la Zona 2 (Grupo 1 en la gráfica de la Figura 5.29) tienen tiempos intermedios en la Zona 1. En las restantes Zonas 3, 4 y 5 no se observan comportamientos especialmente distinguidos en los valores de tiempos de resolución en ninguno de los grupos obtenidos con la métrica euclídea.

En el árbol de decisión de este agrupamiento con el método K-Medias y la utilización de la distancia euclídea, que puede verse en la Figura 5.30, se destacan como variables de separación de grupos los valores de los tiempos empleados en la resolución de las Zonas 1 (muy especialmente) y Zonas 2, 3 y 6 (en menor medida).

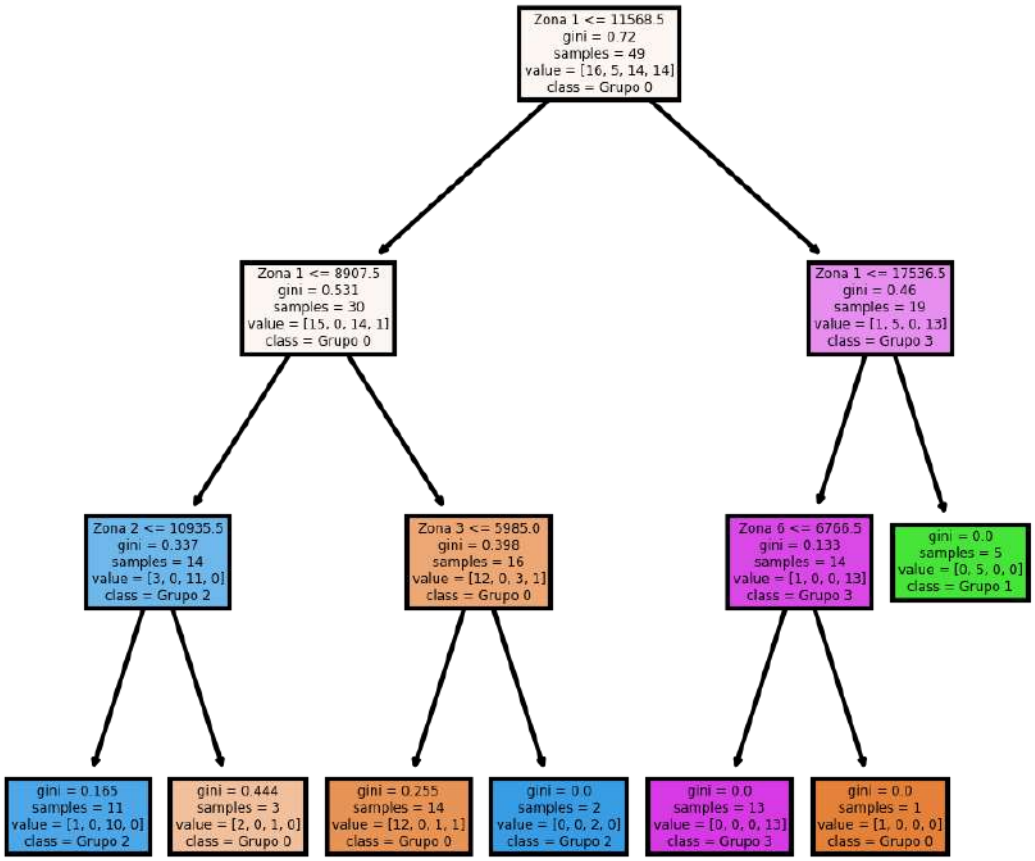


Figura 5.30: Árbol de decisión del agrupamiento de tiempos por zonas

Velocidades de resolución por zonas

En esta sección se analizarán agrupamientos de las resoluciones del TMT-A de acuerdo a dos nociones de velocidad definidas por zona a partir de los tiempos de resolución por trial de los sujetos.

En una primera instancia, utilizaremos una noción de velocidad que llamaremos *velocidad media por trial en la zona* definida como la suma de las velocidades (inverso del tiempo) de cada trial de la zona, dividida por la cantidad de trials que incluye esa zona. Esencialmente, esta cantidad da idea de la velocidad con que el sujeto encuentra cada número de esa zona del test.

La segunda idea de velocidad que se considerará para realizar agrupamientos de las series será la que se llamará en lo que sigue *velocidad regular de avance* y que se calcula como el cociente entre el número de trials de la zona (cantidad de avance) y el tiempo empleado en resolver todos esos trials, es decir, en resolver la totalidad de la zona. Esta idea refiere a la velocidad con que el individuo avanza en la resolución de la secuencia de números de esa zona del test.

Velocidades medias por trial en la zona

Obtenidas las velocidades medias por trial en cada una de las seis zonas para cada uno de los 49 sujetos, se construyeron los gráficos de boxplots de los valores de éstas velocidades medias por trial por zona en los que se detectan las diferencias entre las bajas velocidades con que los sujetos resuelven los trials de las zonas 1 y 3 del TMT-A, es decir, los trials 1 al 5 y trials 11 y 12 del test, y las velocidades de las demás regiones de valores bastante superiores y similares entre sí. Pueden verse estos boxplots en la Figura 5.31.

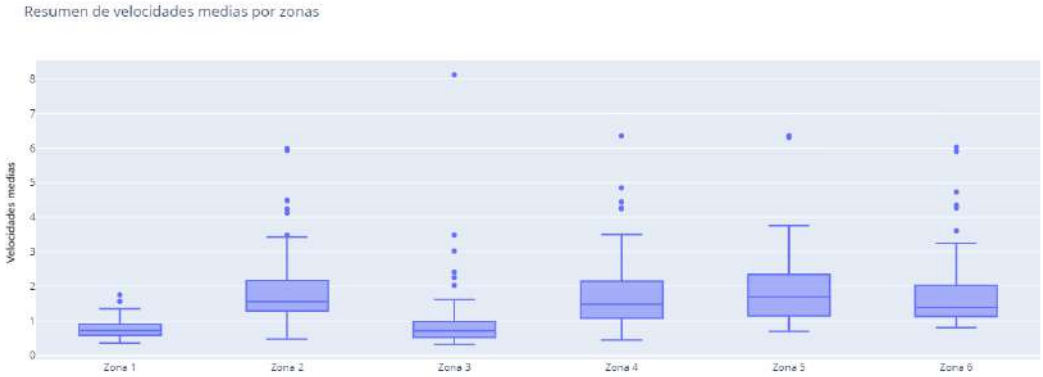


Figura 5.31: Estadística de los valores de velocidades medias por zonas

Luego, se aplicó el método de la inercia al conjunto de datos de velocidades medias por trial en cada zona para obtener un apropiado número de grupos en que pueden agruparse las resoluciones del TMT-A de los individuos bajo estudio. Este procedimiento arrojó como resultado que cinco grupos es un adecuado número de grupos para el conjunto de datos. A continuación se llevo a cabo el proceso de clustering utilizando el método de K-Medias con distancia euclídea y con la medida DTW (mismos algoritmos utilizados en las secciones anteriores) para agrupar los datos de las velocidades medias por trial por zonas en cinco grupos.

Analizando los valores de la mediana de estas velocidades medias en cada grupo obtenido en los distintos agrupamientos, se aprecian diferencias en los valores centrales de los grupos en las zonas segunda, tercera y quinta (en los dos métodos de agrupamiento aplicados) y también en la cuarta zona en uno de los métodos. Ninguno de los agrupamientos evidencia grandes diferencias en los valores de velocidad media por trial en la primer zona del test, es decir, en la resolución de los primeros cinco trials.

Considerando los gráficos de boxplots de las velocidades medias en el caso de agrupamiento donde se utilizó la distancia euclídea, que pueden verse en la Figura 5.32, si bien no se observa una zona en la que todos los grupos se diferencien en sus valores medios de velocidad, pueden distinguirse un grupo caracterizado por valores altos de velocidad media por trial en las zonas 2 y 5 en relación con los valores en las demás zonas, un grupo con velocidades altas en la zona 3, un grupo caracterizado por altos valores de velocidades medias por trial en la zona 4 y otro agrupamiento con valores también notoriamente altos en la zona 6.

Figura 5.32: Gráficos de caja de los grupos de velocidades medias por zonas

Al analizar la composición de todas las resoluciones de cada grupo que resulta del método de K-Medias con la distancia euclídea, que puede visualizarse en la Figura 5.33, se puede advertir que el agrupamiento forma un conglomerado con tres sujetos de altos valores de velocidad media por trial en la zona 3 (el Grupo 1 en la Figura 5.33), un grupo de ocho sujetos con altas velocidades en la última zona 2 y zona 6 (Grupo 3), un grupo de tres sujetos caracterizados por altas velocidades en zonas 2 y 5 (Grupo 5), un grupo de altos valores solo en zona 4 (Grupo 4), y el grupo más numeroso de sujetos con velocidades en la franja media en todas las zonas (Grupo 2 en la Figura 5.33). Los tres sujetos del Grupo 5 con valores especialmente altos en las velocidades de las zonas 2 y 5, evidentemente son distinguibles pues los otros agrupamientos estudiados de 5 grupos también los muestran formando uno de los cinco conglomerados.

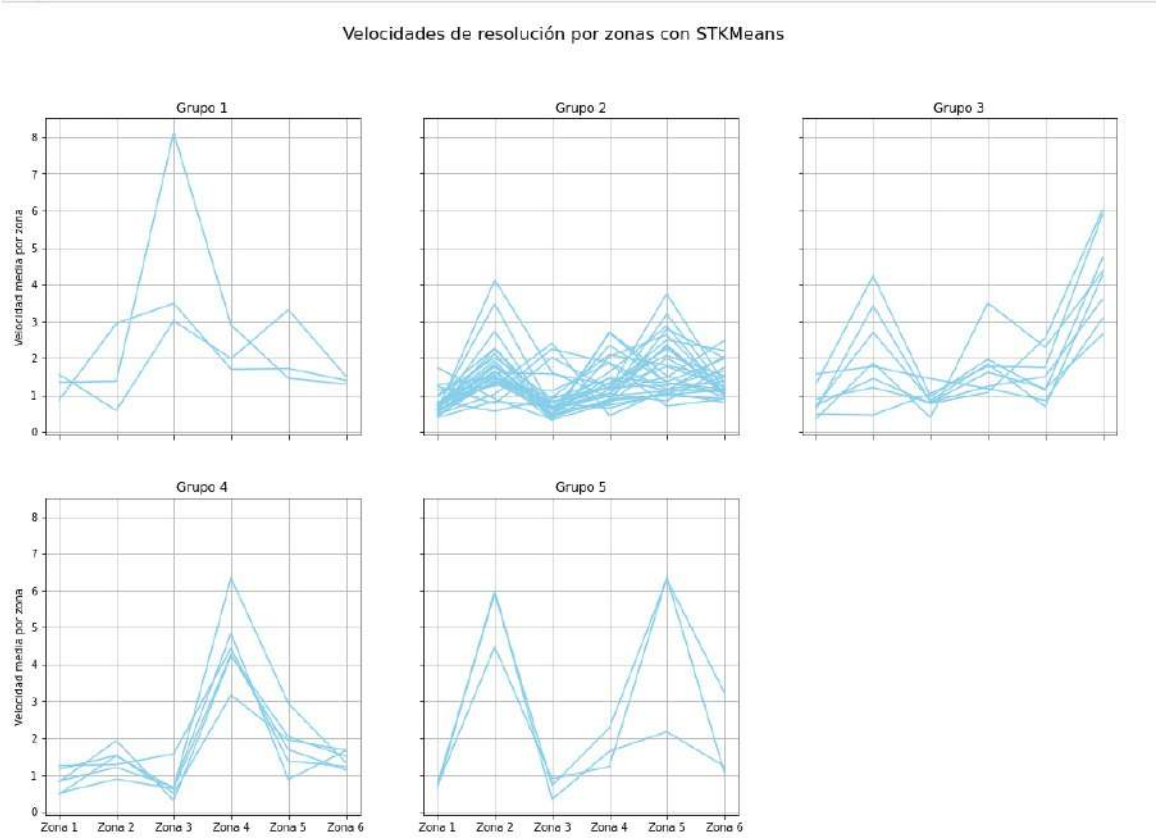


Figura 5.33: Visualización de grupos de velocidades medias por zonas

Al construir el árbol de decisión de este agrupamiento, que se muestra en la Figura 5.34, se advierte que la zona más relevante en cuanto a valores de velocidad media por trial en la formación de los grupos de los sujetos es la zona 6, seguida de la información de las velocidades en las zonas 4 y 2, y finalmente en la zona 3.

Paralelamente, se realizaron pruebas con otras cantidades de grupos utilizando los mismos métodos pero se comprobó que los agrupamientos en tres conglomerados no distinguen particularmente grupos de velocidades medias por trial en ninguna de las zonas y los agrupamientos en dos conglomerados separan los grupos de sujetos según sus valores de velocidades medias por trial en la misma zona 6.

