ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA DEL SOFTWARE

TRABAJO FIN DE GRADO

VISUALIZANDO LA EVOLUCIÓN DE LOS CASOS DE COVID-19 CON TDA

Autor/es:

Casasola Calzadilla, María

Tutor/es:

González Díaz, Rocío

Paluzo Hidalgo, Eduardo

Repositorio de trabajo:

https://github.com/marcascal2/evolucion-casos-COVID-19-BM.git

Segunda convocatoria

Curso 2020/2021

Resumen

A continuación, se expone en este documento la memoria del trabajo de fin de grado basado en el análisis de datos de COVID-19 en España a través del algoritmo BallMapper. La idea general de este trabajo se basa en sacar conclusiones que no habían sido expuestas anteriormente sobre el COVID-19, gracias a la aplicación del algoritmo mencionado, y confirmar otras que ya se habían hallado. Para conseguir esto, antes, se han debido escoger los conjuntos de datos sobre los que realizar las pruebas, realizando una limpieza y preprocesado de los datos y aplicando finalmente el algoritmo BallMapper. Tras la aplicación, se han analizado los resultados obtenidos para concluir con las ideas generales del estudio. Gracias a este estudio se han conseguido hallar nuevas deducciones sobre la situación de pandemia vivida durante el año 2020 y que aún sigue muy presente en nuestras vidas.

Abstract

Next, the report of the final degree project based on the analysis of COVID-19 data in Spain through the BallMapper algorithm is exposed in this document. The general idea of this work is based on drawing conclusions that have not been previously exposed about COVID-19, thanks to the application of the aforementioned algorithm, and confirming others that have already been found. To achieve this, before, the data sets on which the tests were carried out had to be chosen, cleaning and preprocessing the data and finally applying the BallMapper algorithm. After the application, the results obtained have been analyzed to conclude with the general ideas of the study. Thanks to this study, new deductions have been obtained about the pandemic situation experienced during the year 2020 and that is still very present in our lives.

INDICE

INDICE DE ILUSTRACIONES

INDICE DE TABLAS

Introducción

En este informe se pretende realizar un análisis de distintos casos de COVID-19 en España con la aplicación del algoritmo BallMapper (BM). La mayor parte de los datos de entrada a los que aplicaremos el algoritmo BM consisten en información suministrada por el repositorio de GitHub (FERNÁNDEZ CASAL, 2020) que, a su vez, emplea: *datos oficiales disponibles en la pestaña Documentación y Datos de la web Situación de COVID-19 en España (aplicación shiny) del Instituto de Salud Carlos III (ISCIII).* La última extracción de datos para el proyecto se ha realizado el 31 de julio de 2021.

El análisis de datos es un proceso que nos permite interpretar los datos de los que disponemos inspeccionando, tratando y evaluando dichos datos. Los datos pueden ser analizados para probar conjeturas o la invalidez de teorías, pero, sobre todo, para sacar conclusiones pudiendo realizar, en algunos casos, ciertas predicciones evitando riesgos o factores negativos. Sin embargo, la extracción de información de conjuntos de datos de gran dimensión, incompletos y ruidosos suele ser un desafío. Un tipo de análisis de datos es el análisis topológico de datos, en inglés topological data analysis (TDA), se puede describir en términos generales como una colección de métodos de análisis de datos que encuentran estructura en los datos. Estos métodos incluyen agrupamiento, estimación múltiple, reducción de dimensión no lineal, estimación de modo, homología persistente, etc. BallMapper es un algoritmo pensado para realizar este tipo de análisis sobre el conjunto de datos escogido.

Por otra parte, cada año, miles de personas enferman a causa de distintos factores que cursan en forma de brotes. Por ello, detectar lo antes posible la aparición de los brotes y evitar su extensión y agravamiento se ha convertido en este último año en el objetivo principal de muchos países, entre ellos, España.

En diciembre de 2019 surgió un agrupamiento de casos de neumonía en la ciudad de Wuhan (provincia de Hubei, China). El 7 de enero de 2020, las autoridades chinas identificaron como agente causante del brote un nuevo virus de la familia Coronaviridae que posteriormente fue denominado SARS-CoV-2 1 (COVID-19). La secuencia genética fue compartida por las autoridades chinas el 12 de enero. El Comité de Emergencias del Reglamento Sanitario Internacional (RSI, 2005) declaró el brote como una Emergencia de Salud Pública de Importancia Internacional (ESPII) en su reunión del 30 de enero de 2020, y, posteriormente, la OMS lo reconoció como una pandemia global el 11 de marzo de 2020.

Tras la aplicación de BM sobre los distintos conjuntos de datos recogidos a lo largo de los años 2020-21, se han podido corroborar muchas de las conclusiones extraídas con anterioridad por distintos estudios ya realizados. Sin embargo, además de esto, se han hallado detalles que habían podido pasar desapercibidos. Se debe tener en cuenta que es la primera vez que se usa en España un algoritmo de este tipo para sacar conclusiones sobre el COVID-19.

Esta memoria se divide en **7 capítulos** junto con esta previa introducción, una breve explicación de los objetivos y un glosario de términos relevantes para el trabajo. Además, podemos encontrar en la portada del informe la **URL del repositorio en GitHub** que cuenta con el código utilizado para cada caso de análisis. En concreto, los capítulos del documento se organizan de la siguiente manera: el ***Capítulo 1*** sintetiza los recursos utilizados a lo largo del desarrollo y el tiempo invertido en él. El ***Capítulo 2*** recoge aspectos relacionados con la epidemiología y pasos a tener en cuenta. En cuanto al ***Capítulo 3***, este habla sobre los conjuntos de datos y su relación con el estudio. El ***Capítulo 4*** por su parte, recoge todos los conocimientos matemáticos necesarios para comprender el algoritmo utilizado, así como la descripción del propio algoritmo. En el ***Capítulo 5*** ya se realiza una comparación entre dos estudios realizados previamente con BM. Finalmente, el ***Capítulo 6*** es nuestro estudio en cuestión con los diferentes grupos de datos utilizados y el ***Capítulo 7*** narra las conclusiones extraídas después de dicho estudio.

Objetivos del estudio

El estudio trata sobre la comprensión de la propagación de enfermedades, en concreto la Covid-19, a través de la visualización de datos teniendo en cuenta importantes interacciones multidimensionales, como son distintas características de las comunidades autónomas españolas. Con la ayuda del algoritmo BM de análisis de datos topológicos, se pretende comprender cómo se propaga la enfermedad atendiendo a diferentes características, haciendo uso de los datos de contagios recopilados a lo largo de la primera ola de la pandemia en los distintos territorios españoles.

Realizando una síntesis, se puede decir que los objetivos son:

1) El estudio de trabajos que aplican el algoritmo de Mapper a datos de la COVID en E.E.U.U. y U.K.
2) La reproducción de la metodología propuesta en dichos estudios a datos de la COVID en España

Glosario de términos

Términos epidemiológicos

- **Brote**: incremento significativo de casos en relación a los valores esperados.
- **Cuadro clínico**: conjunto de síntomas característicos de una enfermedad que suelen aparecer en las personas que la padecen.
- **Métodos preventivos**: preparación y disposición de métodos que se toman anticipadamente para evitar un riesgo o ejecutar algo.
- **Pandemia**: enfermedad epidémica que se extiende a muchos países o que ataca a casi todos los individuos de una localidad o región.
- **Seroprevalencia**: porcentaje de personas en una población que tienen anticuerpos en la sangre, que indican que han estado expuestas a un virus u otro tipo de organismo infeccioso.
- **Síntoma**: señal o indicio reveladora de una enfermedad.

Términos matemáticos

- **Algoritmos Mapper**: algoritmos que encapsulan la estructura local y global de conjuntos de datos dados y se pueden utilizar en análisis de datos exploratorios. Son un medio de representar conjuntos de datos complejos y de alta dimensión usando los grafos como estructura. Una variante del algoritmo Mapper es BallMapper.
- **Conjunto de datos**: contiene los valores para cada una de las variables que corresponden a cada miembro del conjunto de datos.
- **Grafo de Reeb**: objeto matemático que refleja la evolución del conjunto de nivel de una función de valor real en una variedad diferenciable.
- **Media**: promedio de un conjunto de datos. Es el centro de gravedad de la distribución de datos.
- **Mediana**: punto medio en un conjunto de datos que ha sido dispuesto desde el más pequeño al más grande, es decir, el valor que supera a la mitad de los de la muestra y se ve superado por la otra mitad.
- **Medidas de tendencia central**: son las que representan un punto central en torno al cual se encuentran las observaciones.
- **Medidas estadísticas**: ayudan a generalizar un grupo de datos, a hacer inferencias sobre este, y a compararlo con otros grupos de datos.

- **Moda**: valor que aparece más veces en un conjunto de datos, es decir, el valor que tiene mayor frecuencia.
- **Rango**: anchura de los datos. La diferencia entre el valor más grande y el más pequeño, que indica qué extensión de la recta de los números ocupan los datos de nuestra muestra.
- **TDA**: (Topological Data Analysis) es un enfoque para el análisis de conjuntos de datos utilizando técnicas de topología.
- **Topología computacional**: subcampo de topología con una superposición con áreas de Ciencias de la Computación, en particular, geometría Computacional y teoría de la complejidad computacional.

Capítulo 1. Análisis temporal y costes de desarrollo

Análisis temporal del proyecto

En este apartado se exponen las fases en las que se subdivide el desarrollo del proyecto y el tiempo revisto y real para cada una de ellas divididas en meses. Para ello, se representarán ambas cantidades de tiempo en una tabla de cara a poder realizar una comparación más exhaustiva entre el tiempo estimado y el tiempo real invertido en cada periodo de tiempo.

En primer lugar, la estimación inicial se realizó de cara a la convocatoria de julio y, conforme avanzaba el desarrollo, dicha estimación se modificó en los meses restantes enfocándose en la convocatoria de septiembre. Sin embargo, esto no produjo ningún tipo de inconveniente ya que la estimación inicial se basaba en un reparto equitativo de las horas por semanas, atendiendo al nivel de ocupación que se estimaba tener en cada época.

Las fases en las que se divide el estudio realizado y las estimaciones correspondientes realizadas para cada una son las siguientes:

	DEFINICIÓN/ INICIACIÓN	PLANIFICACIÓN	EJECUCIÓN	CONTROL/ SEGUIMIENTO
INICIO	22/10/2020	04/11/2020	18/11/2020	22/10/2020
FINAL ESTIMADO	-	18/11/2020	31/08/2021	31/08/2021
FINAL REAL	04/11/2020	18/11/2020	01/09/2021	30/08/2021
NÚMERO DE HORAS ESTIMADAS	-	12:00	266:00	15:00
NÚMERO DE HORAS REALES	9:00	20:44	269:10	9:22

Tabla 1. Tabla de tiempos estimados y reales por fases

Definiendo en que consiste cada fase, podría quedar resumido como sigue:

- **Definición/Iniciación:** Consiste en diferentes reuniones para estudiar las diferentes posibilidades y realizar una elección final del tema, además de una investigación previa sobre él.
- **Planificación:** Se realiza una planificación del tiempo estimado que se va a invertir en el proyecto repartido en fases y meses.
- **Ejecución:** Se lleva a cabo el objetivo de la propuesta escogida. Esta fase implica investigación sobre el algoritmo utilizado, aplicación del mismo, extracción de conclusiones y documentación.
- **Control/Seguimiento:** En esta fase se recogen todas las horas dedicadas a reuniones de seguimiento a lo largo del desarrollo.

Por otra parte, la estimación de horas y las horas reales invertidas por meses quedan recogidas en la siguiente tabla, ***Tabla 2***.