Velocidades regulares de avance

En el caso de las series de velocidades regulares de avance, el gráfico de boxplots que se muestra en la Figura 5.35 también indica menores velocidades en las zonas 1 y 3 respecto de las observadas en las otras zonas pero con diferencias menos pronunciadas y con una mayor variabilidad en los sujetos en relación a las variaciones que tienen en cada zonas las velocidades medias por trial. Esta mayor

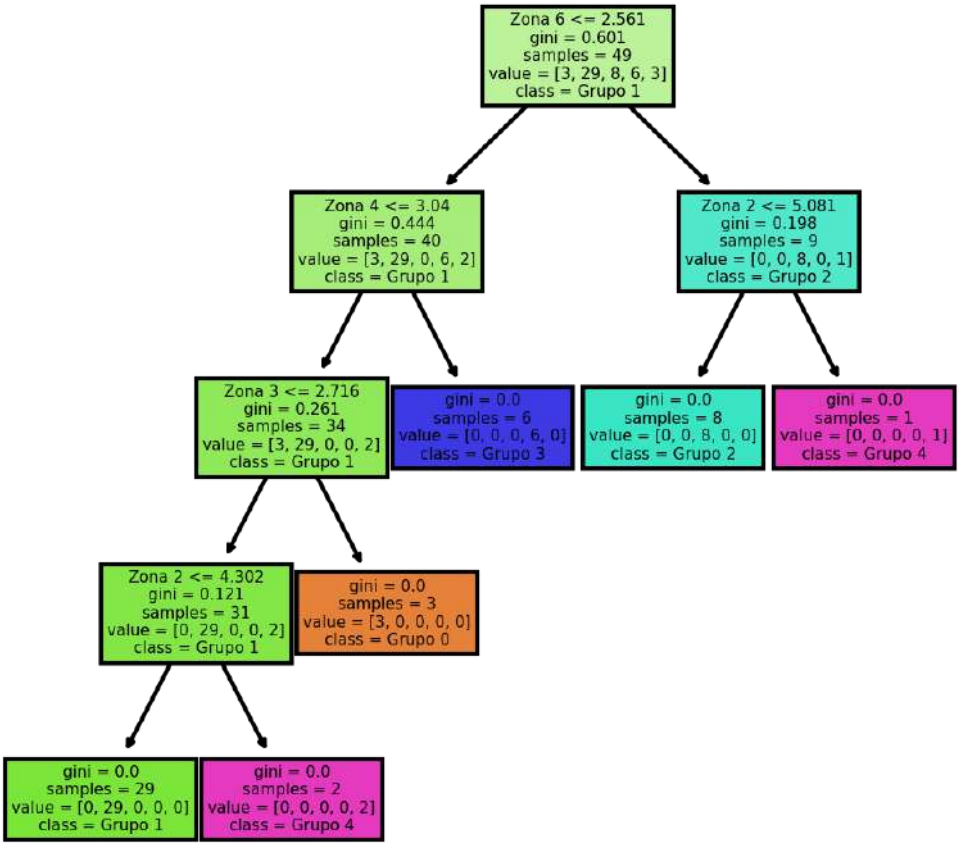


Figura 5.34: Árbol de decisión del agrupamiento de velocidades medias por trial en cada zona

varianza es naturalmente explicable por el hecho de que en este segundo caso de velocidad se considera una velocidad zonal y no por trial, y una mayor franja temporal origina una mayor variabilidad en el conjunto de individuos.

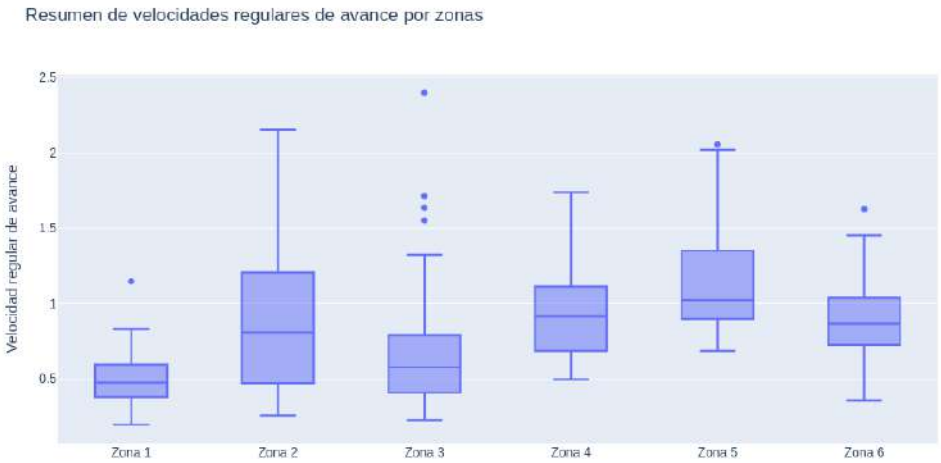


Figura 5.35: Resumen estadístico de las velocidades de avance por zona

Al agrupar por K-Medias con la distancia euclídea y con la medida Dynamic Time Warping, para el caso de 4 grupos que es el indicado por el método de la inercia cuyos gráficos de boxplots pueden versen

en la Figura 5.36, no se observa que ninguno de los agrupamientos separe los grupos por las diferencias en sus valores de velocidades regulares de avance en alguna zona en particular. Éstos gráficos solo indican que el método de K-Medias utilizando la distancia euclídea distingue un grupo de los otros tres por sus altos valores de velocidad de avance en la zona 3 (grupo de boxplot violeta en el primer gráfico de la Figura 5.36).

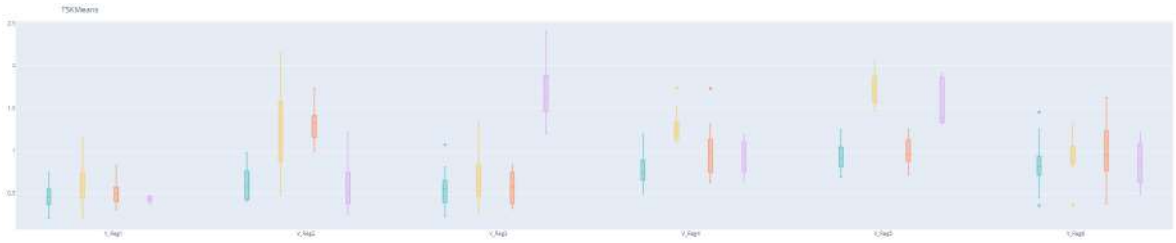


Figura 5.36: Resumen estadístico de las velocidades de avance por zona en cada grupo

Analizando las gráficas de todas las series de velocidades de avance en cada grupo de la Figura 5.37, puede apreciarse un grupo de cinco sujetos caracterizados por sus elevados valores de velocidades regulares de avance en la resolución de las zonas 3 y 5 del test, un grupo con valores considerablemente altos de velocidades en las zonas 2 y 4, otro grupo con velocidades de avance altas en relación a los demás individuos especialmente en las zonas 2 y 5 pero también con velocidades medianamente altas en la zona 4.

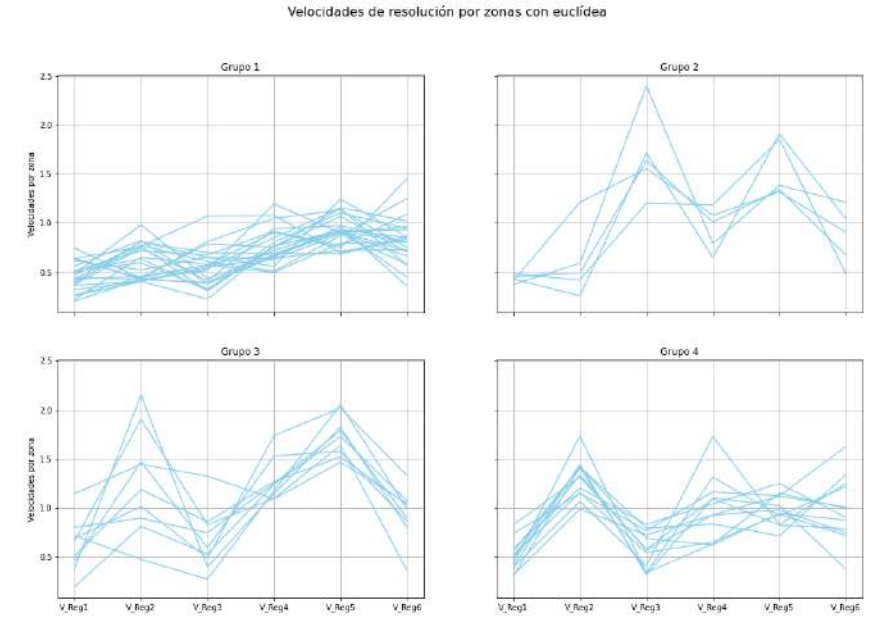


Figura 5.37: Visualización completa del agrupamiento con distancia euclídea en cuatro grupos

En el análisis del agrupamiento para el caso de tres conglomerados, se observa que el método K-Means con la medida Dynamic Time Warping separa a los individuos según los valores de sus velocidades regulares de avance en las regiones 4 y 5, como puede verse en la Figura 5.38.

Figura 5.38: Resumen estadístico de las velocidades de avance por zona en cada uno de los tres grupos

Si se consideran clusterings de dos grupos, pueden verse separaciones claras de los conglomerados según sus valores en las velocidades regulares de avance en las regiones 2, 4 y 5 con DTW, como muestran los gráficos de boxplots de estos grupos en la Figura 5.39.

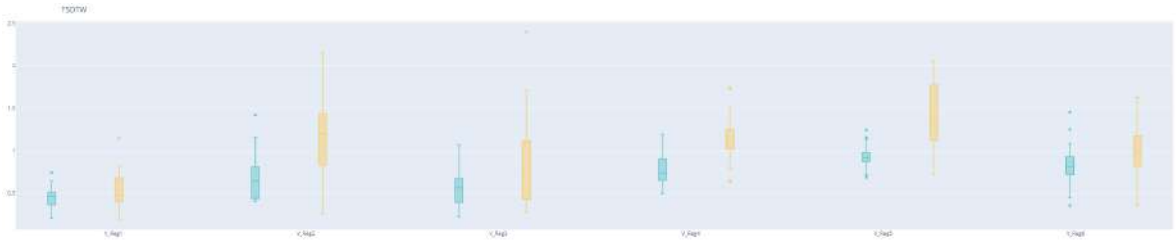


Figura 5.39: Resumen estadístico de las velocidades de avance por zona en cada uno de los dos grupos

Capítulo 6

Secuencias vectoriales y Multimatch

En este capítulo, realizaremos un estudio de agrupamiento de nuestros datos considerando a las trayectorias visuales durante la resolución del test TMT-A como secuencias de vectores fijación-sacada, también llamados scanpaths. Éstas secuencias vectoriales se compararán utilizando el método de análisis multidimensional conocido como Multimatch que, a partir del estudio temporal y espacial bidimensional del scanpath, origina cinco índices de similaridad entre todo par de secuencias vectoriales. Multimatch ha sido utilizado en el estudio de distintos tipos de datos de seguimiento ocular como se refleja en la literatura, por ejemplo, en los trabajos de las referencias [6, 13, 37]. En nuestro caso, procederemos a obtener las similitudes Multimatch para todo par de scanpath de nuestro conjunto de datos y a continuación aplicaremos métodos de clustering a las trayectorias visuales de estudio en función de las medidas de similitud dadas por esos índices, obteniendo agrupamientos definidos a partir de características más globales de los scanpaths en comparación con los producidos por los enfoques de agrupamiento que se han analizado anteriormente.

Comenzaremos la sección describiendo las nociones de fijaciones, sacadas y scanpaths para proseguir con el estudio del agrupamiento.

6.1. Fijaciones y sacadas

El creciente interés que despierta el estudio de los movimientos oculares ha llevado a conocer con amplio detalle la funcionalidad de los distintos elementos que los constituyen: las sacadas, y microsacadas, y las fijaciones (ver, por ejemplo, referencias [10, 21, 30, 32]). De hecho, las métricas más comunes de registro visual se basan en estos conceptos y los investigadores suelen analizar los movimientos oculares en tareas como el TMT en términos de fijaciones y sacadas, y los movimientos oculares pueden caracterizarse como una secuencia de estos elementos.

Las fijaciones son representadas como muestras discretas de puntos de posición cuasi estable hacia donde se dirige la mirada, en algún sentido, son pausas de la mirada sobre regiones informativas de interés y se evidencian por ser segmentos de tiempo de algunos milisegundos en el que el ojo permanece más o menos inmóvil en un punto. Las fijaciones están formadas por un conjunto de datos que presentan una baja velocidad en un entorno espacial pequeño (esto es, poco desplazamiento espacial).

Durante las fijaciones el ojo recopila y envía información al cerebro, son los períodos en los que el sujeto extrae información del objeto que está mirando. Por esto, la cantidad, ubicación y duración de las fijaciones son características que resultan de interés para comprender el proceso cognitivo que se está desarrollando. A partir de las fijaciones podemos inferir, por ejemplo, las dificultades que presenta una tarea para un sujeto.

En el caso particular del TMT puede resultar de interés evaluar si un sujeto revisita una misma región muchas veces sin encontrar su número objetivo, o si hace alguna fijación sobre él y continúa su exploración sin registrarlo, pues estos aspectos podrían estar asociados a la atención que está poniendo en juego el individuo en el desarrollo de la tarea y esto es justamente lo que el test pretende evaluar.

Las sacadas están definidas como los movimientos oculares entre dos fijaciones consecutivas, son movimientos rápidos de la visión que se observan como saltos en velocidad que llevan la fóvea del ojo de

una posición de fijación a otra. Mientras se realiza un movimiento sacádico no se registra información de la escena, aunque puede ser que el cerebro siga procesando en ese período, ver [32]. Las sacadas se caracterizan por su dirección, duración, amplitud y velocidad, y dan información de la precisión con el que el sujeto realiza el movimiento necesario para alcanzar un objetivo, lo cual en algunos casos es de sumo interés, por ejemplo, en salud visual o en la detección de problemas neurológicos.

De esta manera, los movimientos oculares en tareas como el TMT suelen caracterizarse como una secuencia de fijaciones y sacadas que se conoce como *scanpath*. Concretamente, un scanpath es una sucesión de puntos en el plano que corresponden a las posiciones de los centros de cada fijación que realiza el individuo en el orden en que las realiza, y a cada punto se le asigna un peso correspondiente a la duración de la fijación realizada, es decir, un tiempo que indica la duración de la fijación. Así, un scanpath es una sucesión de vectores fijación-sacada que van desde cada una de las posiciones centrales de una fijación hasta la siguiente, y en la que cada fijación tiene asignado un peso que es el tiempo de duración de la fijación. En consecuencia, un scanpath es una simplificación de la trayectoria seguida por la mirada que resume la información más relevante sobre las fijaciones y las sacadas. De este modo, un punto crucial en el estudio de scanpath es la clasificación de los datos registrados por el eye tracker en fijaciones y sacadas.

Cabe mencionar que no existe consenso absoluto sobre cómo determinar los elementos fijación y sacada a partir de los datos de seguimiento ocular, como analiza la referencia [18], y uno de los condicionantes fundamentales es la frecuencia de muestreo de los dispositivos por lo que el proceso de identificación de fijaciones y movimientos sacádicos en los protocolos de seguimiento ocular es una parte esencial del análisis de datos de movimiento ocular y puede tener una impacto dramático en el análisis de alto nivel. La identificación de fijaciones reduce significativamente el tamaño y la complejidad del protocolo de movimiento ocular, lo que resulta útil por al menos dos razones: poco o nada del procesamiento visual se puede lograr durante un movimiento sacádico, por lo que los caminos reales recorridos durante las sacadas suelen ser irrelevantes para muchos intereses de investigación, y los movimientos oculares más pequeños que ocurren durante las fijaciones, como temblores, desviaciones y movimientos postsacádicos, a menudo significan poco en análisis de alto nivel. Por lo tanto, la identificación de la fijación es un método inherentemente estadístico conveniente para minimizar la complejidad de los datos de seguimiento ocular conservando la mayoría de las características esenciales a los efectos de la comprensión del procesamiento cognitivo y visual.

En el trabajo [35] se propone una taxonomía que clasifica algoritmos de identificación de fijaciones y sacadas con respecto a cinco criterios espaciales y temporales. En nuestro caso, para el estudio de los scanpath de las trayectorias de resolución del TMT-A con que contamos para este trabajo de tesis, y la posterior aplicación del método Multimatch prevista, clasificamos los datos obtenidos del dispositivo de seguimiento ocular identificando las fijaciones mediante la utilización de un algoritmo desarrollado por el grupo de investigación NEUFISUR que realiza la clasificación de los datos en fijaciones y sacadas con base en las velocidades de desplazamiento punto a punto. Pueden consultarse algunos detalles técnicos del algoritmo en el apéndice F.

En la Figura 6.1 puede verse un ejemplo de la visualización del conjunto de fijaciones y sacadas del TMT-A en pantalla, en la que la duración de la fijación se representa como un círculo con centro en la posición de la fijación y radio proporcional al tiempo de la fijación. Los segmentos que unen los centros de los círculos son las sacadas que indican la secuencia entre las fijaciones.

6.2. El método Multimatch

El método conocido como MultiMatch, propuesto originalmente por Jarodzka, Holmqvist and Nyström en [19], tiene un enfoque multidimensional y fue diseñado específicamente para calcular la similitud de rutas de exploración visual con un tratamiento vectorial. El método representa las rutas de exploración como una secuencia de vectores geométricos que describen movimientos individuales en un espacio bidimensional: cualquier ruta de exploración está formada por una secuencia vectorial en la que los vectores representan movimientos sacádicos, y la posición inicial y final de los vectores sacádicos representan fijaciones.

Multimatch compara las rutas de escaneo basándose en múltiples características de la trayectoria,

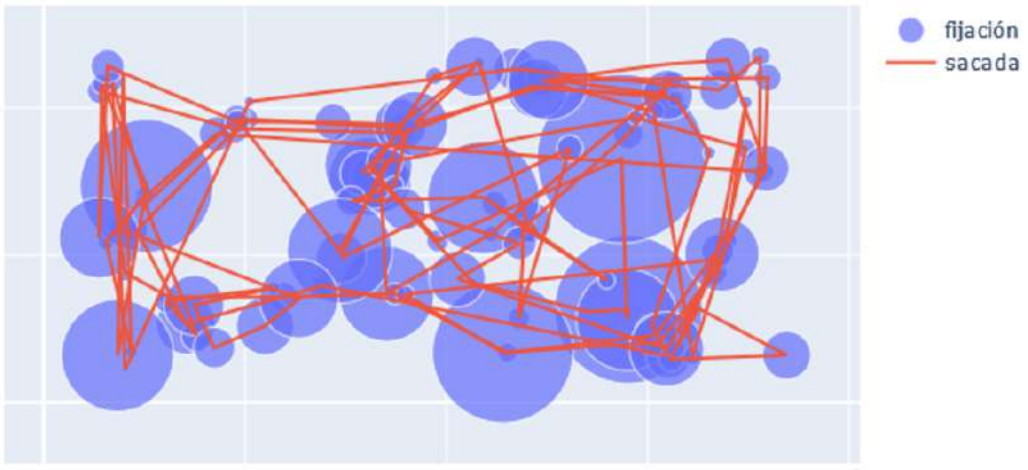


Figura 6.1: Ejemplo de scanpath

su secuencia temporal y su estructura espacial, considerando la forma de cada ruta de escaneo en su totalidad para una evaluación de similitud multidimensional. El método obtiene cinco medidas independientes que capturan la similitud entre diferentes características de las rutas de exploración, a saber, forma, dirección, longitud, posición y duración.

El método Multimatch tiene varias ventajas sobre otros métodos. Quizás la principal de las ventajas de este método sea el hecho de que proporciona varias medidas para evaluar similitud entre las rutas de exploración, y cada medida por sí misma captura una componente única de la similitud entre las ruta de escaneo. Esta técnica proporciona más detalles que otros métodos sobre el tipo de similitud de dos trayectos de exploración según los índices elegidos para realizar la comparación. Además, el método Multimatch se ocupa de la información temporal, no solo de la duración de las fijaciones, sino que también aborda los cambios en tiempo y longitudes de la ruta de exploración. Adicionalmente, las rutas se pueden comparar incluso si tienen una longitud diferente.

En el apéndice G se incluye una descripción del procedimiento del método Multimatch y cuáles son las dimensiones de comparación de secuencias vectoriales que considera.

Al finalizar el proceso Multimatch se tienen cinco índices normalizados que miden las diferencias entre cada par de rutas de escaneo en cinco aspectos o dimensiones distintas. Como resultado, las cinco medidas de similitud tienen rango $[0, 1]$ y los valores más altos de esas medidas indican una mayor similitud entre las rutas de exploración en la dimensión considerada.

6.2.1. Agrupamiento con Multimatch

Volviendo al objetivo de nuestro trabajo de agrupamiento de datos de seguimiento ocular durante la resolución del TMT-A y dado que el método Multimatch proporciona medidas multidimensionales de comparación de rutas de escaneo visual, realizaremos un análisis de agrupamiento de nuestros datos de rutas de resolución visual considerando los índices de similaridad de Multimatch, comparando las trayectorias con un enfoque temporal y bidimensional. Así, el análisis de agrupamientos de las resoluciones a partir de los índices producidos por este método representa un aporte importante al estudio de esta tesis, dado que los enfoques anteriores analizan separadamente la variable temporal y las variables posicionales.

A los efectos de visualizar la tendencia a la formación de grupos de las secuencias de fijación y sacadas de estudio a partir de los índices de Multimatch, empleamos la técnica conocida como "Visual assessment of cluster tendency" (VAT). Ésta técnica permite evaluar visualmente si los datos muestran indicios de algún tipo de agrupamiento. La idea del método es sencilla. Se calcula una matriz de similaridad entre los pares de observaciones, luego se reordena esa matriz de forma que las observaciones similares están situadas cerca unas de otras, es decir, se obtiene una matriz de similaridad ordenada y ésta se representa gráficamente empleando un gradiente de color de acuerdo al valor de las distancias. Si existen

agrupamientos subyacentes en los datos, se forma un patrón de bloques cuadrados en la representación gráfica de la matriz.

Así, el primer paso de nuestro análisis de agrupamiento utilizando Multimatch fue obtener los valores de los índices de similaridad propuestos por el método para cada par de scanpaths que se tiene en estudio. Con los valores de índices obtenidos se construyeron las cinco matrices de similaridad, una en correspondencia con cada índice del método y se sometieron al método VAT para observar la tendencia al agrupamiento de los datos con este enfoque.

En este punto cabe recordar que cada medida de similaridad, y por ende cada matriz construída, recoge la información de una dimensión diferente de comparación de los scanpaths (matrices de comparación de forma de la ruta de escaneo, de dirección y longitud de sacadas, de posición y duración de fijaciones).

En las Figuras 6.2 y 6.3 se muestran los mapas de calor de las cinco matrices asociadas a las medidas de similitud del método Multimatch y los dendogramas correspondientes a los agrupamientos jerárquicos aglomerativos obtenidos con los dos tipos de enlaces (linkages) entre los del tipo single, complete, average, weighted, centroid, median y ward, que dieron los agrupamientos más significativos para nuestros datos.

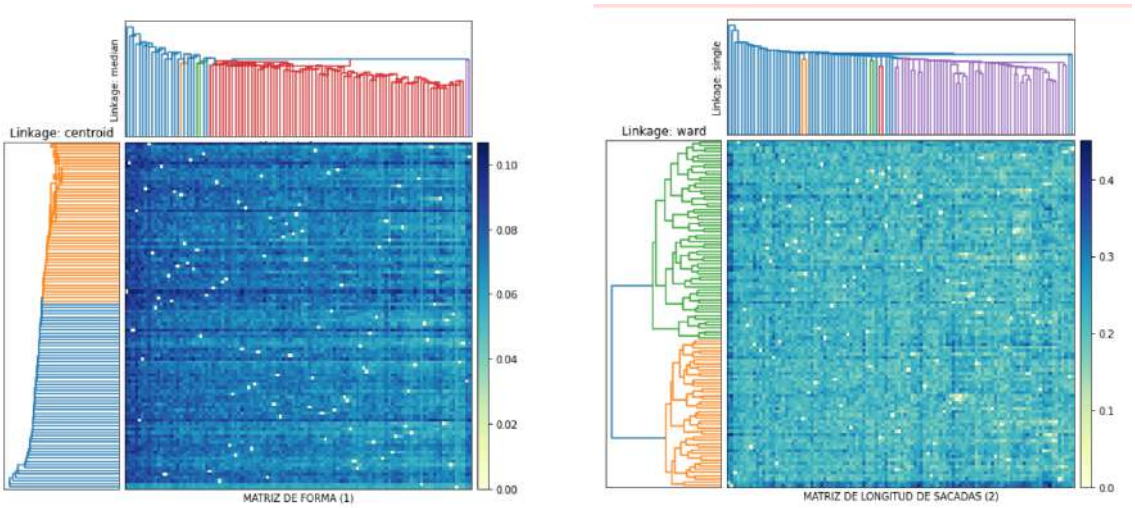


Figura 6.2: Mapa de calor y dendogramas de la matrices de forma y duración de fijaciones

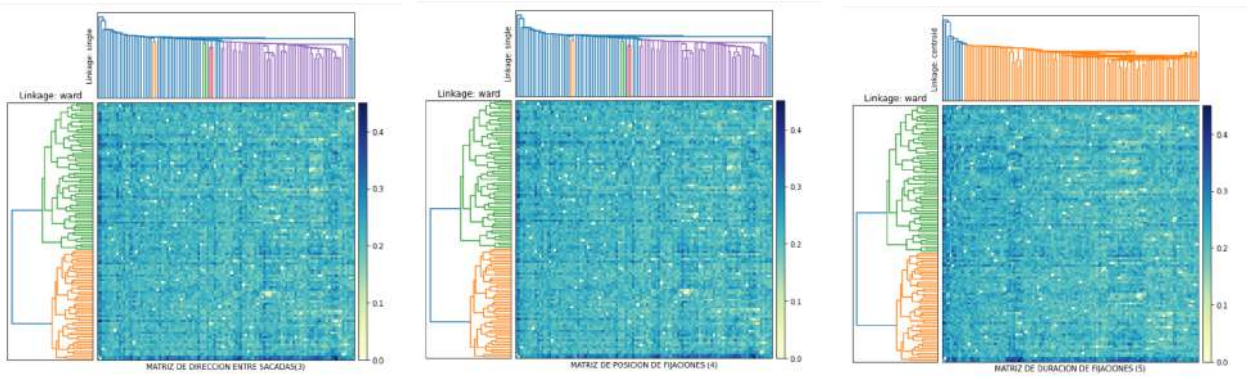


Figura 6.3: Mapa de calor y dendogramas de las matrices de dirección entre sacadas, posición de fijaciones y longitud de sacadas

El primer comentario que puede hacerse a la vista de estos gráficos es que el agrupamiento de los 108 scanpaths en relación a los índices de Multimatch puede realizarse en dos grupos (en general bastante balanceados en número de elementos) o, eventualmente con algún grupo minoritario, en tres clusters.