A su vez, cada mes se ha dividido por semanas, estimando las horas a invertir en cada semana por cada mes. Sin embargo, en este documento quedan expuestas de manera global por meses, como se ha mencionado anteriormente.

MES	HORAS ESTIMADAS	HORAS REALES
OCTUBRE	-	9:00
NOVIEMBRE	36:00	38:58
DICIEMBRE	36:00	30:18
ENERO	21:00	29:07
FEBRERO	31:00	26:20
MARZO	48:00	49:00
ABRIL	25:00	12:42
MAYO	15:00	22:37
JUNIO	20:00	12:40
JULIO	23:00	33:24
AGOSTO	36:00	34:50
TOTAL	300:00	308:16

Tabla 2. Tabla de tiempos estimados y reales por meses

De esta manera, podemos calcular la tabla de desviación temporal de cara a mostrar el error cometido a la hora de la estimación respecto de la realidad.

En nuestro caso, debido a que las fases son muy generales, se procede a realizar la tabla de desviación temporal con los datos recogidos en la ***Tabla 2***.

MES	HORAS ESTIMADAS	HORAS REALES	DESVIACIÓN TEMPORAL
OCTUBRE	-	9:00	-
NOVIEMBRE	36:00	38:58	2:58
DICIEMBRE	36:00	30:18	5:42
ENERO	21:00	29:07	8:07
FEBRERO	31:00	26:20	4:40
MARZO	48:00	49:00	1:00
ABRIL	25:00	12:42	12:18
MAYO	15:00	22:37	7:37
JUNIO	20:00	12:40	7:20
JULIO	23:00	33:24	10:24
AGOSTO	36:00	34:50	1:10

Tabla 3. Tabla de desviación temporal por meses

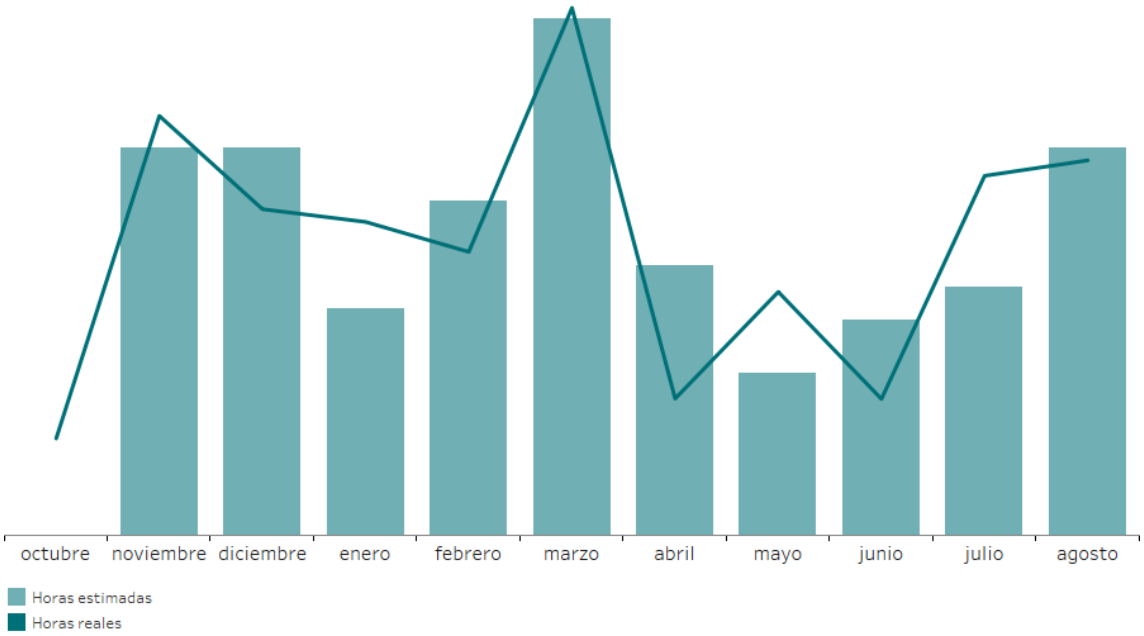


Ilustración 1. Gráfica de desviación temporal por meses

Costes de desarrollo

Los costes a tener en cuenta para el desarrollo de un proyecto son: costes de personal, costes de servicios y amortización.

Costes de personal

En cuanto a los costes de personal, en este proyecto solo se implicaría el coste directo de una persona. Por tanto, en caso de contar con el título del Grado en Ingeniería Informática del Software, podríamos determinar el coste según el número de horas trabajadas:

	Sueldo	*Horas trabajadas*	*TOTAL*
Ingeniero informático	13€/h	308h 16’	**4007,47 €**

Tabla 4. Tabla de costes de personal

Costes de servicios

En este apartado, solo se deben tener en cuenta los servicios básicos consumidos (Internet y luz).

Servicio	*€/mes*	*Meses de consumo*	*TOTAL*
Luz	35.5	10	355 €
Internet	31	10	310 €

TOTAL	**665 €**

Tabla 5. Tabla de costes de servicios

Amortizaciones

El último paso será calcular las amortizaciones, es decir, únicamente los equipos usados para el desarrollo del proyecto. Ya que los sistemas y programas utilizados en la ejecución no han requerido ningún coste.

Se escoge el porcentaje de amortización anual de un 20% como se indica en la Agencia Tributaria, (Gobierno de España. Ministerio de Hacienda y Función Pública., s.f.).

Producto	*€*	*Meses de uso*	*Amortización*	*TOTAL*
HP 2016 Flagship Laptop 15.6" HD, Intel i7-6500U	899	10	20%	**149.83 €**

Tabla 6. Tabla de amortizaciones

Coste total

Sabiendo todo lo expuesto en los apartados anteriores, se muestra a continuación la tabla con los costes totales del proyecto:

Tipo de coste	*Coste total*
Coste de personal	4007,47 €
Coste de servicios	665 €
Amortizaciones	149.83 €
TOTAL	**4822.30 €**

Tabla 7. Tabla de costes totales del proyecto

Capítulo 2. Investigación sobre epidemiología

Debido a la relación tan estrecha del estudio con el concepto de brote epidémico, se debe abordar, antes de empezar, la estrategia correcta a seguir frente a dichos brotes.

Frente a este tipo de situaciones se debe acudir a una prevención primaria, que conlleva una serie de métodos preventivos, como podría ser el uso de mascarillas o, como pasó en marzo del 2020, la declaración de un estado de alarma frente a la expansión exponencial del virus. Dependiendo de la gravedad de la situación, las medidas tomadas varían, pero se considera indispensable la aplicación de medidas de prevención secundarias como es la detección de la aparición de nuevos brotes lo antes posible y

evitar la propagación y agravamiento de la situación. Es en este punto donde entra en materia nuestro estudio, ya que la finalidad principal de éste se enfoca en la extracción de información que ayude a comprender mejor las características que pueden estar relacionadas con la expansión del virus.

Utilidad del estudio de un brote epidémico y pasos a seguir

Principalmente, la finalidad de un estudio de un brote epidémico se basa en conocer sus causas para evitar su propagación pudiendo aplicar medidas efectivas, determinar los factores que condicionan su aparición y ayudar a detectar enfermedades nuevas. Además, produce una mejora del conocimiento global de la población, atendiendo a detalles que han podido pasar desapercibidos.

Siguiendo el documento de protocolo de investigación ante un brote epidémico (Portal oficial de la Administración de la Junta de Andalucía), los pasos para la actuación ante la aparición de un brote epidémico, en resumidas cuentas, deberían ser los siguientes:

1. Verificar la existencia del brote. Debemos advertir la presencia del brote lo antes posible para poder actuar con rapidez.
2. Construir la definición de caso. Deben incluir signos y síntomas comunes y pertinentes de la enfermedad que, individualmente o juntos, configuren un cuadro clínico claro o indicativo de la enfermedad.
3. Recuento de los casos existentes hasta el momento, realizando un análisis de toda la población objeto.
4. Desarrollar una hipótesis orientando los datos en términos de tiempo, lugar y persona, y determinando quién está a riesgo de enfermar.
5. Implementar las medidas de control y prevención necesarias.
6. Recoger cualquier tipo de hallazgo o avance, cualquier tipo de información que pueda aportar nuevas consideraciones.

La información disponible inicialmente nos puede proporcionar una gran cantidad de ayuda, de ahí la importancia de aportar detalles que puedan servir de orientación como, por ejemplo, la fecha de aparición de los síntomas, número de personas afectadas, localización del brote, etc. Sin embargo, la realización de este estudio está basada en el paso número 6, ya que, al no haber considerado algunos aspectos que podrían estar estrechamente relacionados con la aparición de síntomas, es interesante hacerlo ahora y descubrir cualquier tipo de aspecto nuevo de cara a un posible rebrote o nuevas enfermedades.

Capítulo 3. Conjuntos de datos

Un conjunto de datos (*dataset*) son datos usualmente presentados de manera tabular. Cada columna del conjunto representa una variable en particular y cada fila corresponde a un miembro dado del conjunto de datos en cuestión. Recoge valores para cada variable, como, por ejemplo, la altura, el peso o el color de un objeto. Cada valor se conoce como un dato. Cada conjunto de datos puede incluir datos para uno o más miembros, correspondiendo al número de filas. Además, cada uno tiene varias características que definen su estructura y propiedades. Estos incluyen el número y tipos de atributos o variables y las medidas estadísticas que se pueden aplicar a él.

Existen distintos tipos de variables con los que puede contar un conjunto de datos. Por un lado, se encuentran las variables cualitativas, también llamadas variables categóricas o atributos, que son aquellas que no son expresadas por números. Dentro de estas podemos ramificar aún más los tipos en: ordinales, puras o dicotómicas. Y por otro lado tenemos las variables cuantitativas, o también conocidas como numéricas, que son las que necesitan números para expresar valor. También pueden subdividirse en dos tipos: discretas o continuas.

Continuando con los conceptos que rodean la idea de *conjunto de datos*, podemos hablar ahora de las medidas estadísticas. Los resultados de aplicar los cálculos de estas medidas a los conjuntos de datos nos muestran características de ellos de forma resumida y previa a su exhaustivo análisis, (Servicio de Actualización Académica y Capitalización en Estadística. (SERAACE)): *Los datos obtenidos de la muestra son de utilidad para obtener estimaciones de los valores de la población*.

Podemos diferenciar tres grupos diferentes de parámetros estadísticos descriptivos:

- **De centralización (o posición central)**: muestran aspectos basados sobre todo en datos centrales.
- **De dispersión**: dan una valoración de cómo de dispersos están los datos.
- **De posición**: informan sobre la distribución de una variable exponiendo a cuántos de sus valores supera un dato dado.

En este estudio se decide realizar un análisis descriptivo previo para cada conjunto de datos utilizados en el ***Capítulo 6***. Para ello, se han decidido aplicar los parámetros de centralización: moda, media y mediana, y el parámetro de dispersión: rango. Podemos

encontrar la definición completa de dichas medidas en el apartado ***Glosario de términos***, en concreto en el subapartado *de* ***términos matemáticos***.

En cuanto a la preparación y preprocesado de un conjunto de datos, se conforma de distintas tareas entre las que podemos mencionar la integración y unificación de los datos, la depuración y la ingeniería de características o variables. Todas estas tareas componen el ciclo de vida del modelado y exploración de un conjunto de datos. Se parte de una fase de entendimiento de los datos y de su extracción de distintas fuentes. Posteriormente, se procede a una unificación y transformación en un único conjunto de datos, así como a una selección de aquellas variables relevantes para el estudio. Para determinar qué variables mantener, es importante entender el problema y los datos obtenidos.