También se observa que el índice de similitud que se ocupa de comparar la forma general de las

trayectorias, que esencialmente mide diferencias entre pares de sacadas alineadas, es considerablemente mayor que los índices de similaridad respecto a las demás características distinguidas del método Multimatch en todos los pares, es decir, las similitudes de forma entre los scanpaths de análisis son considerablemente mayores que las otras similitudes analizadas con el método Multimatch en los datos de estudio. En consecuencia, los agrupamientos producidos por los linkages utilizados en relación a la matriz de forma separan a los scanpaths en dos grupos de una manera más clara y con un mayor balance de clase que en los otros casos.

Es de destacar también el hecho de que con los enlaces ward y single, los índices correspondientes a longitud de sacadas, dirección de sacadas y posición de fijaciones agrupan a los datos de la misma manera, esto lleva a pensar que estos tres índices captan iguales características de los 108 scanpaths que son a su vez, diferentes de las características que leen los índices de forma de la trayectoria y duración de las fijaciones visuales.

Análisis de los grupos

Con el objetivo de interpretar los agrupamientos generados por el método Multimatch de manera similar a lo realizado en el caso de los agrupamientos anteriores, es decir, utilizando boxplots de variables de referencia por grupos, se obtuvieron variables de significancia en este enfoque vectorial que considera a la trayectoria visual como una secuencia de fijaciones y sacadas, para observar posibles caracterizaciones de los grupos en relación a esas variables.

Con este propósito, luego de la aplicación del algoritmo de clasificación de fijaciones y sacadas diseñado por el grupo Neufisur, para cada una de las trayectorias recorridas durante la resolución del test por parte de los individuos que están bajo estudio en este trabajo de tesis, se consideraron diversas variables descriptivas que usualmente se consideran al trabajar con datos de seguimiento ocular junto con la idea de fijación y sacada, y se ensayó un análisis de clustering de las resoluciones en función de estas variables. Dichas variables se detallan a continuación.

- **Tiempo total** de resolución del test.
- **Número de fijaciones** (NFij).
- **Mediana de la duración de las fijaciones** (DMFij).
- **Desvío estándar de duración de las fijaciones** (DFij desv).
- **Mediana de los desvíos de la abscisa de los puntos que forman la fijación** (xMAD). Este valor refleja la mediana del ancho de las fijaciones.
- **Mediana de los desvíos en la ordenada de los puntos que forman la fijación** (yMAD). Este valor refleja la mediana del alto de las fijaciones.
- **Mediana de las medianas de la abscisa de los puntos que forman la fijación** (xmed). Este valor refleja la mediana en x de los puntos de las fijaciones.
- **Mediana de las medianas de la ordenada de los puntos que forman la fijación** (ymed). Este valor refleja la mediana en y de los puntos de las fijaciones.
- **Mediana de la longitud recorrida con la secuencia de fijaciones** (TrayMFij).
- **Desvío estándar de la longitud recorrida con la secuencia de fijaciones** (TrayFij desv).
- **Mediana de la duración temporal de las sacadas** (DurMSac).
- **Desvío estándar de la duración de las sacadas** (DurSac desv).
- **Mediana de la amplitud o longitud de las sacadas** (AmplMSac).
- **Desvío estándar de la amplitud de las sacadas** (AmplSac desv).

- **Mediana de la abscisa del desplazamiento de las sacadas** (MDx).
- **Desvío de la abscisa del desplazamiento de las sacadas** (Dx desv).
- **Mediana de la ordenada del desplazamiento de las sacadas** (MDy).
- **Desvío de la ordenada del desplazamiento de las sacadas** (Dy desv).
- **Mediana de la velocidad de las sacadas** (VelMSac).
- **Desvío estándar de la velocidad de las sacadas** (VelSac desv).
- **Mediana del ángulo de las sacadas** (AngMSac).
- **Desvío del ángulo de las sacadas** (AngSac desv).

Luego de definir y obtener las variables mencionadas, se llevó a cabo un estudio de correlación entre esas variables descriptivas de fijaciones y sacadas y se observaron varios conjuntos de variables fuertemente correlacionadas como puede verse en la Figura 6.4.

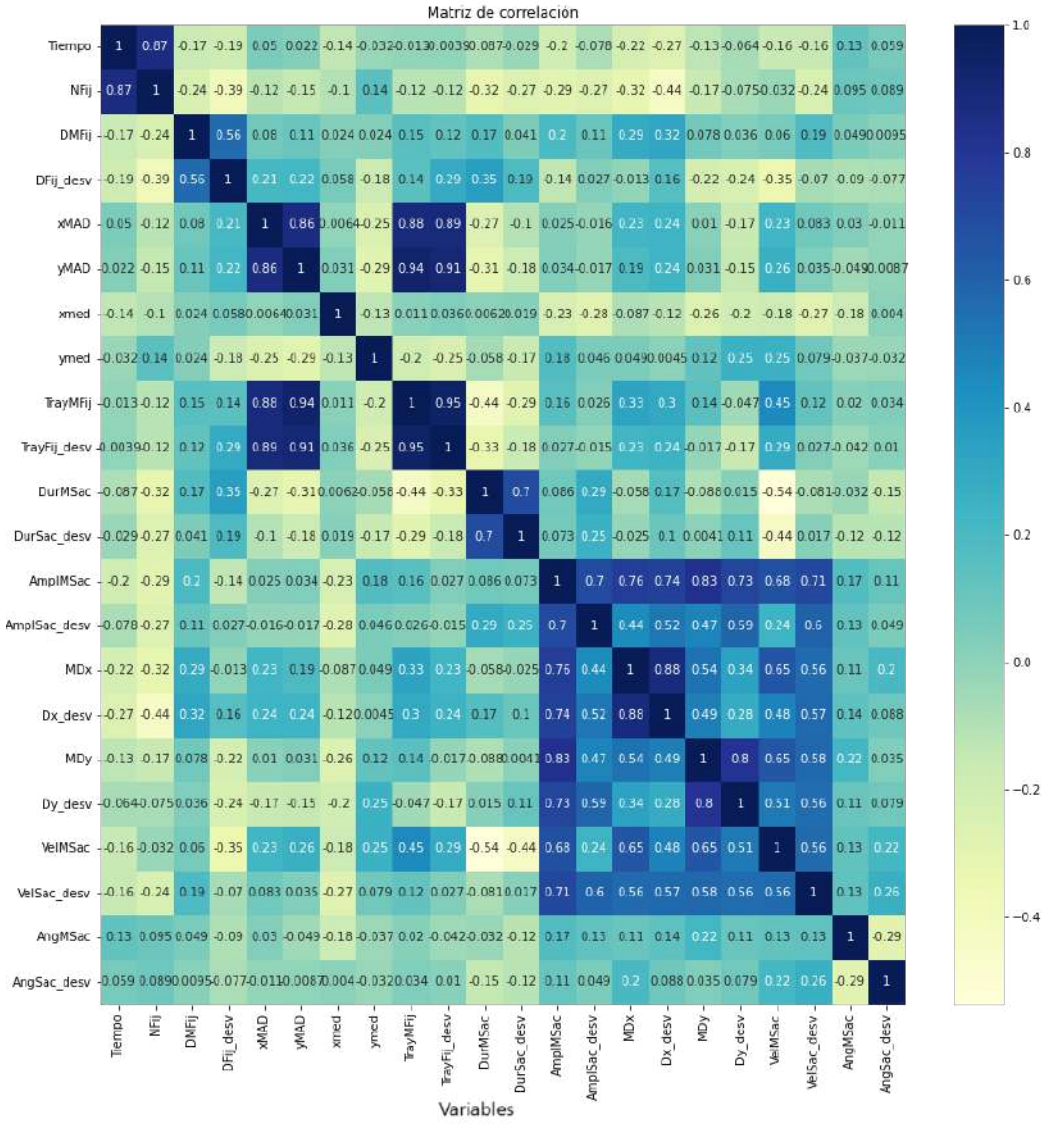


Figura 6.4: Mapa de calor de correlación cruzada de variables

Por lo tanto, se continuó con el estudio de los agrupamientos originados con el método Multimatch en relación a estas variables, considerando solamente una variable de cada grupo altamente correlacionado,

concretamente no considerando las variables Tiempo, Mediana de la duración de las fijaciones, Mediana de los desvíos en y de los puntos que forman la fijación, Desvío estándar del camino recorrido en las fijaciones, Mediana del camino recorrido en las fijaciones, Desvío estándar de la duración de las sacadas, Desvío estándar de la amplitud de las sacadas, Mediana de la componente x del desplazamiento de las sacadas, desvío de la componente x del desplazamiento de las sacadas, Mediana de la componente y del desplazamiento de las sacadas, Desvío de la componente y del desplazamiento de las sacadas, Mediana de la velocidad de las sacadas, y Desvío estándar de la velocidad de las sacadas.

Los resultados del análisis de los grupos generados por Multimatch con respecto a estas variables derivadas de fijaciones y sacadas fueron los siguientes: los grupos obtenidos con el enlace ward a partir de la similaridad de forma y la de dirección de sacadas se diferencian en el valor medio de velocidades medias puntuales bidimensionales; los clusterings generados con la matriz de longitud de sacadas muestran diferencias en sus valores de la variable tiempo de resolución y valor medio de velocidades medias puntuales verticales en y; también se obtienen grupos con diferencia en los valores de la media de velocidades medias puntuales en y al originar el agrupamiento a partir de la similaridad de posición de fijaciones y una no tan notoria diferencia en el valor medio de velocidades medias puntuales bidimensionales.

Por otro lado, teniendo en cuenta que los índices de Multimatch refieren a las características de la trayectoria considerada como una secuencia vectorial de fijaciones y sacadas, se realizó un análisis de los agrupamientos obtenidos con este método en relación a las variables usuales de estudio de fijaciones y sacadas de un scanpath que se detallaron al comienzo de esta subsección. Lo que pudo observarse es que al clusterizar en dos grupos con la similaridad de forma del scanpath y el enlace ward se distinguen diferencias en las variables desvío de amplitud de sacadas y velocidad media de las sacadas; al considerar la matriz asociada a las longitudes de las sacadas, se observan diferencias en los grupos respecto a los valores de la variable tiempo de resolución y número de fijaciones; el índice asociado a la dirección de las sacadas arroja diferencias en los grupos, aunque no muy contundentes, en los valores de las variables mediana de los desvíos de las componentes horizontal y vertical (en x e y) de los puntos de una fijación, en la mediana del camino recorrido en una fijación y en el desvío estándar de la amplitud de las sacadas; el agrupamiento en relación a la posición de las fijaciones separa los grupos con diferencias en las variables correspondientes al número de fijaciones y la mediana del desvío de la componente horizontal de los puntos que forman una fijación.

Respecto al agrupamiento de acuerdo al índice de duración de las fijaciones, sus grupos no reflejan diferencias en los valores de ninguna de las variables de fijaciones y sacadas consideradas, lo que es natural dado que la información sobre la duración temporal de las fijaciones que lee Multimatch no está medida por ninguna de las variables elegidas para caracterizar los grupos en relación a fijaciones y sacadas.

A partir de las matrices de índices Multimatch también se llevó a cabo un estudio de agrupamiento utilizando el método Spectral para dos grupos y se observó los siguiente. La matriz de forma origina grupos que tienen diferencias en los valores de las variables mediana de la amplitud de las sacadas y de la abscisa del desplazamiento de las sacadas; en los grupos generados por la matriz de longitud de sacadas se observan diferencias en relación a los valores de las variables tiempo de resolución y número de fijaciones; también tienen diferencias en el número de fijaciones los grupos correspondientes a la matriz que maneja información de la posición de las fijaciones; y el agrupamiento Spectral asociado a la dirección de las sacadas muestra diferencias en los grupos en los valores del desvío de amplitud de sacadas y desvío horizontal de desplazamiento de las sacadas.

6.2.2. Multimatch de datos homogéneos

En esta sección, se describen los resultados obtenidos para los 49 sujetos que se consideraron en la sección 4.11 al aplicarles las mismas técnicas detalladas en la sección anterior a partir de las matrices de similaridad generadas por el método Multimatch.

A este subconjunto de resoluciones de estudio, al igual que para el conjunto de 108 trayectorias original, se le aplicó el método jerárquico aglomerativo utilizando las matrices de índices de Multimatch y obteniendo los dendogramas correspondientes a cada posible método de enlace ("linkage"). Así, se obtuvieron agrupamientos significativos de los scanpaths y se cotejaron los grupos con las variables ya mencionadas que son de interés de los especialistas en seguimiento ocular. Este análisis arrojó los siguientes resultados.

En el caso de la matriz comparativa de la forma general de la ruta de escaneo, al aplicársele los métodos de agrupamiento (en dos y tres grupos) Spectral y aglomerativo con distancia euclídea y enlaces ward y completo, y analizar sus gráficos de boxplots se observó que tres de los agrupamientos (espectral, aglomerativo con ward y con completo en dos grupos) agrupan a los sujetos según sus valores en las variables longitud total de la trayectoria, velocidad media y velocidad media de las sacadas, como puede verse en el gráfico de boxplot de la Figura 6.5 para uno de esos métodos.