En la depuración de los datos también influye el correcto entendimiento de las circunstancias que rodean el estudio. El preprocesado consiste en rellenar o descartar valores desconocidos de las variables, encontrar “outliers”, eliminar características que cuyo uso sea ilegal o inmoral… Por otro lado, la ingeniería de características conlleva una serie de decisiones por parte del científico de datos, como puede ser la normalización de una variable o cualquier otra transformación de estas que pueda beneficiar el estudio.

El estudio y previsualización de un conjunto de datos es interesante simplemente por el hecho de conocer un poco más nuestro conjunto, sobre todo cuando se trata de conjuntos amplios, y, además, según las medidas estadísticas aplicadas, incluso podríamos tener una idea de si los resultados que se obtienen en el estudio posterior del conjunto son correctos y esperados. De hecho, la elección de las variables mencionadas anteriormente viene provocada por la necesidad de tener una idea general de cada columna en cada caso. De esta forma podríamos ver la diferencia entre los rangos de valores de cada una y si se considera necesaria la normalización.

Capítulo 4. Estudio del algoritmo BallMapper

Tras una serie de consideraciones, se decidió implementar el código en R, (Dlotko, BallMapper, 2019), debido a que se trata de un lenguaje bastante adecuado para la estadística, ya que permite manipular los datos de forma precisa. Además, puede leer prácticamente cualquier tipo de datos. Hasta cierto punto, es compatible con grandes

conjuntos de datos, aunque podemos advertir ciertas limitaciones de este software como, por ejemplo, la lentitud, la cual resta efectividad.

Además, como ya describe el Dr. Curran J. M. en su libro (Curran, 2011), R cuenta también con ventajas como el tratarse de un lenguaje gratis, además de no tener que estar instalado en los directorios del sistema. Es la elección de muchos estadísticos profesionales, extensible y cuenta con un sistema de gráficos de alta calidad. Es más, de cara al cálculo de posibles medidas estadísticas sobre los conjuntos de datos obtenidos para este informe, *R también admite un enfoque de las estadísticas basado en la interfaz de línea de comandos (CLI) mucho más robusto y útil* (Curran, 2011).

Destacando lo más importante, R tiene capacidades avanzadas de representación gráfica, por lo que nos permite realizar gráficos de forma que podamos presentar los resultados de forma vistosa, algo a tener muy en cuenta de cara al uso del algoritmo BM, cuya visualización es esencial para sacar las conclusiones acertadas. Aunque también podemos destacar en este punto una desventaja de R, no soporta gráficos en tres dimensiones.

Paweł Dłotko y BallMapper CRAN package

El código utilizado para este estudio ha sido desarrollado por Paweł Dłotko, matemático e informático, cuya principal área de especialización es la topología computacional. Desde enero de 2013 hasta enero de 2015 trabajó como investigador postdoctoral en la Universidad de Pennsylvania, aunque anteriormente había sido profesor asistente en la Universidad Jagiellonian de Cracovia. Es, además, uno de los desarrolladores de la biblioteca Gudhi. Actualmente trabaja en el área de topología computacional aplicada.

Como bien describe nuestro autor: (DLOTKO, 2019): *Ball Mapper es un resumen de formas para el análisis de datos topológicos. Ball Mapper proporciona un resumen topológicamente preciso de datos en forma de gráfico abstracto.*

El primer paso, por tanto, se fijó en el estudio del código escogido y su fundamento matemático, para entender en profundidad cómo funciona, aunque el mismo autor proporciona dos ejemplos para una primera visualización de la salida del algoritmo.

Estudio del algoritmo BallMapper con TDA

El análisis topológico de datos (TDA) tiene como objetivo utilizar métodos rigurosos de topología, como grafos Reeb o Mapper, así como la homología persistente, para analizar datos complicados, grandes y de alta dimensión. *La homología es un formalismo matemático para hablar de manera cuantitativa e inequívoca sobre cómo está conectado un espacio. En comparación con la mayoría de los otros formalismos competidores, la homología tiene algoritmos más rápidos, pero captura menos información topológica. […] En la práctica, tener algoritmos rápidos es una ventaja definitiva y ser insensible a cierta información topológica no es necesariamente un inconveniente,* (Edelsbrunner & Harer, 2010).

Con el grafo de Reeb, el análisis de espacios topológicos estaba cubierto para los datos continuos, es decir, este algoritmo trata de explicar de forma sintetizada la estructura que puede tener una superficie o figura 3D.

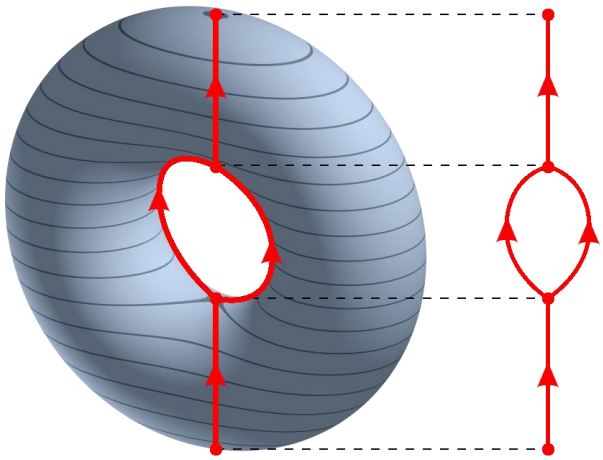


Ilustración 2. Ejemplo de aplicación del Grafo de Reeb sobre una superficie (WIKIPEDIA)

Sin embargo, la necesidad de realizar estas estimaciones con datos discretos, como lo es una nube de puntos, hizo que surgiera entonces el algoritmo Mapper. Este algoritmo se introdujo en los métodos topológicos para el análisis de conjuntos de datos de alta dimensión y se describe típicamente como una forma de aproximar un grafo de Reeb, (Goldfarb, 2018). De esta manera, mientras estos algoritmos se centran en describir la forma de la nube de puntos que tenemos como objeto de estudio, BM nos describe esta nube atendiendo a las características que tienen en común los puntos que la forman y las relaciones que comparten.

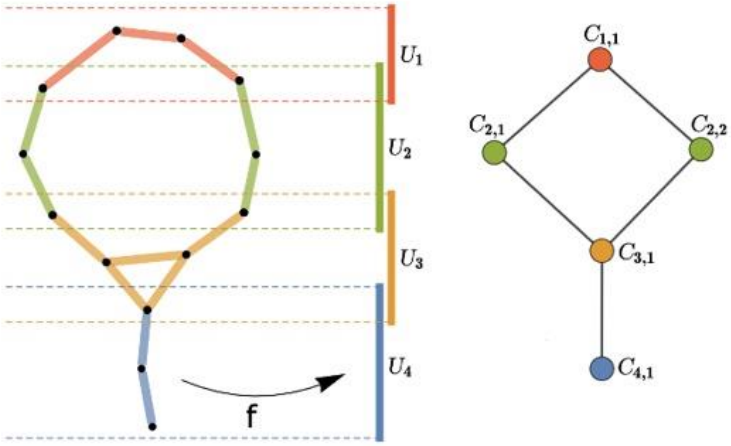


Ilustración 3. Ejemplo de la aplicación del algoritmo Mapper (MINERVINO)

Para explicar BallMapper, necesitaremos antes exponer los conceptos necesarios para su comprensión.

Espacio R^N

En este proyecto vamos a trabajar con el espacio $\mathrm{R^N}$, espacio vectorial y consiste en el producto cartesiano de N copias de R. La interpretación geométrica de los elementos de R^{N} se plantea como puntos o vectores, además, para N = 2 tenemos un plano y para N = 3 el espacio tridimensional que somos capaces de percibir, sin embargo, para N > 4, perdemos la intuición geométrica. En R^{N} disponemos de las operaciones de suma y producto por escalares, utilizadas en los cálculos para la aplicación del algoritmo BM.

Símplices y complejos simpliciales

Un **símplice** es un conjunto de puntos en un espacio afín, de manera que un k-símplice (k-simplex) es el conjunto de puntos definidos por una envolvente conexa de k+1 puntos independientes.

Dada una nube de puntos, se puede construir un **complejo simplicial** K, que es una colección de símplices tales que:

- para cada símplice s del complejo, todos los símplices que lo componen forman parte del complejo, $s \subset K$.
- dados dos símplices del complejo $s_1, s_2 \in K$, la intersección es vacía u otro símplice del complejo s_x, $s_1 \cap s_2 = \emptyset$ ó $s_1 \cap s_2 = s_x$.

Mapper

La idea básica puede denominarse agrupamiento parcial, en el sentido de que un paso clave es aplicar algoritmos de agrupamiento estándar a subconjuntos del conjunto de datos original, de manera que, siguiendo una función de filtrado (que no es más que una forma continua de recorrer los puntos) obtengamos un recubrimiento del conjunto de datos mediante subconjuntos no disjuntos, y luego comprender la interacción de los grupos parciales formados de esta manera entre sí. Esta construcción produce una imagen "multiescala" del conjunto de datos. De hecho, se puede construir una familia de complejos simpliciales así. Este hecho permite evaluar hasta qué punto las características son "significativas", es decir, no son características producidas por ruido en el dato de entrada. Luego, se cuentan las componentes conexas de nuestro conjunto, que representarán los nodos del grafo, y se forman las relaciones entre ellas a través de aristas, en función de las intersecciones entre los clústeres.

No intentamos obtener una representación totalmente precisa de un conjunto de datos, sino más bien una imagen de baja dimensión que sea fácil de entender y que pueda señalar áreas de interés. En resumen, se obtiene una representación que muestra cómo se puede recorrer la nube de puntos de forma continua.

Ball Mapper

Se podría explicar a grosso modo el funcionamiento de BM partiendo de la nube de puntos mencionada anteriormente. Sobre ella aplicamos una función de filtrado, que en este caso va implícita en el algoritmo, ya que se engloban distintos puntos en una misma bola si reúnen las propiedades necesarias de similitud para pertenecer a un mismo conjunto que depende del valor de un parámetro ε. Por tanto, se define un algoritmo de agrupación, de manera que los nodos del grafo tienen peso y se representan de mayor tamaño si hay más puntos en su ε-*bola* (bola de radio ε). De esta manera, cada clúster equivale a un conjunto de vértices en una misma componente conexa.

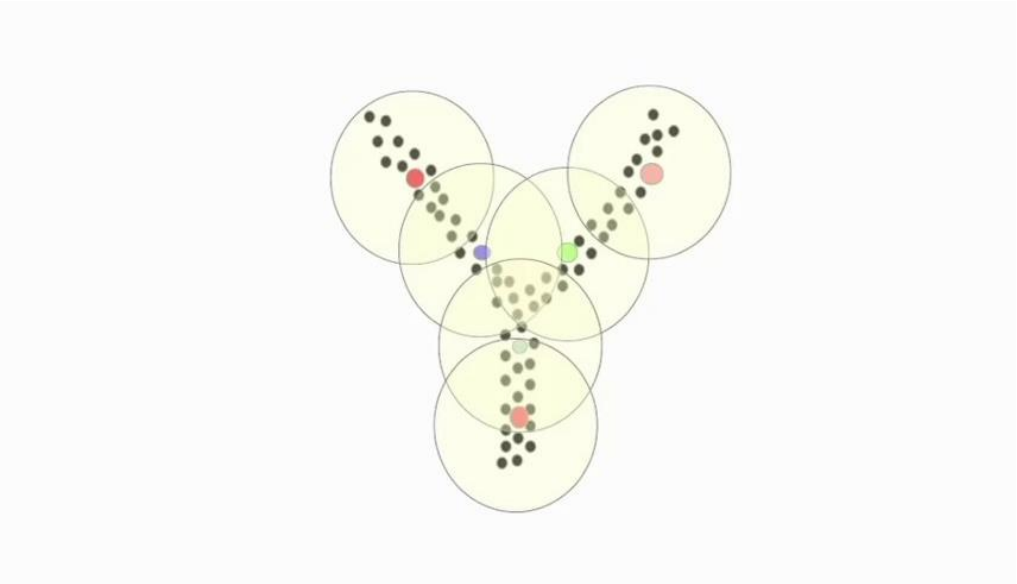


Ilustración 4. Ejemplo de aplicación del algoritmo BallMapper (DLOTKO, IMAGEN APLICACIÓN DE BALLMAPPER SOBRE UNA NUBE DE PUNTOS)

Se seleccionan los puntos mayores de manera que estén a distancia menor que ε, y todos los puntos negros que están dentro de la ε-bola centrado en un punto "gordo" pertenecen a su mismo conjunto. Así, se obtiene un recubrimiento por subconjuntos no disjuntos.