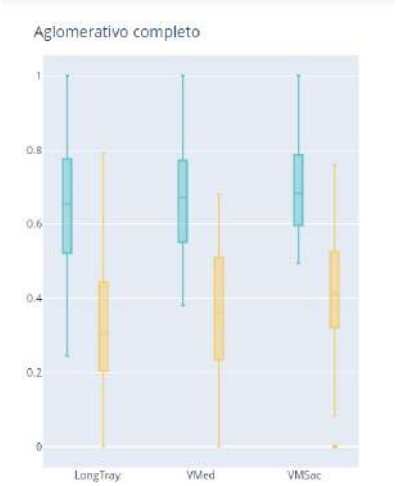


Figura 6.5: Variables separadoras en grupos según forma de la ruta de escaneo visual

Al considerar los valores de similitud dados por el índice de comparación de la longitud de las sacadas de la trayectoria visual, se notó especialmente el caso del método espectral en dos grupos que separa claramente a las resoluciones de los individuos en relación a los valores de sus variables tiempo de resolución, número de fijaciones, longitud de la trayectoria, mediana y desvío mediano absoluto de la duración de las fijaciones y cantidad de fijaciones fuera de los números del test, lo que puede verse en sus gráficos de boxplots como, por ejemplo, el expuesto en la Figura 6.6.

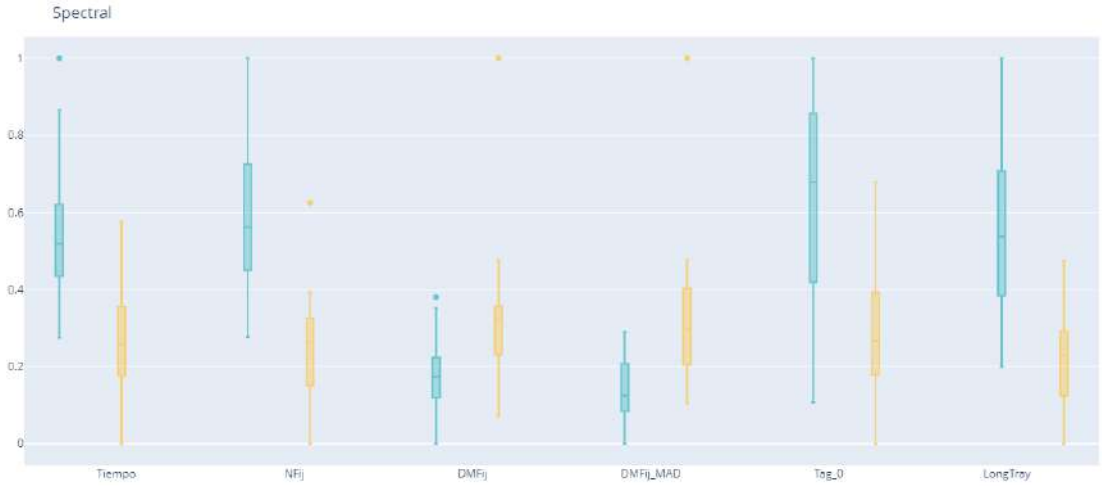


Figura 6.6: Variables separadoras en método espectral de la matriz correspondiente a longitud de sacadas

Luego de proceder al agrupamiento aglomerativo con métrica euclídea y enlace ward a partir de la matriz asociada al índice de medición de dirección de las sacadas de Multimatch y analizar los valores de las variables de posible caracterización de los grupos en este caso, se detecta que al agrupar en dos grupos las diferencias apreciables tienen que ver con valores de las variables velocidad media de desplazamiento y mediana de la velocidad de las sacadas, y en el caso de tres conglomerados se distinguen los tres grupos por sus valores en exactamente las mismas variables, hecho que puede visualizarse en el gráfico de boxplot de la Figura 6.7.

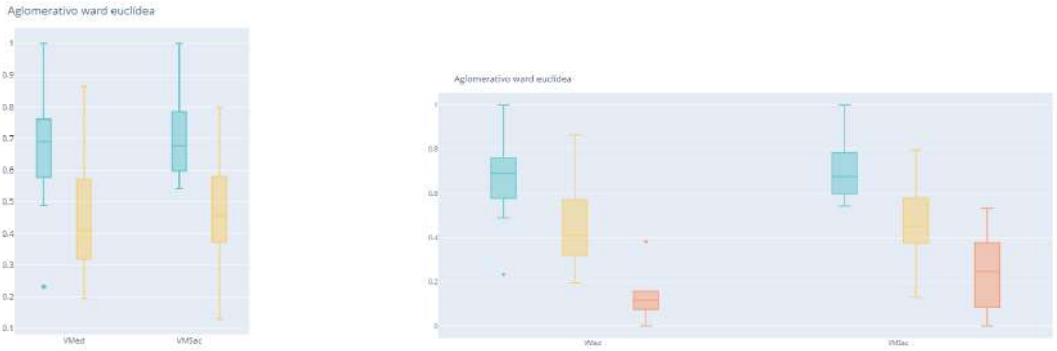


Figura 6.7: Gráficos de boxplots de los grupos de la matriz de dirección de sacadas

Al aplicar los métodos de agrupamiento mencionados para dos grupos a la matriz comparativa de trayectorias según la posición de las fijaciones de la mirada de los individuos durante el test, se ven diferencias intergrupo en los valores de las variables tiempo de resolución de la prueba, número de fijaciones y longitud total de la trayectoria, como puede verse en la Figura 6.8.

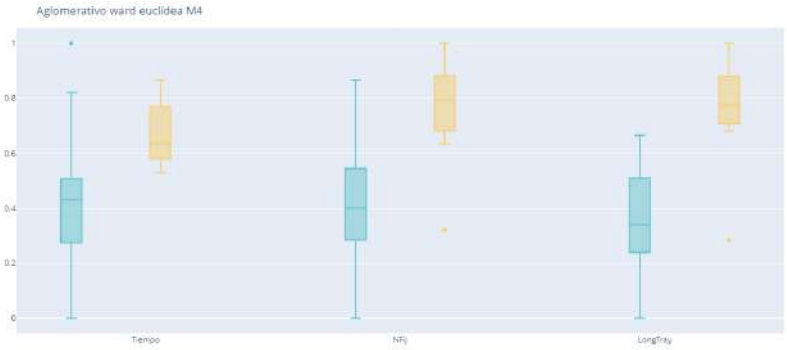


Figura 6.8: Boxplots de agrupamientos de la matriz correspondiente a la posición de fijaciones

Considerando la posibilidad de tres conglomerados de individuos a partir de la misma matriz de posición de fijaciones, el método que arroja algo de claridad en la separación de los grupos de acuerdo a las variables estudiadas es el método aglomerativo también con enlace ward y distancia euclídea, que muestra separaciones de los tres grupos de acuerdo a la longitud de la trayectoria. El gráfico de boxplots de este agrupamiento es el de la Figura 6.9.

Analizando los agrupamientos de la matriz que contiene la información comparativa de resoluciones brindada por el índice de similaridad asociado a la duración de sus fijaciones durante el test TMT-A, se observa que separando los individuos en tres grupos el método Spectral es el que muestra diferencias en los grupos respecto a las medianas de velocidades de resolución y a las medianas de velocidades de sacadas. Este hecho puede verse en la Figura 6.10.

Si se consideran dos grupos en este caso, el método aglomerativo con distancia euclídea, ya sea con enlace ward o completo, también muestra diferencias en los grupos de acuerdo a los valores de las variables medianas de velocidades de resolución y de velocidades de sacadas.

Por otro lado, más allá de la separación con respecto a las variables derivadas que fueron indicadas en lo anterior, en este enfoque Multimatch, la formación de grupos en función de cada uno de los índices

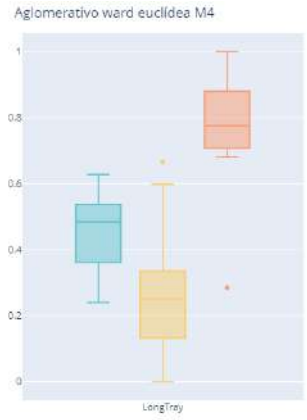


Figura 6.9: Variables separadoras del agrupamiento de la matriz asociada a la comparación de acuerdo a la posición de las fijaciones

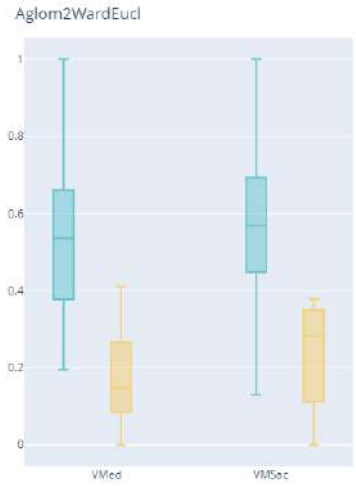


Figura 6.10: Variables destacadas en grupos de acuerdo a duración de las fijaciones

tiene la interpretación natural dada justamente por el índice de la matriz de similaridad que origina la agrupación. Es decir, las resoluciones del test que pertenecen al mismo grupo, por ejemplo asociado al índice de forma, son las que tienen características más parecidas en relación al aspecto general de la secuencia de vectores fijación-sacada. En otras palabras, a diferencia de los otros métodos utilizados en esta tesis, en el enfoque Multimatch de alguna manera se conoce el criterio de agrupamiento de cada matriz y, por lo tanto, las características distintivas de cada grupo respecto a los demás grupos del mismo índice de agrupamiento están explicitadas por la propia definición del índice de Multimatch.

Capítulo 7

Aportes del trabajo

En este trabajo de tesis de maestría se han obtenido diversos agrupamientos de datos de seguimiento ocular durante la realización del test TMT-A, desde distintos enfoques. Se han conformado agrupamientos de los datos según variables definidas a partir de las mediciones del dispositivo de eye tracking, relacionadas a la visión integral de la trayectoria espacio-temporal de la resolución (Sección 4); agrupamientos de acuerdo a la posición de la mirada en cada momento de la medición durante el recorrido visual (Subsección 5.1); grupos asociados al tiempo de resolución de cada búsqueda particular de números del test (Subsección 5.2.1); grupos asociados a los tiempos empleados en la resolución de zonas distinguidas del test (Subsección 5.2.2) y agrupamientos obtenidos considerando a la trayectoria visual como una secuencia vectorial de fijaciones y sacadas (Sección 6), según distintas características de similaridad entre las trayectorias (de forma general, de dirección y longitud de las sacadas, y de posición y duración de las fijaciones).

De esta manera, se ha llevado a cabo un estudio de clustering de datos de seguimiento ocular que brinda un amplio horizonte de agrupamiento de datos de este tipo desde distintas ópticas, permitiendo la incorporación de nuevas trayectorias visuales de resolución del TMT-A a los grupos obtenidos, en función de la información que sea de interés en el análisis de ese individuo o de esa resolución en particular.

Por otra parte, los métodos de detección de trayectorias anómalas que hemos llevado a cabo en la Sección 4.7 han demostrado ser de utilidad para revelar registros atípicos en los datos de seguimiento ocular en casos de observaciones con errores de medición, escaneos visuales de resolución sesgados a ciertos lugares de la pantalla, resoluciones incorrectas o incompletas del test, y trayectorias de resolución que han empleado tiempos atípicamente largos o extemadamente cortos. Todas estas características surgen como información que puede ser de interés de los especialistas al analizar el desempeño de un individuo en la prueba.

Cabe mencionar también la importancia e interés que puede tener un estudio como el de esta tesis que incluye un análisis de la información general de las trayectorias visuales de resolución los sujetos, por un lado, y un estudio de la información más detallada que puede aportar un dispositivo de seguimiento ocular electrónico, dado que ambos enfoques pueden brindar herramientas clínicas en distintos contextos. El aporte que pueda hacerse respecto al análisis más grueso de la información global de resolución del test (tiempos, velocidades, etc.) será de mayor utilidad para los profesionales que no tengan acceso a un dispositivo para realizar seguimiento ocular, y el análisis más detallado de los datos con una buena cantidad de información espacial puede arrojar información útil para aquellos que sí cuenten con un eye tracker en el consultorio y posean esa información. Además, el análisis realizado en este trabajo se adapta perfectamente al uso de eye trackers de bajo costo que podrían ser incorporados sin mucho esfuerzo económico en consultorios psicopedagógicos y psicológicos ya que, en definitiva, la principal motivación de procesar la información del seguimiento ocular es mejorar la evaluación provista por el test utilizado comúnmente en pruebas neuropsicológicas, considerando que el análisis de los datos del TMT digitalizado puede convertirse en una herramienta poderosa para colaborar en el diagnóstico de una gran variedad de patologías mediante la construcción de nuevas medidas de diagnóstico clínico.

Capítulo 8

Trabajo futuro

Algunas líneas de investigación futura que pueden continuar el trabajo de esta tesis son las siguientes:

- Estudiar índices de correlación en las trayectorias vistas como series temporales, ya sea en el enfoque de series de posiciones, series de trials o series por zonas.
- Realizar una comparación de los valores de las variables que caracterizan a los distintos grupos en los agrupamientos con los valores de las mismas variables de la solución ideal (los mínimos movimientos visuales necesarios para recorrer todos los números del test conociendo la posición de los números de antemano), por ejemplo, calculando distancias de cada resolución a esa resolución ideal.
- Estudiar las anomalías en cada serie de tiempo para detectar posibles dificultades de cada sujeto en hallar un número o dificultades colectivas en alguna zona del test.
- Analizar las coocurrencias de las diferentes trayectorias de individuos en los grupos formados por los distintos agrupamientos, es decir, sujetos que están agrupados frecuentemente juntos en los distintos enfoques.
- Llevar a cabo nuevos procesos de ensambles de los agrupamientos obtenidos.
- Realizar estudios similares a los de esta tesis para el caso del test TMT-B y analizar diferencias entre ambas pruebas.

Apéndice A

Métodos de selección de variables

Principal Feature Analysis (PFA). El método de Análisis Principal de Variables fue propuesto en [25] por Lu et al. y utiliza el mismo criterio que el de componentes principales. Sin embargo, éste es un método más eficiente computacionalmente que el Análisis de Componentes Principales (PCA) para reducir la dimensión de un conjunto de variables. El método se trata de eligir un subconjunto del conjunto original de atributos que contenga la mayoría de la información esencial. La idea de esencial se refiere tanto al sentido de maximizar la variabilidad del atributo en el espacio de menor dimensión como a minimizar el error de reconstrucción del problema.

El punto de principal ventaja de PFA sobre PCA es que requiere menos sensores o menos funciones a calcular y el subconjunto de variables elegido mantiene las propiedades de optimalidad del método PCA, conservando el significado de las variables originales.

LASSO (Least Absolute Shrinkage and Selection Operator). Este método fue introducido por primera vez en [39] por Robert Tibshirani. Está basado en una fórmula estadística cuyo propósito principal es la selección de variables y regularización de modelos de datos. El método LASSO regulariza los parámetros del modelo reduciendo los coeficientes de regresión, incluso reduciendo algunos de ellos a cero. Si dos atributos están correlacionados linealmente, su presencia simultánea aumentará el valor de la función de costo, por lo que la regresión de Lasso intentará reducir el coeficiente de la variable menos importante a cero para seleccionar las mejores variables. La reducción de coeficientes de LASSO aumenta la precisión de la predicción del modelo ya que reduce la varianza y minimiza el sesgo, es una herramienta útil para eliminar todas las variables que son irrelevantes y que no están relacionadas con la variable de respuesta.

Spectral feature selection (SPEC). Esta técnica estudia la manera de seleccionar variables de acuerdo a la estructura del grafo inducido por un conjunto de similaridades de instancias consideradas dos a dos. El método SPEC evalúa la relevancia de un atributo vía la medición de su capacidad de preservar la similaridad de la muestra especificada previamente. Más concretamente, asumiendo las similitudes entre cada par de muestras almacenadas en una matriz, la técnica de selección de variables espectral estima la relevancia del atributo midiendo la consistencia de las variables con el espectro de una matriz derivada de la matriz de similitud. Más detalles del método pueden consultarse en [45].

Logistic Regression. La regresión logística puede considerarse una extensión de los modelos de regresión lineal, con la particularidad de que el dominio de salida de la función está acotado al intervalo $[0, 1]$ y que el procedimiento de estimación, en lugar de mínimos cuadrados, utiliza el procedimiento de estimación de máxima verosimilitud. Este método se utiliza para comprender la relación entre la variable dependiente y una o más variables independientes mediante la estimación de probabilidades con una ecuación de regresión logística. Un modelo de regresión logística puede utilizarse para predecir las probabilidades de las clases sobre la base de las variables de entrada, luego de ordenarlas de acuerdo a un ranking de importancia relativa. Est técnica selecciona las variables a considerar basada en una puntuación del atributo en ese ranking de importancia relativa respecto al resultado del aprendizaje.

Aunque la técnica de regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo y la asignación final se hace en función de las probabilidades predichas. Más detalles del método pueden consultarse en [28].

Apéndice B

Técnicas de detección de atípicos

One Class Support Vector Machine. El concepto de máquinas de soporte vectorial de una clase utilizado para la detección de valores atípicos fue propuesto en 1999 en [36]. El enfoque se basa en la idea de que los datos se separan en una única categoría y los outliers son detectados separados del resto de las observaciones. Básicamente la técnica separa del origen todos los registros y maximiza la distancia del hiperplano de los datos al origen. El modelo está entrenado para descubrir los parámetros del hiperplano que maximiza la distancia desde el origen. Se espera que todas las muestras de un lado del hiperplano sean "inliers", mientras que todas las muestras restantes se consideran valores atípicos. La construcción del hiperplano resulta en una función binaria que captura las regiones del espacio de muestras donde yace la densidad de probabilidad de los datos, asignando el valor 1 a una "pequeña" región (que captura los puntos de entrenamiento) y -1 en otro caso. Las observaciones que sean asignadas con el valor -1 se consideran atípicos fuera de la distribución de los datos. El método de máquinas de soporte vectorial de una clase es capaz de ayudar a identificar los valores atípicos con una probabilidad de error muy pequeña.

Isolation Forest. Este método, propuesto en [40], es un método de detección de anomalías muy potente que se basa en un marco general de aprendizaje en conjunto. Los árboles de decisión son un método práctico de particionar un conjunto de datos. Comenzando desde la raíz, para cada nodo, se seleccionan una variable y un umbral, y las muestras se dividen en dos subconjuntos. El método toma ventaja de dos propiedades cuantitativas de las observaciones atípicas: son unas pocas instancias minoritarias y tiene valores de los atributos que son muy diferentes de los de las instancias normales. Estas características hacen a las anomalías más susceptibles de estar aisladas que los puntos normales y una estructura de árbol puede aislar efectivamente una instancia cerca de la raíz del árbol de decisión, mientras que los puntos normales son aislados más profundamente al final del árbol. Estas propiedades de los árboles son la base del método para detectar atípicos conocido como Isolation Tree o Isolation Forest. Este método construye un ensamble de árboles para un dataset dado en el que las anomalías son aquellas instancias que tiene camino promedio de corta longitud en el árbol. En efecto, como es posible observar, los inliers normalmente pertenecen a regiones de alta densidad que requieren más partición para aislar la muestra. Por el contrario, los valores atípicos que se encuentran en regiones de baja densidad se pueden detectar con menos pasos de partición porque la granularidad que se requiere es proporcional a la densidad de los conglomerados. Por lo tanto, se construye un Isolation Forest con el objetivo de medir la longitud de ruta promedio para todos los inliers y compararlo con el requerido por todas las muestras. Cuando dicha longitud es más corta, la probabilidad de ser un valor atípico aumenta. Por lo tanto, después de entrenar el Isolation Forest, se puede calcular una puntuación de anomalía considerando la longitud del camino promedio para una muestra dada. Cuando tal puntaje es cercano a 1, podemos concluir que la probabilidad de una anomalía es también muy grande. Por el contrario, valores de puntuación muy pequeños indican que las observaciones son en cambio posibles inliers.

Gaussian mixtures. El método de mezcla gaussiana, cuyos detalles pueden consultarse en [34], es un modelo en el que se considera que las observaciones siguen una distribución probabilística formada

por la combinación de múltiples distribuciones normales. El método puede entenderse como una generalización de K-means con la que, en lugar de asignar cada observación a un único cluster, se obtiene una distribución de probabilidad de pertenencia a cada uno. Ajustar un modelo de mezcla gaussiana consiste en estimar los parámetros que definen la función de distribución de cada componente normal. Usar un modelo de mezcla gaussiana para la detección de anomalías es bastante simple: una vez aprendidos los parámetros, se puede calcular la densidad de probabilidad que tiene cada observación de pertenecer a cada componente y al conjunto de la distribución. Observaciones con muy poca densidad de probabilidad pueden considerarse como anomalías, es decir, cualquier punto muy alejado de las funciones gaussianas puede considerarse un atípico. Debe definirse qué umbral de densidad se desea utilizar. Esta técnica puede ayudar a detectar un comportamiento poco probable o lejano del comportamiento nominal de las observaciones.

Apéndice C

Métodos para determinar el número de grupos

Método del codo. Uno de los métodos que se utilizan para determinar el número óptimo de conglomerados, ajustando el modelo con un rango de valores para el número de grupos del agrupamiento por K-Medias, es el conocido como "Elbow method" o método del codo. Este método traza el valor de la función de costo producida por diferentes valores del número k de grupos. El método calcula y grafica una puntuación promedio para todas las agrupaciones según el número de grupos k. Si k aumenta, la distorsión promedio disminuirá, cada grupo tendrá menos instancias constituyentes y las instancias estarán más cerca de sus respectivos centroides. Sin embargo, las mejoras en la distorsión promedio disminuirán a medida que aumente k. Se denomina "codo" al valor de k en el que la mejora en la distorsión disminuye más, y es en el valor que debemos dejar de dividir los datos en grupos adicionales. De esta manera, si el gráfico de líneas se asemeja a un brazo, entonces el "codo" (el punto de inflexión en la curva) es una buena indicación de que el modelo subyacente se ajusta mejor en ese punto.

Gap statistics. Es otro método para aproximar el número "correcto" de grupos. La idea básica de este método es elegir el número de grupos en el que ocurre el mayor salto en la distancia dentro del grupo, en función del comportamiento general de muestras extraídas uniformemente. Puede tenerse una idea de qué tan bueno es cada combinación de valores de k y su error asociado, si se puede comparar el error de nuestro agrupamiento con el error esperado para el mismo k bajo una distribución de referencia nula. En esencia, el método consiste en encontrar el valor de k para el cual la diferencia entre el error de nuestra agrupación y el esperado bajo una distribución nula es mayor. Este proceso se lleva a cabo mediante una métrica de error (la suma de cuadrados de distancias dentro del grupo) con respecto a nuestra elección del número de grupos k.

Criterio de información Bayesiano ("Bayesian information criterion" o BIC). Es un criterio de selección de modelos basado parcialmente en la función de probabilidad, que equilibra el número de parámetros del modelo y el número de datos frente a la función de máxima verosimilitud. El método busca encontrar el número de parámetros del modelo que minimiza el BIC, en nuestro caso, el número de grupos que minimiza el índice BIC.

Mezcla Gaussiana Bayesiana. Esta técnica aplicada para hallar el número óptimo de grupos es capaz de dar ponderaciones iguales o cercanas a cero para los clústeres innecesarios, es decir, el algoritmo permite la eliminación de los grupos innecesarios automáticamente.

Rendimiento o "performance". Un procedimiento usual para decidir el mejor número de grupos de un análisis de clustering es comparar las características de los agrupamientos considerando diferente número de grupos y realizando la elección del posible número adecuado de clusters para nuestro conjunto de datos, asistidos por un estudio de la calidad del agrupamiento que se obtiene al considerar con una misma técnica de clustering diferente número de grupos para el mismo dataset.

Comprender el desempeño de los métodos de aprendizaje no supervisados es inherentemente mucho más difícil que comprender el de los métodos de aprendizaje supervisado porque, a menudo, no hay "mejor" solución clara. Para el aprendizaje supervisado hay muchas métricas de resultados sólidos pero, desafortunadamente, para los métodos de agrupamiento no tenemos etiquetas en las que confiar y necesitamos comprender cuán "diferentes" son nuestros grupos. En este contexto, se piensa en qué propiedades tendría idealmente un buen clustering, grupos compactos y bien separados unos de otros. La evaluación de resultados está asociada al uso de índices de validez que realicen una buena comparación entre distintos agrupamientos, índices que se utilizan para medir la calidad de los resultados de la agrupación. Existen varios de éstos índices, pueden encontrarse más de 30 de ellos en la literatura, que intentan hacer estas mediciones basados en el agrupamiento modelo a partir de los valores de atributos y las medidas de un buen clustering, y varios estudios han concluido que algunos de éstos índices de validación se desempeñan bien en un amplio rango de problemas. Utilizaremos algunos de ellos en nuestro trabajo, por lo que comentaremos a continuación brevemente de qué se trata cada uno de ellos.

El método más común para evaluar el rendimiento de un algoritmo de agrupamiento es el método del **índice de la silueta**. Esta técnica proporciona un índice por instancia y una representación gráfica global que muestra el nivel de cohesión interna y de separación de los clusters. La métrica de Silhouette funciona analizando qué tan bien encaja un punto dentro de su grupo y mide la máxima varianza de grupo. El coeficiente de la silueta varía de -1 a 1 e, idealmente, si la puntuación de silueta promedio en su agrupación es 1, entonces se habrán logrado agrupamientos perfectos y habrá una mínima confusión sobre qué punto pertenece a qué grupo. Un coeficiente cercano a 0 significa que la muestra está cerca de un límite de clúster y un índice cercano a –1 significa que la instancia puede haber sido asignada al clúster incorrecto.

El **índice de homogeneidad** u "Homogenity Score" es un método que mide la superposición de grupos y se basa en el supuesto de que un conglomerado debe contener solo muestras que tengan exactamente la misma etiqueta.

El **índice Davies-Bouldin** se define como la medida de similitud promedio de cada grupo con su grupo más similar, donde la similitud es la relación entre las distancias dentro del grupo y las distancias entre grupos. Este método calcula el máximo de las distancias intergrupo e intragrupo. Los grupos que están más separados y menos dispersos darán como resultado una mejor puntuación. La puntuación mínima es cero y los valores más bajos indican un mejor agrupamiento.

El **índice Calinski-Harabasz**, también conocido como "Variance Ratio Criterion" o criterio de relación de varianza, es una medida de qué tan similar es un objeto a los de su propio grupo (cohesión) en comparación con los de otros grupos (separación). La cohesión se estima en función de las distancias desde los puntos de datos en un grupo a su centroide de grupo y la separación se basa en la distancia de los centroides de grupo desde el centroide global. El índice mide el radio de distancia intergrupo e intragrupo y se define como la relación entre la varianza o dispersión dentro de un grupo y la dispersión entre grupos. Un valor más alto del índice significa que los conglomerados son densos y están bien separados.

Apéndice D

Técnicas de agrupamiento

K-Medias (o K-Means). El algoritmo K-Means o K-Medias es un algoritmo simple, propuesto por Stuart Lloyd en 1957, capaz de agrupar un conjunto de datos de manera muy rápida y eficiente, a menudo en unas pocas iteraciones.