ALGORITMO 1: ALGORTIMO BALLMAPPER:

Entrada: $espacio\_referencia,\ \varepsilon$

Salida: $grafo$

for c_i ***in*** $espacio\_referencia$:

if $c_i = \{\{C_\alpha\}\}_{\{\alpha \epsilon A\}}$:

$\{X_\alpha\} = construye\_subconjuntos(c_i)$

endif

$agr = aplica\_agrupamiento(\epsilon, X_\alpha)$

$construye\_complejo\_simplicial(agr)$

endfor

Comparación entre Mapper y Ball Mapper

El método Mapper tradicional comienza con un conjunto de datos X y una función de valor real $f: X \rightarrow R$, que es la que hemos llamado función filtro. Esta función puede ser una función que refleja las propiedades geométricas del conjunto de datos, como es el resultado de un estimador de densidad, o puede ser una función que refleje las

propiedades de los datos que se están estudiando. En el primer caso, se intenta obtener información sobre las propiedades cualitativas del conjunto de datos en sí, y en el segundo caso se intenta comprender cómo estas propiedades interactúan con funciones interesantes en el conjunto de datos.

La función $f: X \rightarrow R$ utilizada en el Mapper es continua. Entonces, para cada $x \in X$ y $\varepsilon > 0$ existe $\delta > 0$ tal que $f\,(B\,(x, \varepsilon)) \subset B\,(f\,(x), \delta)$. En otras palabras, se supone que los puntos cubiertos por cada bola en BM tienen valores similares de la función f. En este caso, podemos esperar alguna correspondencia entre los gráficos de Mapper y BallMapper.

La conectividad de los dos métodos puede ser drásticamente diferente si el crecimiento de la función f es demasiado grande en comparación con la densidad de muestreo. En esos casos pueden existir regiones desconectadas de Mapper que están conectadas en BM, ya que este último no depende de la función, sino solo de la proximidad de puntos en X (BM conserva una noción de proximidad, que es a menudo una propiedad deseable, porque las funciones de distancia a menudo codifican la similitud o proximidad entre los puntos).

Capítulo 5. Comparación de estudios de EEUU y UK

Ambos artículos, (DŁOTKO & RUDKIN, 2020) y (Chen & Volic, 2021), proporcionan un estudio sobre la aplicación de los algoritmos BallMapper y Mapper a los datos de COVID-19 recopilados en Reino Unido y Estados Unidos, respectivamente. Como se explica en los dos estudios, se ha demostrado que estos algoritmos capturan una serie de tendencias en la propagación de COVID-19 proporcionando información más completa que la que ofrecen otras técnicas más estándar de visualización de datos como t-SNE, por ejemplo.

Sin embargo, las diferencias entre ambos estudios son considerables. Teniendo en cuenta que ambos países cuentan con densidades demográficas y superficies desiguales, la elección de los datos introducidos en cada estudio y la decisión de cómo abordar dicho problema están relacionados con estos factores.

Discrepancias

De manera global, podemos observar que el estudio en Reino Unido está guiado por la relación establecida entre los contagios y ciertas consideraciones de ingresos, patrones de trabajo y edad de la población que pueden ayudar a comprender la propagación. No obstante, la atención de este estudio se centra en la combinación de cada característica asociada a la densidad de población. Estas características son la edad media, el producto interior bruto (PIB), el valor añadido bruto (VAB), las horas trabajadas y el salario bruto. En cada caso, el gráfico muestra el número total promedio de casos reportados hasta la fecha dada para cada bola.

Por otro lado, atendiendo a la investigación en Estados Unidos, la representación de Mapper captura la evolución de la propagación en todo el país, facilitando la comparación de datos en el tiempo y el espacio. La posición de un determinado condado de EE. UU. en el gráfico se determina, en parte, en relación con los condados circundantes, es decir, en este estudio se considera únicamente la posición geográfica de cada punto y el día en el que se registró el número de contagios.

En cuanto a los datos introducidos en cada uno, podemos observar que, en el trabajo realizado en U.K, el algoritmo BM produce una representación bidimensional, aunque no se especifica la manera de codificar los datos a introducir. En cuanto al caso de Estados Unidos, también se cuenta con una representación bidimensional como resultado de la aplicación de Mapper, pero los puntos de datos iniciales están formados por 4 dimensiones, información sobre la ubicación geográfica (latitud y longitud), la fecha y el número de casos acumulativos confirmados de COVID-19 en ese condado en esa fecha. Específicamente, se usa una colección de datos de series temporales sobre el número de COVID-19 confirmados a partir del 22/1/20.

La discrepancia entre los datos introducidos conlleva distintas maneras de normalizar valores, ya que, por ejemplo, en el estudio de EE. UU. es necesaria una normalización de 4 dimensiones distintas. Para la aplicación de BM en UK, simplemente los ejes del conjunto de datos se normalizan en el intervalo [0, 1], para superar el desafío de tener un eje dominante. No obstante, como se explica en el estudio de EE. UU, *se necesita compensar el efecto de distorsión inducido por las diferencias numéricas de cada coordenada de los vectores de puntos, normalizando cada una de ellas por separado,* (Chen & Volic, 2021).

Atendiendo a la coloración de nodos, también podemos encontrar en este aspecto una gran diferencia entre ambos documentos. En el caso de Reino Unido, el color de los puntos contiene la mayor parte de la información, ya que se trata del resultado numérico de casos de Covid-19. Se agrega una barra de color al gráfico en el que los valores más bajos son rojos y los más altos son morados. Sin embargo, como se explica en el artículo de Estados Unidos, los colores en los gráficos Mapper no tienen ningún significado matemático. Los nodos se colorean de acuerdo con cómo se ordenan los datos, y esto se hace en base a la información geográfica, de modo que los condados cercanos tengan el mismo color. Por lo tanto, se pueden usar los colores para ayudar a distinguir ubicaciones geográficas.

Finalmente, cabe destacar el proceso de filtrado de información en el artículo correspondiente a Estados Unidos, el cual no es necesario y no se realiza en el estudio de Reino Unido. Esta diferencia simplemente está relacionada con el hecho de aplicar Mapper en el caso de EE. UU, el cual no lleva implícita la filtración, y en el caso de UK aplicar BM, algoritmo que no necesita una previa filtración.

Como se explica en el documento, algunos condados tienen un número desproporcionadamente grande de casos acumulativos de COVID-19, los cuales aumentan significativamente el rango de los datos y, por lo tanto, se les da más peso en el proceso de normalización. Como resultado, su presencia en la nube de puntos reduce la importancia numérica de otros puntos de datos y los hace bastante indistinguibles en el gráfico Mapper. Para resolver este problema, se ha filtrado el conjunto de datos produciendo diferentes gráficos con o sin estos valores atípicos. Esto proporciona información más útil sobre los lugares que no están en la primera posición en términos de casos totales de COVID-19 pero que aún muestran tendencias preocupantes o importancia regional.

Capítulo 6. Aplicación de Ball Mapper

Una vez han sido explicados todos los conceptos necesarios para entender el algoritmo BM, se pasan a realizar una serie de análisis sobre diferentes conjuntos de datos, con el objetivo de sacar conclusiones que aún no han sido contempladas. A continuación, se presenta un experimento pequeño e ilustrativo que sigue una orientación similar al estudio que se realizó para la población de UK, (Instituto de Salud Carlos III, 2020).

Estudio de la seroprevalencia

El 6 de Julio de 2020 el Instituto de Salud Carlos III, en colaboración con el Ministerio de Sanidad y los Servicios de Salud de todas las Comunidades Autónomas, publica el informe final de la primera fase del Estudio Nacional de sero-Epidemiología de la infección por SARS-CoV-2 en España, del que han sido extraídos los datos para el primer análisis de este estudio, (Instituto de Salud Carlos III, 2020).

(Instituto de Salud Carlos III, 2020. Pág. 1): *ENE-COVID es un amplio estudio longitudinal sero-epidemiológico, de base poblacional, cuyos objetivos son estimar la prevalencia de infección por SARS-CoV-2 mediante la determinación de anticuerpos frente al virus en España y evaluar su evolución temporal.*

Esta primera ronda, que va desde el 27 de abril hasta el 11 de mayo de 2020, incluye 68.296 participantes en total. (Instituto de Salud Carlos III, 2020. Pág. 27): *ENE-COVID representa a la población general española no institucionalizada. Colectivos tan importantes como las personas mayores y las personas dependiente que viven en residencias o en otro tipo de instituciones no están representadas en este estudio.* **[…]** *tampoco representa adecuadamente a otros colectivos de especial interés, como son los profesionales sanitarios, las personas que trabajan en residencias y otro tipo de centros asistenciales, las fuerzas de seguridad, los conductores de transporte público y otros. Aunque el estudio cuenta con participantes de estos sectores esenciales, la muestra resulta insuficiente para caracterizarlos adecuadamente, lo que requeriría también cuestionarios específicos con mayor grado de detalle.*

En la ***Ilustración 5*** se visualiza el conjunto de datos que se usará de objeto en el primer estudio realizado con el algoritmo BM. Dicho conjunto cuenta con las columnas referentes al sexo (*sex*), la edad (*age.cat*), la cantidad de individuos estudiados en cada caso (*pop*), la media de la prevalencia del virus en cada grupo (*prev*), el número de infectado con respecto al conjunto de individuos estudiados para cada caso (*infect*), el número de individuos fallecidos a causa del virus (*death.covid*) y el número de individuos fallecidos debido a causas externas al Covid-19 (*death.excess*) del conjunto entero de datos.

	sex	age.cat	pop	prev	infect	death.covid	death.excess
2	1	1	2205512	0.0324910	71659.29	3	32
3	1	2	2557872	0.0364694	93284.06	3	0
4	1	3	2479132	0.0575589	142696.11	18	0
5	1	4	2978715	0.0469104	139732.71	48	3
6	1	5	3916706	0.0533675	209024.81	192	168
7	1	6	3493827	0.0526599	183984.58	705	601
8	1	7	2598212	0.0489133	127087.12	1904	2065
9	1	8	1783664	0.0468757	83610.50	4145	5114
10	1	9	993308	0.0459045	45597.31	5299	7497
12	6	1	2078268	0.0423326	87978.49	2	11
13	6	2	2396681	0.0438495	105093.26	3	22
14	6	3	2404106	0.0571521	137399.71	17	10
15	6	4	3012421	0.0520345	156749.82	29	71
16	6	5	3877832	0.0533281	206797.41	103	91
17	6	6	3563505	0.0517353	184359.00	318	369
18	6	7	2803368	0.0500889	140417.62	749	875
19	6	8	2138101	0.0462329	98850.61	1986	2646
20	6	9	1605845	0.0499314	80182.09	3704	5203

Ilustración 5. Conjunto de datos para el estudio de la seroprevalencia

Se realiza un tratamiento previo de los datos de manera que se obtienen las **medidas de tendencia central** que facilitarán el estudio inicial del conjunto de datos y nos servirá para demostrar si las variables son significativas y dicho conjunto es consistente. Estas medidas son: las media, la mediana, la moda y el rango.