El número k en el método de k-medias es especificado al inicio de su aplicación, y es el número de centroides a buscar para que el algoritmo asigne cada observación a exactamente uno de los k grupos que hallará el algoritmo a través de cálculos de distancia euclideana por pares en un proceso iterativo. El algoritmo forma los grupos minimizando la variación dentro del clúster (también conocida como inercia) de manera tal que la suma de las variaciones dentro del clúster en todos los k clústeres sea lo más pequeña posible. Este algoritmo tiene dos limitaciones principales: la métrica es siempre euclideana y no es muy robusto a valores atípicos. Por otro lado, es necesario ejecutar el algoritmo varias veces para evitar soluciones subóptimas y debe especificarse el número de grupos al iniciarlo. Además, K-Means no se comporta muy bien cuando los grupos tienen diferentes tamaños, diferentes densidades o formas no esféricas.

Mini Batch K-Means. Una variante importante del algoritmo K-Means, conocido como Mini Batch K-Means, se propuso en 2010 por David Sculley. Este algoritmo es una extensión del método K-Medias estándar pero en lugar de utilizar el conjunto de datos completo en cada iteración y calcular las medias globales, el algoritmo propuesto es capaz de usar mini-lotes, moviendo ligeramente los centroides en cada iteración. Este proceso acelera el algoritmo K-Means y hace posible agrupar enormes conjuntos de datos, pero su inercia es generalmente un poco peor que la de K-Means, especialmente a medida que aumenta el número de conglomerados.

BIRCH (Balanced iterative reducing and clustering using hierarchies). Este algoritmo tiene una dinámica ligeramente más compleja que Mini Batch K-Means y en su parte final emplea un método de agrupamiento jerárquico. Puede ser más rápido que K-Means por lotes, con resultados similares, siempre que el número de funciones no sea demasiado grande.

Agrupamiento Jerárquico. En la agrupación jerárquica se genera una secuencia de configuraciones de agrupamiento que se pueden organizar en la estructura de un árbol. La ventaja de la agrupación jerárquica es que no requiere que el número de clusters esté predefinido. En su lugar, podemos elegir el número de clusters después de ejecutar la agrupación jerárquica. Otra gran utilidad del método jerárquico es la posibilidad de representación del dendograma ya que, teniéndolo a la vista, se puede interpretar cómo los puntos de datos se relacionan entre sí y decidir subjetivamente en qué "nivel" deben existir los grupos para determinar dónde cortar el dendograma para elegir el número de clusters en el algoritmo. Sin embargo, una desventaja es que la agrupación jerárquica requiere calcular la matriz de distancia de todos los pares de observaciones, lo que puede ser una operación que requiere mucho tiempo para grandes conjuntos de datos. Las distancias más frecuetemente utilizadas son la distancia euclideana, la distancia de Manhattan, la distancia de Minkowski (parametrizada con un valor p). Una vez que sea ha elegido una distancia y se ha definido una métrica, el siguiente paso es definir una

estrategia de fusión que se denomina enlace o "linkage". El objetivo de un método de vinculación es descubrir los grupos que deben ser fusionados en uno solo en cada nivel de la jerarquía. Los criterios de vinculación o "linkage" lidian con el concepto de determinar el modo en que se calculan la distancias entre los grupos y depende del tipo de problema a resolver. Las opciones de enlaces más populares son: "linkage average" (enlace promedio o centroide) que refleja la búsqueda de centroide que utiliza K-Means y minimiza la distancia promedio entre grupos considerando todos los pares posibles; "single linkage" (enlace único) cuyo criterio de vinculación es la distancia mínima entre un par de puntos de dos grupos; "complete linkage" (enlace completo) que funciona encontrando la distancia máxima entre un par de puntos entre dos grupos, es decir, combina grupos basado en los puntos más lejanos entre los dos grupos; y el enlace Ward que está basado en la distancia euclideana, tiene en cuenta todos los grupos y dos de ellos son seleccionados con el objetivo de minimizar la suma de las distancias al cuadrado. Determinar qué criterio de linkage es mejor para un problema depende en gran medida del conjunto de datos en particular.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Es otro algoritmo popular de agrupación con un enfoque diferente basado en la estimación de la densidad local que agrupa los datos en función de la densidad de puntos. El enfoque DBSCAN amplía la evaluación de la métrica de distancia incorporando también la noción de densidad y permite al algoritmo identificar grupos de formas arbitrarias. Si hay grupos de datos que existen en la misma área que otros, pueden verse como miembros del mismo grupo.

DBSCAN agrupa puntos cuando están muy juntos, a partir de la elección de una distancia y de un número mínimo de puntos que deben existir dentro de un grupo. En DBSCAN, podemos etiquetar puntos explícitamente como valores atípicos y evitar tener que agruparlos. Comparado con los otros algoritmos de agrupamiento, DBSCAN es mucho menos propenso a la distorsión causada por valores atípicos en los datos, en función de los hiperparámetros elegidos. Otra característica de DBSCAN es que no es necesario dar explícitamente el número de clusters esperados al inicio del método. En comparación con K-Medias y agrupaciones jerárquicas, DBSCAN puede verse como potencialmente más eficiente, ya que solo tiene que mirar cada punto de datos una vez. Si bien DBSCAN es muy poderoso, no es infalible y puede verse como potencialmente exagerado, dependiendo de cómo sean los datos originales. Además, si la densidad varía significativamente entre los grupos, puede ser imposible que pueda capturar todos los clústeres correctamente. Sin embargo, combinado con K-Means y clustering jerárquico, DBSCAN completa una sólida caja de herramientas cuando se trata de la tarea de aprendizaje sin supervisión para el agrupamiento de datos.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Es una versión jerárquica de DBSCAN. El algoritmo toma el algoritmo DBSCAN y lo convierte en un algoritmo de agrupamiento jerárquico. En otras palabras, agrupa según la densidad y luego vincula los clústeres basado en la densidad en función de la distancia, de forma iterativa.

Mean-Shift. Este algoritmo tiene algunas características comunes con DBSCAN, puede encontrar cualquier cantidad de grupos de cualquier forma y tiene solo un hiperparámetro: el llamado ancho de banda. El método se basa en la estimación de la densidad local. A diferencia de DBSCAN, Mean-Shift tiende a particionar los grupos cuando tienen variaciones de densidad interna y es complejo computacionalmente.

K-Medoides. Es un método propuesto en 1987 por Kaufman L. y Rousseeuw P.J. como alternativa a K-Means. Tanto K-Medias como K-Medoides crean agrupaciones basadas en todas las observaciones en los datos, donde la distancia entre los centros de los conglomerados y las observaciones en los conglomerados se minimiza. La principal diferencia entre ellos es la elección de los centroides o centros de conglomerados, que en K-Medias se toman como la media de todas las observaciones en el conglomerado, mientras que los k medoides son centros que son observaciones reales y que siempre pertenecen al conjunto de datos. K-medoides es más robusto a valores atípicos y a veces es más eficiente que K-Means.

Además, este algoritmo no está estrictamente vinculado a la métrica euclideana, por lo tanto, puede aprovechar al máximo el potencial de las funciones de distancia alternativas.

Fuzzy C-Means. Es una variación del algoritmo de K-Medias que asocia a cada muestra un vector que generalmente representa la probabilidad de que esa muestra pertenezca a cada uno de los grupos. En este método, un clúster no se considera una partición mutuamente excluyente, sino un conjunto flexible que puede superponerse a algunos de los otros grupos. Todas las muestras se asignan a todos los clústeres pero un vector de peso determina la nivel de membresía con respecto a cada uno de ellos. Los clusters contiguos pueden definir propiedades parcialmente superpuestas; por lo tanto, con este método una muestra dada puede tener una ponderación no nula para dos o más clústeres.

Mezcla Gaussiana (o "Gaussian mixtures"). La mezcla gaussiana es un modelo o proceso generativo y probabilístico que asume que las instancias se generaron a partir de una mezcla de varias distribuciones gaussianas cuyos parámetros se desconocen. Este método puede considerarse el padre de K-Means, porque la forma en que funciona es muy similar; pero, al contrario de ese algoritmo, dada una muestra y un conjunto de grupos (que se representan como distribuciones gaussianas), proporciona un vector de probabilidad. El modelo está entrenado para maximizar la probabilidad, dado un número predefinido de componentes. En este modelo, cada grupo puede tener una forma elipsoidal diferente, tamaño, densidad y orientación. El modelo de mezcla gaussiana es un método que puede utilizarse para estimar densidad, agrupamiento y detección de anomalías, y funciona muy bien en grupos con formas elipsoidales.

Propagación de afinidad (o "Affinity propagation"). Este algoritmo utiliza un sistema de votación en el que las instancias "votan" por instancias similares para ser sus representantes y una vez que el algoritmo converge, cada representante y sus votantes forman un grupo. Este método puede detectar cualquier número de grupos de diferentes tamaños pero también tiene cierta complejidad computacional.

Agrupación espectral. Es una técnica muy popular que realiza una proyección del conjunto de datos en un nuevo espacio. Este algoritmo toma una matriz de similitud entre las instancias y reduce la dimensionalidad. Luego usa otro algoritmo de agrupamiento en este espacio de baja dimensión. La agrupación espectral puede capturar estructuras de agrupaciones complejas y puede utilizarse para subdivir grupos y resolver problemas no convexos.

K-Shape. Es un algoritmo de agrupamiento especialmente diseñado para series temporales que se basa en un procedimiento iterativo de refinamiento, que crea grupos homegéneos y bien separados (ver [31]). Como medida de distancia, K-Shape usa una versión normalizada de la correlación cruzada para considerar las formas de las series de tiempo mientras las compara. La técnica calcula el centroide del cluster, que es usado en la siguiente iteración para actualizar la asignación de las series a los grupos, y supera a todos los métodos escalables, incluso a algunos no escalables, en términos de precisión.

Dynamic Time Warping. En los casos en que se intenta agrupar las series temporales de acuerdo a la información de su forma, una posible elección de medida de similitud entre series temporales es la conocida como Dynamic Time Warping (o DTW) (ver [1]). Esta medida de similitud no es una métrica de distancia porque no satisface la desigualdad triangular pero aún así tiene el potencial de ser una respuesta viable a la tarea de proporcionar una medida de similaridad flexible e interpretable entre series temporales. Utilizando un estiramiento o una compresión de segmentos de los datos temporales, DTW determina un emparejamiento óptimo entre cualquier par de series temporales según un costo mínimo de alineamiento. De esta manera, las series que exhiben patrones similares en diferentes períodos de tiempo, se consideran similares. Otra característica y ventaja importante de DTW es el hecho de que esta técnica expresa la distancia entre series de distinta longitud, lo que en nuestro caso de estudio es un tema crucial al pensar nuestros datos como series temporales. El método DTW se puede optimizar mediante un parámetro de ancho de banda, en el que se calculan solo los valores de la matriz cercanos a

la diagonal. Esta versión de DTW se llama FastDTW. Una fuerte limitación de Dynamic Time Warping es que no es diferenciable en todas partes debido al operador de mínimo que se usa en los cálculos. Soft-DTW (ver [3]) es una variante del método que se ha introducido como una forma de mitigar esta limitación y propone reemplazar este operador de mínimo por un mínimo suave.

Apéndice E

Métodos de comprensión de clusterings

Caracterización univariada consiste en evaluar la importancia de las variables, tomadas individualmente, en la construcción de la estructura del agrupamiento. La idea es realizar mediciones de la proporción de la varianza de cada variable explicada por la pertenencia a un cierto grupo, la comparación de los valores medios de las variables en cada grupo entre sí y en relación a la media global de la variable, etc. Estas características pueden analizarse mediante los gráficos de boxplots, que es una técnica útil de caracterización univariada para sugerir las propiedades de (dis)similaridad de los distintos grupos. En la visualización de los boxplots por grupos, pueden observarse las cajas y llegar a algunas conclusiones. Sin embargo, los diagramas de caja solo muestran información unidimensional. La información sobre la estructura espacial específica de los datos en el espacio multidimensional (que incluye todas las variables consideradas) no puede observarse claramente en el análisis de boxplots.

Caracterización multivariada: Son técnicas que permiten analizar la interacción entre las distintas variables en cada cluster, que a veces están muy correlacionadas, y qué porcentaje explican de la varianza de las variables cada uno de los grupos.

Visualización en las dimensiones de mayor varianza: La intuición aquí es que para dar significado a un clúster se debe ver qué tan diferente es de otros clústeres. Una forma eficiente de ver en qué se diferencia un grupo de otro es centrarse en los atributos que varían más. Una forma de obtener las variables con más varianza es tomar el resultado del Análisis de Componentes Principales (PCA), que retiene efectivamente la dimensión que varía más y comprime las dimensiones que varían menos.

Aprendizaje automático: La idea aquí es pensar en los grupos como clases y en las variables como atributos de entrada. Esto convierte efectivamente el problema de determinar el significado de un cluster en un problema de clasificación. Luego podemos intentar usar un algoritmo de aprendizaje automático para la clasificación, con el fin de "aprender automáticamente" la relación entre la entrada (variables o features) y la salida (clúster). Interpretar esta relación aprendida entre la entrada y la salida nos dará una idea del significado de los grupos.

Hay varias opciones de algoritmos de aprendizaje automático que pueden utilizarse para este fin. Sin embargo, como nuestro objetivo es interpretar la relación aprendida por la máquina entre la entrada y la salida, es mejor elegir un algoritmo que sea interpretable. Uno de los algoritmos de aprendizaje automático altamente interpretable es el árbol de decisión, que muestra las principales variables que distinguen a los grupos. Al analizar el árbol de decisión, se pueden resaltar las características del grupo correspondiente. Por ejemplo, podríamos generar automáticamente las reglas o condiciones de asignación de las observaciones a los distintos grupos en función de las variables originales y el resultado de la agrupación como etiqueta.