	sex	*age.cat*	*pop*	*prev*	*infect*	*death.covid*	*death.excess*
Media	3.5	5	2604838	0.0482	127472.5	1068.222	1376.556
Mediana	3.5	5	2518502	0.0494	132243.4	147.500	129.500
Moda	1	1	2205512	0.0324	71659.29	3	0
Rango	5	8	2923398	0.0250	163427.5	5297	7497

Tabla 8. Estadísticas descriptivas para el conjunto de datos del estudio de seroprevalencia

Caso 1: (ε = 0.5)

	sex	age.cat	pop	prev	infect	death.covid	death.excess
2	0	0.1111111	0.05296516	0.0000000	0.02256640	8.120179e-05	0.0020671835
3	0	0.2222222	0.06836092	0.1587050	0.04129073	8.120179e-05	0.0000000000
4	0	0.3333333	0.06492051	1.0000000	0.08407536	1.299229e-03	0.0000000000
5	0	0.4444444	0.08674893	0.5752137	0.08150943	3.735282e-03	0.0001937984
6	0	0.5555556	0.12773282	0.8327981	0.14150767	1.542834e-02	0.0108527132
7	0	0.6666667	0.10925586	0.8045708	0.11982598	5.708486e-02	0.0388242894
8	0	0.7777778	0.07012351	0.6551127	0.07055994	1.544458e-01	0.1333979328
9	0	0.8888889	0.03453324	0.5738295	0.03291465	3.364190e-01	0.3303617571
10	0	1.0000000	0.00000000	0.5350867	0.00000000	4.301259e-01	0.4843023256
12	1	0.1111111	0.04740545	0.3925977	0.03669678	0.000000e+00	0.0007105943
13	1	0.2222222	0.06131796	0.4531094	0.05151602	8.120179e-05	0.0014211886
14	1	0.3333333	0.06164238	0.9837721	0.07948934	1.218027e-03	0.0006459948
15	1	0.4444444	0.08822165	0.7796225	0.09624410	2.192448e-03	0.0045865633
16	1	0.5555556	0.12603429	0.8312264	0.13957903	8.201380e-03	0.0058785530
17	1	0.6666667	0.11230032	0.7676870	0.12015018	2.565976e-02	0.0238372093
18	1	0.7777778	0.07908744	0.7020093	0.08210247	6.065773e-02	0.0565245478
19	1	0.8888889	0.05001975	0.5481871	0.04611067	1.611043e-01	0.1709302326
20	1	1.0000000	0.02676375	0.6957264	0.02994607	3.006090e-01	0.3361111111

Ilustración 6. Primer conjunto de datos para el estudio por individuos

Como se puede observar, las columnas están normalizadas para que los resultados obtenidos no presenten distorsión debida a diferencias en los rangos de las diferentes variables. La normalización es muy común en las tareas de preprocesado de datos, pero no es siempre necesaria. Sin embargo, sí es útil en casos como este en el que las variables presentan escalas muy diferentes.

Tras el tratamiento del conjunto de datos, se pasa a la elección del valor de ε. Para ello, se decide probar con el valor intermedio 0.5, ya que todas las columnas están normalizadas de 0 a 1 y este es el primer caso del estudio.

Una vez aplicado BM con un valor para ε = 0.5, se observa que el recubrimiento obtenido queda de la siguiente manera:

Nº bola	*Grupo(s) cubierto(s)*
1	Hombres entre 0-19 años
2	Hombres entre 20-39 años
3	Hombres entre 40-59 años
4	Hombres entre 60-69 años
5	Hombres entre 70-79 años
6	Hombres mayores de 80 años
7	Mujeres entre 0-19 años
8	Mujeres entre 20-39 años
9	Mujeres entre 30-59 años
10	Mujeres entre 30-39 y 50-69 años
11	Mujeres entre 70-79 años
12	Mujeres mayores de 80 años

Tabla 9. Recubrimiento para el problema por individuos. Caso 1

Antes de comenzar con el análisis general de las representaciones obtenidas, podemos atender a ciertas características que se deducen mirando las agrupaciones obtenidas. Por ejemplo, las mujeres entre 30-39 años se encuentran englobadas en tres bolas (bolas 8, 9 y 10). Por tanto, como se produce solapamiento, sabemos que se formará una arista que una estas dos bolas, marcando visualmente dicho solapamiento. Este fenómeno está muy relacionado con el valor de épsilon escogido, ya que es el valor que marca en qué medida se pueden considerar '*similares*' dos miembros del conjunto de datos para agruparlos en la misma bola o no.

Al aplicar dicho algoritmo tomando como valores cada una de las columnas de nuestro conjunto de datos sucesivamente, se obtienen 7 ilustraciones con la misma representación de la distribución por clústeres, en las cuales van variando los colores de cada uno de ellos en función de los distintos valores que se toman. Gracias a estas variaciones se pueden obtener determinadas conclusiones como las que se describen posteriormente. Además de esto, también podemos observar variaciones en el tamaño de las bolas, lo que indica cuántos miembros del conjunto de datos están agrupados en un mismo conjunto.

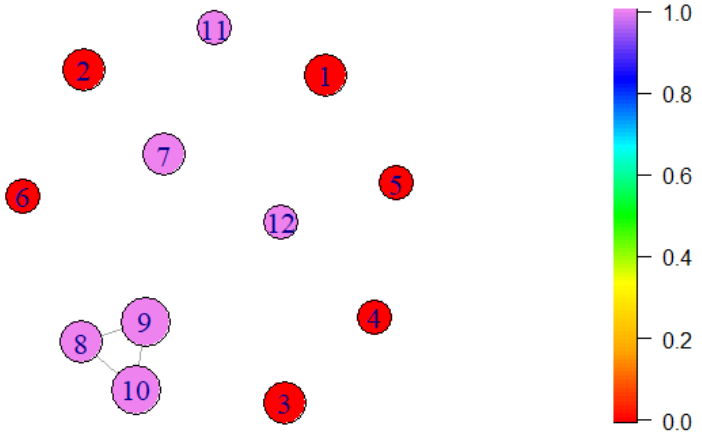


Ilustración 7. Columna referida al sexo. estudio por individuos. Caso 1

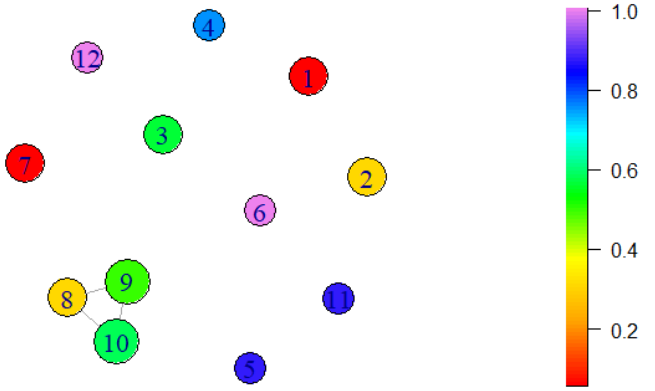


Ilustración 8. Columna referida a la edad. Estudio por individuos. Caso 1

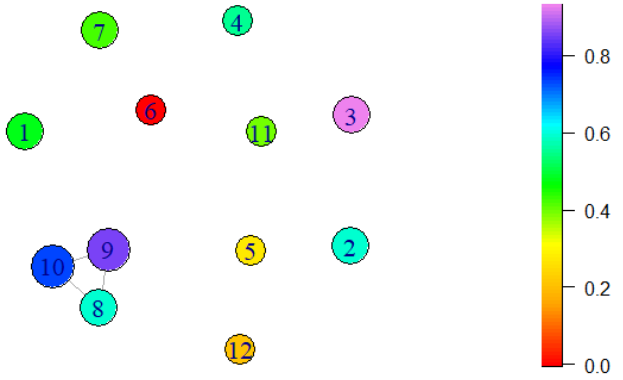


Ilustración 9. Columna referida a la población. Estudio por individuos. Caso 1

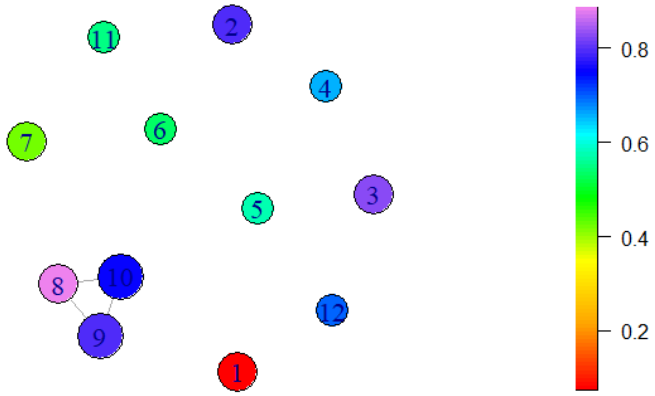


Ilustración 10. Columna referida a la prevalencia. Estudio por individuos. Caso 1

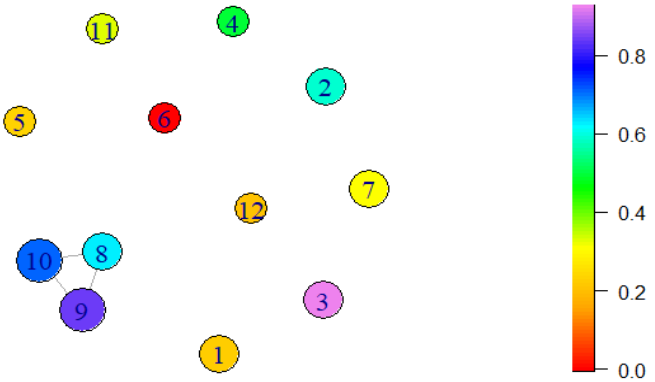


Ilustración 11. Columna referida al número de infectados. Estudio por individuos. Caso 1

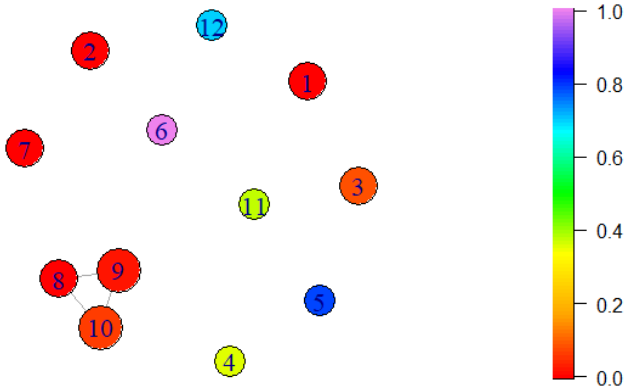


Ilustración 12. Columna referida al número de fallecimientos por covid. Estudio por individuos. Caso 1

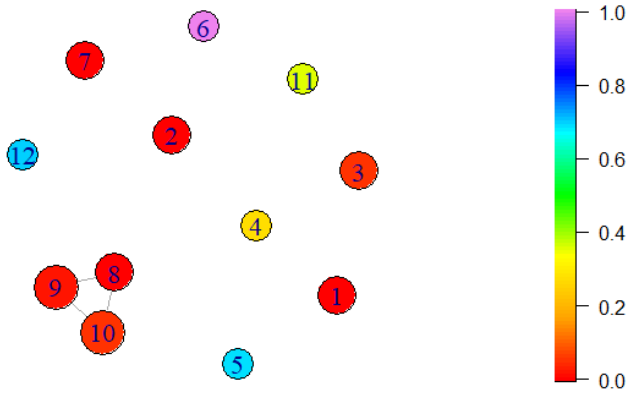


Ilustración 13. Columna referida al número de fallecimientos por causas ajenas al covid. Estudio por individuos. Caso 1

Análisis de individuos femeninos

Como podemos ver, los grupos de mujeres entre 20 y 69 años, que corresponden a las bolas 8-9-10, están muy relacionados. Además, atendiendo al gráfico correspondiente a la población, ***Ilustración 9***, esta agrupación representa un porcentaje alto del conjunto de individuos estudiado, con un intervalo del [0.6, 0.9]. Nos centraremos, por tanto, en analizar estos tres grupos inicialmente.

Si nos fijamos en las visualizaciones correspondientes al número de infectados y al número de muertes por COVID, podemos ver que el número de mujeres infectadas es alto, puesto que entra dentro del rango [0.6, 0.8], sin embargo, en contraste con este valor, el número de muertes es relativamente pequeño, entrando en el intervalo [0, 0.1]. Por tanto, se puede observar que la mayoría de las mujeres que se infecta no muere por COVID. Además, el número de exceso de muertes tampoco es alto, por lo que se puede afirmar que la mayor parte de mujeres infectadas supera el virus, aunque la prevalencia del virus en ellas es más alta (rango entre [0.7, 1]).

Centramos ahora nuestra atención en las bolas que cubren el resto de las mujeres, que son: la bola 7 correspondiendo a las de edad entre 0-19 años, la 11 a las mujeres entre 70-79 años, y la 12 a las mayores de 80 años.

Como podemos observar, estos tres conjuntos cubren también una parte considerable de la población de estudio, siendo la bola 12 la de menor proporción. La infección en estos casos es pequeña, entrando dentro del intervalo [0.1, 0.3]. Sin embargo, aunque haya menos infecciones en estos rangos de edades, la prevalencia del virus mantiene una proporción inversa con este valor. Como podemos ver en la visualización referida a la prevalencia, ***Ilustración 10***, el valor de la medida aumenta con respecto al valor de

infección, es decir, aunque haya menos cantidad de mujeres infectadas, el virus perdura más. Además, como ya sabíamos, las muertes por COVID en los grupos de mujeres de mayor edad son considerablemente altas con respecto al número de infecciones.

Análisis de individuos masculinos

Fijándonos primero en el porcentaje de población masculina del estudio al que corresponde cada grupo, podemos ver que la mayor parte de los hombres que participan están entre 20-59 años. Si nos fijamos únicamente en estos grupos, podemos observar que la infección en ellos es alta, comprendiendo un intervalo entre [0.6, 1]. Además, la prevalencia del virus en dichos individuos es también alta, pero no tanto como el número de infectados. Las muertes, sin embargo, son pocas, entrando dentro del rango [0, 0.1].

En orden de número de individuos, sigue el grupo de edad entre 60-69 años. Este conjunto muestra unas estadísticas un poco más bajas para la infección, es decir, en la porción de población que cubre estos individuos no se contagian todos, pero la cantidad de muertes no decae como en el caso anterior. Pasa algo parecido con los hombres de 70-79 años, que, aunque corresponden a una proporción menor de la población, el número de muertes por COVID es considerablemente alto.

Podemos concluir que es más peligroso el contagio entre hombres de 60-79 años.

En cuanto a los hombres entre 0-19 años (bola 1), aunque cubren un porcentaje medio de la población, el número de infecciones es más pequeño y el número de muertes es prácticamente nulo. Además de esto, es el grupo que cuenta con menor prevalencia ante el virus.

Por último, los hombres mayores a 80 años son pocos en este estudio y, por tanto, los contagios tienen una proporción baja también respecto a los demás grupos. Sin embargo, aunque el exceso de muertes es alto debido a la edad, las muertes por COVID lo son también en comparación al número de infecciones.

Conclusiones caso 1

- La mayoría de las **mujeres entre 20 y 69 años** de este estudio que se **infecta no muere** por COVID.
- La mayor parte de **mujeres infectadas supera el virus**, aunque la **prevalencia** del virus en ellas es más **alta**. Esto es algo que refuta alguna de las conclusiones del estudio previo. (Instituto de Salud Carlos III, 2020. Pág. 16): *No se aprecian diferencias importantes entre hombres y mujeres.*

- En los **hombres** que participan y están **entre 20-59 años**, la **infección** y **prevalencia** del virus en dichos individuos es **alta**.
- Los **hombres mayores de 60 años**, aunque corresponden a una proporción pequeña de la población estudiada, cuentan con un **número de muertes** por COVID es considerablemente **alto**.
- Para los **jóvenes entre 0-19 años** el **número de infecciones** es más **pequeño** y el **número de muertes** es prácticamente **nulo**.

Algo que hay que tener en cuenta mientras se realizan distintos experimentos sobre nuestro conjunto de datos inicial (Instituto de Salud Carlos III, 2020. Pág. 16): *Todas estas estimaciones han de ser interpretadas teniendo en cuenta que se trata de información autorreportada, siempre sujeta a un posible sesgo de recuerdo. Los resultados sobre contacto con casos COVID-19 confirmados o posibles también están limitados por el conocimiento que el participante pueda tener de dicha situación, lo que a su vez depende del tipo de relación entre ambos.*

Caso 2: (ε = 0.6)

Tras realizar el análisis del caso anterior, se decide fijar ahora ε = 0.6, valor con el cual surgen ya más conexiones entre grupos.

Una vez aplicado BM al conjunto de datos de la ***Ilustración 6***, con un valor para ε de 0.6, se observa que el recubrimiento obtenido queda de la siguiente manera:

Nº bola	***Grupo(s) cubierto(s)***
1	Hombres entre 0-19 años
2	Hombres entre 20-39 años
3	Hombres entre 40-59 años
4	Hombres entre 50-69 años
5	Hombres entre 70-79 y mayores de 80
6	Mujeres entre 0-19 años
7	Mujeres entre 10-39 años
8	Mujeres entre 30-59 años
9	Mujeres entre 30-39 y entre 50-79 años
10	Mujeres entre 70-79 y mayores de 80

Tabla 10. Recubrimiento para el problema por individuos. Caso 2

De la misma manera que en el caso anterior y para todos los casos posteriores relativos a este estudio, se aplica el algoritmo tomando como valores cada una de las columnas de nuestro conjunto de datos sucesivamente y se obtienen 7 ilustraciones de nuevo.

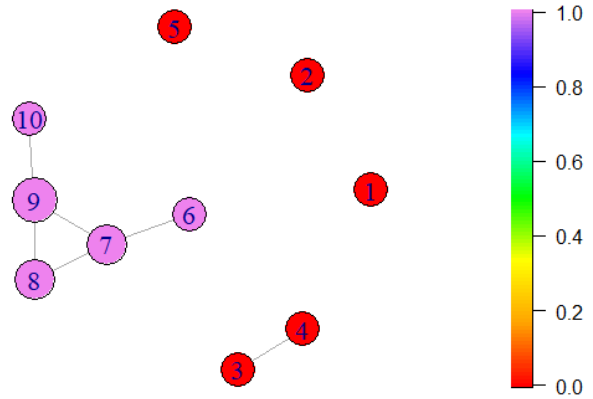


Ilustración 14. Columna referida al sexo. Estudio por individuos. Caso 2

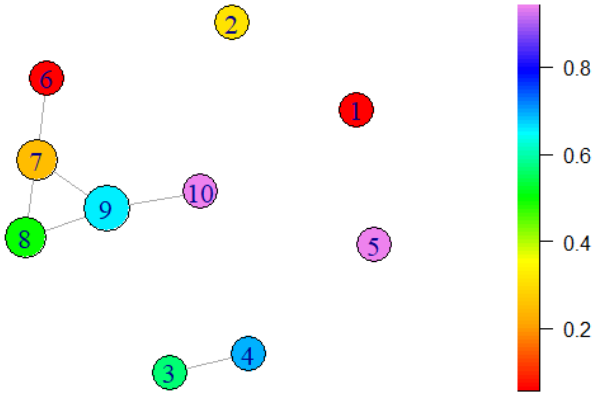


Ilustración 15. Columna referida a la edad. Estudio por individuos. Caso 2

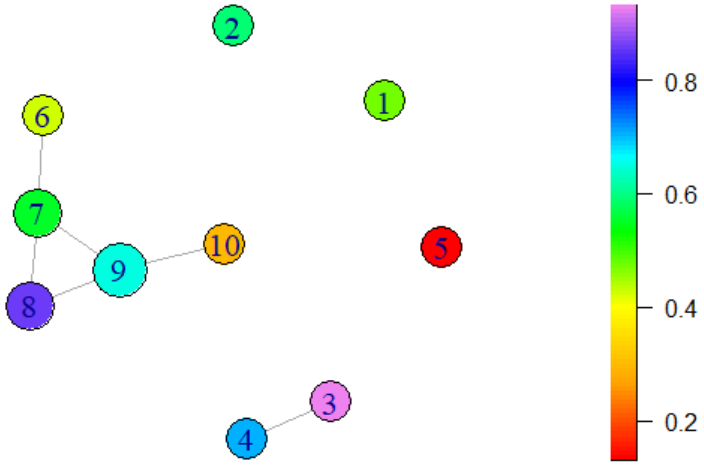


Ilustración 16. Columna referida a la población. Estudio por individuos. Caso 2

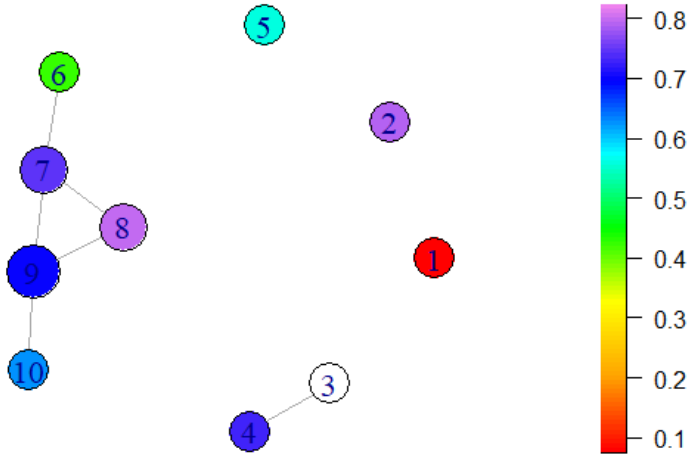


Ilustración 17. Columna referida a la prevalencia. Estudio por individuos. Caso 2

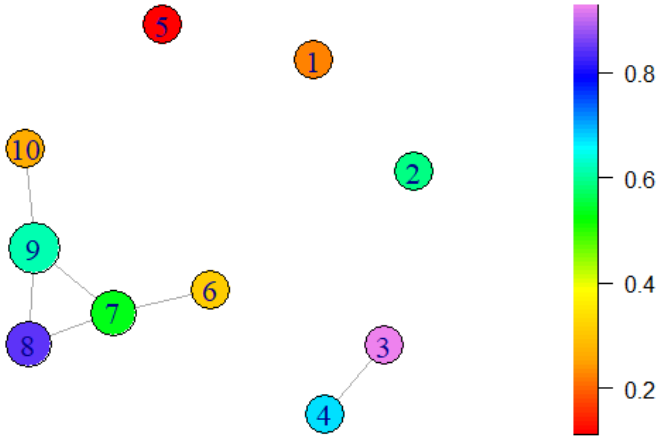


Ilustración 18. Columna referida al número de infectados. Estudio por individuos. Caso 2