Apéndice F

Algoritmo de clasificación de fijaciones y sacadas

El algoritmo de clasificación de datos que utilizamos en esta tesis es un algoritmo desarrollado por el grupo de investigación NEUFISUR que realiza la clasificación de los registros en fijaciones y sacadas con base en las velocidades de desplazamiento punto a punto. El algoritmo consta de dos etapas de procesamiento de la señal. La primera etapa hace una clasificación global y se basa en las velocidades punto a punto del registro. Se propone un umbral de velocidad que se define a partir del ajuste del histograma de velocidades con media móvil de 3 puntos (el punto en consideración, el anterior y el siguiente), y luego se etiqueta cada punto muestreado según la velocidad media calculada con los tres puntos. Si la velocidad es mayor que el umbral, el punto pertenece a una sacada, si es menor se lo considera parte de una fijación. Este procesamiento se realiza sobre la señal y luego se ajusta mediante la aplicación de algunos criterios. El más importante de ellos es el que determina que si una fijación tiene un solo punto, se reetiqueta como punto de una sacada.

Luego de aplicado el algoritmo de clasificación de puntos registrados en fijaciones y sacadas, para cada serie temporal $\{(x(t), y(t))\}$ se consideran todas las fijaciones y sus correspondientes duraciones de forma que los datos iniciales se transforman en un nuevo conjunto de datos, esto es, un nuevo scanpath $\{(\overline{x}, \overline{y}, T)\}$ en el que el par $(\overline{x}, \overline{y})$ indica la mediana de la posición de cada fijación en píxeles normalizados (PN) (que incluye entre 5 y 30 datos muestreados) y T indica la duración de la fijación en esa posición (en milisegundos).

Apéndice G

Método Multimatch

El procedimiento del Multimatch puede describirse mediante los cinco pasos que se detallan a continuación.

Paso 1: Representación de las rutas de exploración como secuencias vectoriales.

A partir de una ruta de escaneo considerada como una secuencia de fijaciones y sacadas, una sacada idealizada se representa como el vector que une una fijación y la siguiente en la línea de tiempo. Las coordenadas cartesianas de las fijaciones en la pantalla son, por tanto, los puntos inicial y final de cada una de las sacadas, y el punto final de una de ellas es el punto inicial de la siguiente. Así se tiene la representación de la ruta de exploración como una secuencia temporal de vectores.

Paso 2: Simplificación de la ruta de exploración.

En cada secuencia vectorial obtenida se combinan iterativamente las fijaciones sucesivas si están a una distancia determinada como mínima o dentro de un umbral direccional dado, es decir, los recorridos de exploración se simplifican en función del ángulo y la longitud de los vectores que lo forman. Se agrupan dos o más movimientos sacádicos si los ángulos entre dos movimientos sacádicos consecutivos están por debajo de un umbral angular y las fijaciones intermedias son más cortas que un umbral de duración, o si la amplitud de los movimientos sacádicos sucesivos está por debajo de un umbral de longitud y la duración de la fijación circundante. Como tal, los movimientos sacádicos pequeños quedan contenidos localmente, y los movimientos sacádicos en la misma dirección se suman para formar movimientos sacádicos más grandes y menos complejos (ver [7]). Luego de este proceso, sacádicos en la misma dirección y fijaciones cercanas a otras se simplifican, por lo tanto, se necesita la determinación de un umbral iterativamente. Este proceso se repite hasta que no se realicen más simplificaciones y ayuda a reducir la complejidad de la ruta de escaneo conservando su estructura espacial y temporal.

Paso 3: Alineación temporal.

Siguiendo a este proceso de simplificación, las rutas de exploración simplificadas se alinean temporalmente, utilizando un enfoque de programación dinámica, para encontrar pares de vectores sacádicos comparables. El objetivo no es necesariamente alinear dos vectores sacádicos que constituyen el mismo componente en su respectiva secuencia vectorial, sino esos dos vectores que son los más similares conservando el orden temporal. Para hacerlo, todos los posibles emparejamientos de movimientos sacádicos se evalúan en similitud por su forma. La alineación se calcula optimizando la diferencia vectorial entre las rutas de escaneo.

Paso 4: Selección de la ruta de exploración.

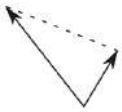
Una vez alineado todo par de scanpaths, se utiliza el conocido como algoritmo de Dijkstra para encontrar el camino más corto desde los primeros vectores sacádicos de cada ruta de escaneo hasta los últimos vectores sacádicos de cada par de scanpaths. La ruta "más corta" se define como la conexión entre fijaciones que tienen las menores longitudes de sacadas.

Paso 5: Cálculo de índices de similitud.

En esta etapa final, cada secuencia fijación-sacada se compara vectorialmente de a pares con todas las demás (aún cuando puedan diferir en longitud) en cinco dimensiones diferentes, calculando cinco medidas o índices de similitud entre cada par de rutas de exploración alineadas y simplificadas. Las

dimensiones de comparación son las siguientes:

- *Forma*: mide la similitud de la forma general de las rutas de exploración de a pares. Se calcula como la diferencia vectorial entre pares sacádicos alineados, normalizada por la diagonal de la pantalla y promediado sobre las rutas de exploración. Esta medida es sensible a las diferencias espaciales en las posiciones de fijación.



- *Dirección*: mide la distancia angular entre vectores sacádicos alineados. Es otra similaridad de forma cuando las amplitudes sacádicas son diferentes. Se calcula como el ángulo diferencia entre movimientos sacádicos alineados, normalizados por π y promediado sobre las rutas de exploración. Esta medida es sensible a la dirección de la sacada solamente, pero no a la amplitud o localización de la fijación.

- *Longitud*: mide la similitud en amplitud sacádica tomando el valor absoluto de la diferencia de longitud entre los puntos finales de los vectores sacádicos alineados, normalizados por la diagonal de la pantalla y promediado sobre las rutas de exploración. Esta medida es sensible a la amplitud sacádica solamente, no a la dirección de la sacada ni a la ubicación o duración de las fijaciones.

- *Posición*: mide la similitud en términos de la distancia euclideana entre fijaciones, tomando la diferencia de posición entre fijaciones alineadas, normalizada por la diagonal de pantalla y promediada entre las rutas de exploración. Esta medida es sensible tanto a amplitudes como a direcciones sacádicas.

- *Duración*: mide similitud en el tiempo de procesamiento. Se calcula como el valor absoluto de la diferencia en la duración de las fijaciones alineadas, normalizada por la duración máxima y promediada entre las rutas de exploración. Esta medida es insensible a la posición de la fijación y a la amplitud de la sacada.

Para obtener una descripción más detallada del algoritmo Multimatch, pueden consultarse la publicaciones originales de Dewhurst et al. [7] y Jarodzka et al. [19].

Bibliografía

[1] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, 1994.

[2] S. Cavaco, A. Goncalves, C. Pinto, E. Almeida, F. Gomes, I. Moreira, J. Fernandes, and A. Teixeira-Pinto. Trail making test: Regression-based norms for the portuguese population. *Archives of Clinical Neuropsychology*, 28:189–198, 2013.

[3] M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series, 2017.

[4] J. Dahmen, D. Cook, R. Fellows, and M. Schmitter-Edgecombe. An analysis of a digital variant of the trail making test using machine learning techniques. *Technol Health Care*, 25(2):251–264, 2017.

[5] J. A. Del Punta, G. Gasaneo, and L. U. Ancarani. Generalized sturmian functions used for a discrete wavelet construction. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 7(5-6):541, 2018.

[6] R. Dewhurst, T. Foulsham, H. Jarodzka, R. Johansson, K. Holmqvist, and M. Nyström. How task demands influence scanpath similarity in a sequential numbersearch task. *Vision Research*, 149:9–23, 2018.

[7] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behaviour Research Methods*, (44):1079–1100, 2012.

[8] A.G. Douglass. *Eye movements in neurocognitive disorders and frontotemporal dementia.* PhD thesis, 2016.

[9] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 06 2003.

[10] R. Engbert, K. Mergenthaler, P. Sinna, and A. Pikovskyb. An integrated model of fixational eye movements and microsaccades. In *Proceedings of the National Academy of Sciences*, volume 108, pages 765 – 770, 2011.

[11] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML*, 2003.

[12] B. Fischer and J.M. Buhmann. Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1411–1415, 2003.

[13] T. Foulsham, R. Dewhurst, M. Nyström, H. Jarodzka, R. Johansson, G. Underwood, and K. Holmqvist. Comparing scanpaths during scene encoding and recognition: a multi-dimensional approach. *Journal of Eye Movement Research*, 5(4):1–14, 2012.

[14] A. L. N. Fred and A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):835–850, jun 2005.

[15] A.L.N. Fred and A.K. Jain. Combining multiple clustering using evidence accumulation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 27:835—-850, 2005.

[16] S. Garbutt, A. Matlin, J. Hellmuth, A. K. Schenk, J. K. Johnson, H.J. Rosen, D. L. Dean, J. Kramer, J. M. Neuhaus, B. L. Miller, S. G. Lisberger, and A. L. Boxer. Oculomotor function in frontotemporal lobar degeneration, related disorders and alzheimer's disease. *Brain*, 131:1268 – 1281, 2008.

[17] R. Ghaemi, Md. N. Sulaiman, H. Ibrahim, and N. Mustapha. A survey: Clustering ensembles techniques. *International Journal of Computer and Information Engineering*, 3(2):365 – 374, 2009.

[18] K. Holmqvist and R. Andersson. *Eye tracking: A comprehensive guide to methods, paradigms and measures*. Lund, Sweden: Lund Eye-Tracking Research Institute., USA, second edition, 2017.

[19] H. Jarodzka, K. Holmqvist, and M. Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, 2010.

[20] D. Kimura, T. Ohtoshi, H. Bizen, A. Imai, M. Notoya, and K. Yamada. A study on visual search during the trail making test: Analysis using an eye tracker. *Neuroscience and Medicine*, 9:116–122, 2018.

[21] E. Kowler. Eye movements: The past 25 years. *Vision Research*, 51(13):1457 – 1483, 2011.

[22] M.L. Lai, M.J. Tsai, F.Y. Yang, C.Y. Hsu, T.C. Liu, S. W.Y. Lee, M.H. Lee, G.L. Chiou, J.C. Liang, and C.C. Tsai. A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, 10:90 – 115, 2013.

[23] I. Linari, G.E. Juantorena, A. Ibáñez, A. Petroni, and J.E. Kamienkowski. Unveiling trail making test: visual and manual trajectories indexing multiple executive processes. *Nature Scientific Reports*, (12):15, 2022.

[24] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.

[25] Y. Lu, I. Cohen, X.S. Zhou, and Q. Tian. Feature selection using principal feature analysis. *ACM Multimedia*, 2007.

[26] B. Luna, K. Velanova, and C. Geier. Development of eye-movement control. *Brain and cognition*, 68(3):293–308, 2008.

[27] L.E. Margulis, M.R. Squillace Louhau, and A.R. Ferreres. Baremo del Trail Making Test para Capital Federal y Gran Buenos Aires. *Revista Argentina de Ciencias del Comportamiento*, 10(3):54–63, 2018.

[28] T. L. (coord.) Martínez. *Técnicas de análisis de datos en investigación de mercados*. 2000.

[29] U. P. Mosimann, R. M. Müri, D. J. Burn, J. Felblinger, J. T. O'Brien, and I. G. McKeith. Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies. *Brain*, 128(6):1267–1276, 03 2005.

[30] J. Otero-Millán, X. G. Troncoso, S. L. Macknik, I. Serrao-Pedraza, and S. Martinez-Conde. Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *Journal of Vision*, 8(14):21, 2008.

[31] J. Paparrizos and L. Gravano. K-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 1855–1870, New York, NY, USA, 2015. Association for Computing Machinery.

[32] K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457 – 1506, 2009.

[33] R. M. Reitan and D. Wolfson. *The Halstead-Reitan neuropsychological test battery: theory and clinical interpretation (2nd ed.)*. Neuropsychology Press, Tucson, 1993.

[34] D. Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009.

[35] D. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*, pages 71 – 78, 2000.

[36] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[37] S. Tanggaard. Mental imagery in high-functioning autism spectrum disorder. scanpath analyses. Master's thesis, University of Oslo, 2016.

[38] L. Tao, Q. Wang, D. Liu, J. Wang, Z. Zhu, and L. Feng. Eye tracking metrics to screen and assess cognitive impairment in patients with neurological disorders. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 41(7):1697—1704, July 2020.

[39] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[40] K. Ting, F. Liu, and Z. Zhou. Isolation forest. In *ICDM 2008. Eighth IEEE International Conference on Data Mining*, pages 413–422, Los Alamitos, CA, USA, dec 2008. IEEE Computer Society.

[41] A. Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, page 331, USA, 2003. IEEE Computer Society.

[42] A. Topchy, A. K. Jain, and W. Punch. *A Mixture Model for Clustering Ensembles*, pages 379–390. 2004.

[43] A. Topchy, A.K. Jain, and W. Punch. Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

[44] A. Weingessel, E. Dimitriadou, and K. Hornik. An ensemble method for clustering. 2003.

[45] A. Z. Zheng and L. Huan. *Spectral Feature Selection for Data Mining*. Chapman and Hall/CRC, Minnesota, 2012.