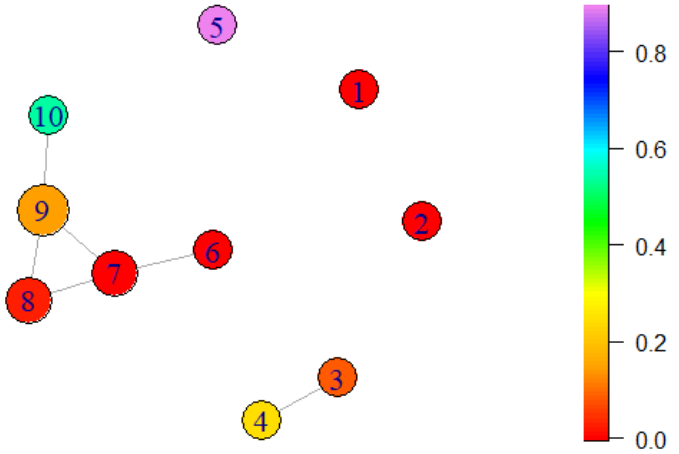


Ilustración 19. Columna referida al número de fallecimientos por covid-19. Estudio por individuos. Caso 2

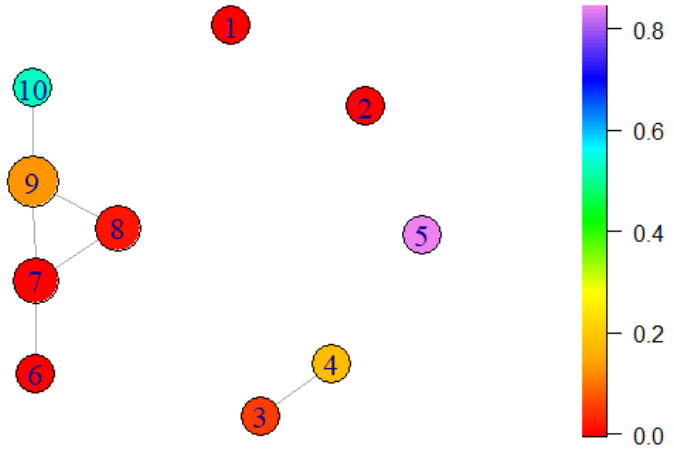


Ilustración 20. Columna referida al número de fallecimientos por causas ajenas al covid-19. Estudio por individuos. Caso 2

Como se puede observar, se forman **cinco clústeres**, el mayor correspondiente a las bolas que cubren todos los grupos de mujeres, otro formado por dos bolas, correspondientes a los grupos de hombres de entre 40 y 69 años y tres más individuales, con los grupos restantes. Estos clústeres corresponden a las bolas entre las que se produce solapamiento. Por tanto, nos centraremos sobre todo en el análisis de estos dos clústeres, ya que puede que sean los que más información nueva puedan aportar. Esta vez no se considera necesario el análisis por separado de ambos grupos según el sexo, debido a que el mayor clúster que se forma engloba individuos tanto del sexo masculino, como del sexo femenino. Por ello, puede que se encuentre una mayor correlación entre conjuntos sin la necesidad de que tengan que pertenecer a un sexo o a otro.

Siguiendo un orden, comenzamos por fijarnos en la representación del algoritmo correspondiente a la prevalencia, ***Ilustración 17***. Como se puede observar, la prevalencia del virus es muy alta en todos los grupos exceptuando la bola 1 (Hombres 0-19) y la bola 6 (Mujeres 0-19). Aunque es verdad que la tasa de prevalencia es más reducida en ambos grupos, la diferencia entre estos dos valores es aproximadamente de 0.4, siendo un valor comprendido entre [0.4, 0.5] en la bola 6 y teniendo un valor prácticamente nulo en la bola 1. Por tanto, podríamos destacar que el virus permanece más tiempo en las mujeres que en los hombres en un rango de edad de entre 0 y 19 años.

Además, algo que salta a simple vista es el hecho de que, en este gráfico, la bola 3, representante del grupo de hombres entre 40-59 años, no tiene color. Si analizamos las bolas que cubren los grupos masculinos, nos damos cuenta de que era una de las bolas

con solapamiento, ya que cuenta con el grupo de 50-59 años incluido también en la 4, pero es la única que cubre el grupo de 40-49 años. Si seguimos analizando las distintas representaciones intentaremos averiguar la causa de este fenómeno y si existen más anomalías relacionadas con este grupo.

Continuando con nuestro análisis, vamos a pasar a realizar ahora un contraste entre las gráficas correspondientes a la población y al número de infectados. Generalizando, podemos observar que ambas visualizaciones son muy parecidas, ya que casi todas las bolas cuentan con un valor en la gráfica de infección un poco más pequeño del que tienen en la de población, aunque, las bolas 2 y 3, que cubren los grupos masculinos de entre 20-39 años y 40-59 años respectivamente, cuentan con el mismo color en ambas. Podemos entender, entonces, que ambas variables están muy relacionadas.

Sin embargo, existe también una excepción en este caso. Se trata de la bola número 1 que cuenta con una tasa de infectados muy baja (entre [0, 0.2]) en comparación con la tasa de población, que cuenta con un valor entre el intervalo [0.4, 0,6]. Como comentamos en la primera parte, nos dimos cuenta de que el virus permanecía más tiempo en el grupo femenino de entre 0 y 19 años que en el masculino de esta misma edad, ahora podemos añadir a esta reflexión que, además de permanecer más, la tasa de infectados es mayor en el grupo femenino contando con menos porcentaje de población estudiada que el grupo masculino.

Finalmente, vamos a hacer un contraste entre las representaciones de infección y número de muertes por COVID. Principalmente, podemos ver que las bolas 1, 2, 6, 7 y 8 cuentan con un valor numérico prácticamente nulo, independientemente de la tasa de infección. Pero, si nos fijamos ahora en las bolas 5 (Hombres mayores de 70 años) y la 10 (Mujeres mayores de 70 años), podemos ver que la tasa de infección en el grupo masculino es aparentemente nula y el número de muertes por el virus tiene el valor máximo, por lo que se podría afirmar que es el grupo que cuenta con más riesgo de muertes por COVID. Además, atendiendo al grupo femenino de este rango de edad, podemos ver que el número de muertes es también elevado comparado con el número de infectados, pero debemos tener en cuenta que el número de infectados en este grupo es mayor al del grupo masculino.

Conclusiones caso 2

A las conclusiones extraídas en el caso anterior podemos añadir las siguientes:

- En cuanto a la **prevalencia** del virus, los grupos de edad **entre 0-19 años** cuentan con el **porcentaje más pequeño** de esta medida, (Instituto de Salud Carlos III, 2020. Pág. 16): *Los niños muestran prevalencias más bajas de anticuerpos.* Sin embargo, algo que no se había observado antes es que se observan **diferencias** considerables entre **hombres y mujeres**, siendo en los **hombres**, la tasa de prevalencia, prácticamente **nula**, mientras que en las **mujeres** cuenta con un **valor medio** en el rango de valores.
- La **tasa de infectados es mayor en el grupo femenino** contando con menos porcentaje de población que el grupo masculino.
- Para los **individuos masculinos mayores de 70** años incluidos en los datos para nuestro estudio, la **tasa de infección es aparentemente nula** y el **número de muertes por el virus cuenta con el valor máximo**, es decir, la posibilidad de morir al ser contagiado por el COVID-19 a partir de esta edad es más alta en hombres que en mujeres.

Caso 3: (ε = 0.55, reducción del conjunto de datos)

Continuando con los análisis anteriores, vamos a realizar ahora una prueba eliminando la columna referida al exceso de muertes, ya que no proporciona información relevante sobre el tema estudiado, y reduciendo el valor de ε a 0.55. Se procede a la variación de ε puesto que el simple hecho de reducir el conjunto de datos manteniendo dicho valor en 0.6 no aportaba mayor información que la analizada en el ***Caso 2***.

Quedarían, por tanto, nuestros datos de estudio de la siguiente forma:

	sex	age.cat	pop	prev	infect	death.covid
3	0	0.000	0.4146558	0.0000000	0.1594712	0.0001887861
4	0	0.125	0.5351868	0.1587050	0.2917915	0.0001887861
5	0	0.250	0.5082524	1.0000000	0.5941399	0.0030205777
6	0	0.375	0.6791436	0.5752137	0.5760071	0.0086841608
7	0	0.500	1.0000000	0.8327981	1.0000000	0.0358693600
8	0	0.625	0.8553468	0.8045708	0.8467808	0.1327166321
9	0	0.750	0.5489858	0.6551127	0.4986298	0.3590711724
10	0	0.875	0.2703553	0.5738295	0.2325997	0.7821408344
11	0	1.000	0.0000000	0.5350867	0.0000000	1.0000000000
13	1	0.000	0.3711298	0.3925977	0.2593271	0.0000000000
14	1	0.125	0.4800486	0.4531094	0.3640511	0.0001887861
15	1	0.250	0.4825884	0.9837721	0.5617316	0.0028317916
16	1	0.375	0.6906733	0.7796225	0.6801335	0.0050972248
17	1	0.500	0.9867025	0.8312264	0.9863707	0.0190673966
18	1	0.625	0.8791813	0.7676870	0.8490719	0.0596564093
19	1	0.750	0.6191630	0.7020093	0.5801980	0.1410232207
20	1	0.875	0.3915967	0.5481871	0.3258528	0.3745516330
21	1	1.000	0.2095291	0.6957264	0.2116216	0.6988861620

Ilustración 21. Segundo conjunto de datos para el estudio por individuos

Aplicando nuestro algoritmo con ε = 0.55, se observa que el recubrimiento obtenido queda de la siguiente manera:

Nº bola	***Grupo(s) cubierto(s)***
1	Hombres entre 0-19 años
2	Hombres entre 20-39 años
3	Hombres entre 40-59 años
4	Hombres entre 30-39 y 60-69 años
5	Hombres entre 70-79 y mayores de 80
6	Mujeres entre 0-19 años
7	Mujeres entre 20-39 años
8	Mujeres entre 30-59 años
9	Mujeres entre 30-39 y entre 50-79 años
10	Mujeres entre 70-79 y mayores de 80

Tabla 11. Recubrimiento para el problema por individuos. Caso 3

Inicialmente, observando simplemente el recubrimiento producido a partir del cambio del valor de ε y la reducción de datos, podemos ver que, esta vez, no se produce tanto solapamiento entre algunas bolas, por ejemplo, en la bola 4, queda excluido el grupo

masculino entre 50-59 años, que antes estaba englobado tanto en la bola 3 como en la 4 y ahora solo está en la 3, aunque continua la inclusión del grupo referido al rango de edad 30-39 años, que está incluido también en la bola 2.

En los conjuntos femeninos pasa algo parecido. Si observamos la bola 7, ya no cuenta con el grupo de entre 10-19 años, que solo queda dentro de la bola número 1, pero, sin embargo, el grupo correspondiente a 30-39 años aún sigue siendo englobado en tres bolas diferentes, como en el caso anterior, las bolas 7, 8 y 9. Podríamos confirmar nuestra intuición del caso 1 de que los grupos femeninos comparten más correlación que los masculinos, debido a que sigue existiendo este solapamiento mencionado.

Veamos si, aunque la variación ha sido pequeña, podemos extraer más información en este caso.

Las representaciones para este caso son las siguientes:

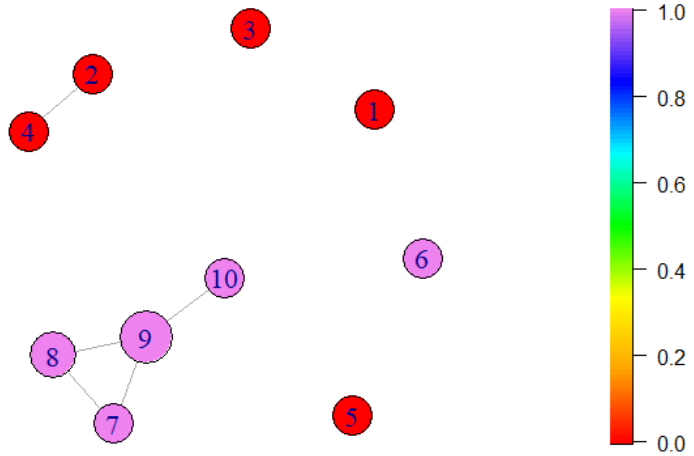


Ilustración 22. Columna referida al sexo. Estudio por individuos. Caso 3

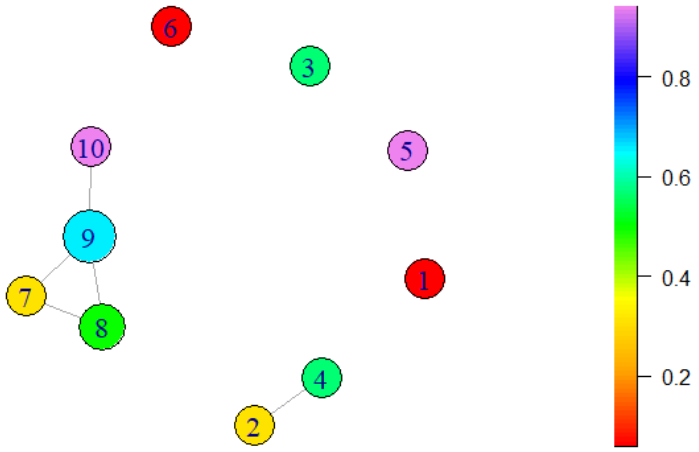


Ilustración 23. Columna referida a la edad. Estudio por individuos. Caso 3

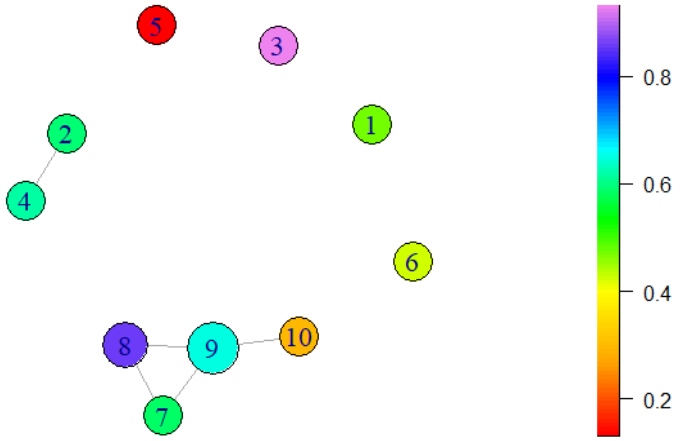


Ilustración 24. Columna referida a la población. Estudio por individuos. Caso 3

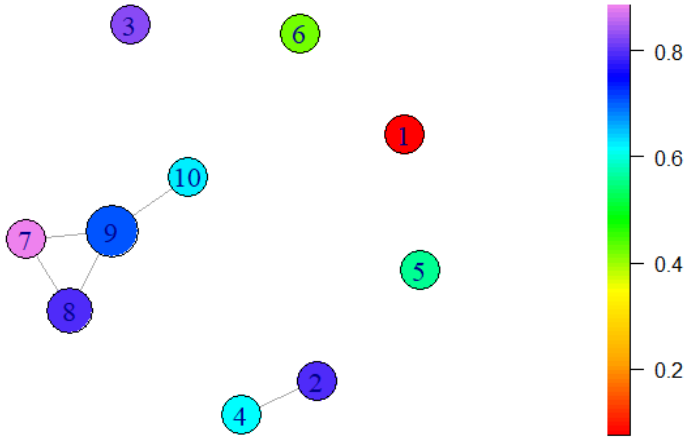


Ilustración 25. Columna referida a la prevalencia. Estudio por individuos. Caso 3

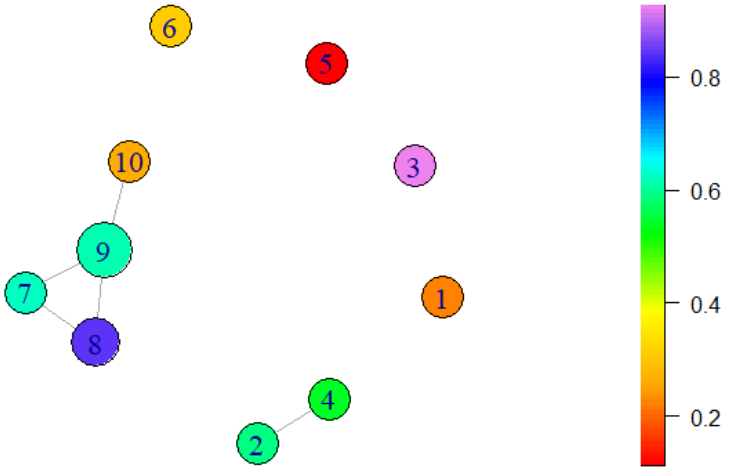


Ilustración 26. Columna referida al número de infectados. Estudio por individuos. Caso 3

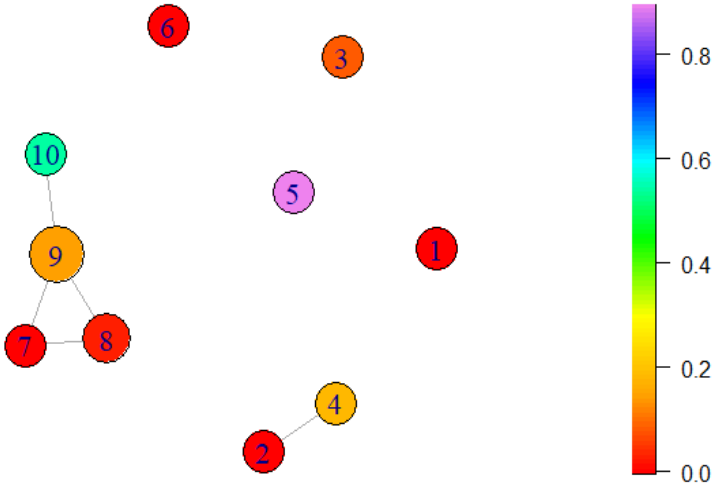


Ilustración 27. Columna referida al número de fallecimientos por covid-19. Estudio por individuos. Caso 3

En primer lugar, observemos la gráfica referida a la prevalencia, ***Ilustración 25***. Recordamos que nuestro problema principal a resolver viene de la mano de esta gráfica, ya que en el ejemplo anterior no conseguíamos que la bola 3 (representativa, en los tres casos, del grupo masculino entre 40-59 años) estuviera coloreada. Como podemos ver, ahora sí que cuenta con su respectivo color, marcando un valor de 0,9 aproximadamente. Si hubiera cambiado la distribución de individuos que se incluyen en la bola, podríamos detectar rápidamente el problema. Sin embargo, esto no ha ocurrido, aunque podemos atender a los cambios producidos en otras bolas en cuanto al nuevo recubrimiento. Como hemos mencionado, el cambio producido es referido al grupo masculino comprendido entre 50-59 años, por lo que podríamos tener sospecha de que el decolorado anterior de la bola en cuestión vendría arraigado a este grupo y al cambio de ε, por supuesto.

Siguiendo con el estudio de la prevalencia, en las bolas que cubren los grupos de edad 0-19 no se encuentra ninguna discrepancia, seguimos advirtiendo que, en esta edad, las mujeres cuentan con un nivel más alto de prevalencia del virus que los hombres, para los que es prácticamente nula.

El único cambio significativo que podemos percibir es el hecho de que las bolas número 2 y 7 se diferencian en valor de manera que el valor de la primera es más pequeño que el de la segunda, cosa que antes era al revés. Si nos fijamos en los cambios de recubrimiento mencionados, vemos que ahora el grupo femenino de 10-19 años no está incluida en la bola 7, por lo que el valor ha aumentado para ésta.

Finalmente, como en los casos anteriores, lo único que podemos sacar como conclusión de la prevalencia, es que en las mujeres suele ser mayor que en los hombres,

independientemente de la edad, y que en los grupos de edad comprendidos entre los 20-69 años prevalece, también, más tiempo que en los individuos entre 0-19 y mayores de 70 años.

Pasemos, ahora, al análisis de la visualización correspondiente al número de infectados respecto a la población, ***Ilustración 26*** e ***Ilustración 24*** respectivamente.

Los resultados para esta gráfica son prácticamente iguales a la gráfica anterior y, como ya sabíamos, esta variable está fuertemente relacionada con la variable de población. Como podemos observar, las edades correspondientes al rango de entre 0-19 años, cuentan con una tasa de infección menor que la de población, pero las demás edades tienen un número de infectados muy parecido, o igual, al número de individuos del estudio en cada edad. Podría confirmarse, entonces, que es más difícil contagiarse contando con una edad reducida.

Por último, cabe resaltar un detalle que puede pasar desapercibido. En el conjunto de estudio que queda excluido en este caso (individuos entre 0-19 años) se observa que la tasa de infección tiene un valor mayor para el grupo femenino que para el grupo masculino, mientras que, en cuanto a la tasa de población, pasa lo contrario, el valor en este caso sería mayor para el grupo masculino que para el femenino. Por tanto, a nuestra última reflexión podríamos añadir que, por lo menos, en los rangos de edad pequeños, sería más fácil contagiarse siendo del sexo femenino que siendo del sexo masculino.

Finalmente, nos centramos en la última representación, la ***Ilustración 27***, referida al número de muertes por COVID-19. Como podemos notar, esta gráfica, al igual que la relativa a la tasa de infección, no ha variado mucho del caso anterior. Podemos observar que, para los grupos de edades comprendidos entre 0-59 años de este estudio, la tasa de muertes por el virus es generalmente nula, incluso teniendo en cuenta que la tasa de infección para todas las edades a partir de 19 años es elevada. En cuanto vamos acercándonos a edades más grandes esta tasa comienza a incrementarse. Pero esto ya lo sabíamos. Lo que no sabíamos es que, si nos fijamos en los valores, aunque sean prácticamente nulos para la mayoría de las bolas, éstos siempre son más elevados en los grupos de edad referidos a individuos masculinos que en los referidos a individuos femeninos. Podríamos concretar que es más frecuente morir por el virus siendo hombre antes que siendo mujer**.**

Conclusiones generales del estudio de la seroprevalencia

Generalizando, con estos análisis del estudio para la seroprevalencia a través del algoritmo BM podemos sacar diferentes conclusiones:

- La **prevalencia en las mujeres** suele ser **mayor** que en los hombres, independientemente de la edad.
- En los grupos de edad comprendidos **entre los 20-69 años prevalece el virus más tiempo** que en los individuos entre 0-19 y mayores de 70 años.
- Es **más probable contagiarse** siendo del **sexo femenino** que siendo del sexo masculino.
- Conforme **más reducida es la edad** de los individuos, **más difícil es contagiarse**.
- Cuanto **más nos acercamos a edades grandes**, la **tasa de mortalidad** comienza a **incrementarse**. Incluso hemos observado como estos **valores** siempre son **más elevados en los grupos de edad masculinos** que en los femeninos.

Estudio por comunidades autónomas.

Una vez realizado un pequeño experimento con nuestro algoritmo, pasamos a estudiar conjuntos de datos más grandes, como el de este estudio, de cara a corroborar algunas de las conclusiones extraídas en los casos anteriores y descubrir otras nuevas.

En esta ocasión, se han extraído los datos oficiales de COVID-19 en España, gracias a un repositorio de GitHub, (FERNÁNDEZ CASAL, 2020), en el que se explica tanto los datos que contiene, como de dónde se sacan.

Para cada caso de este estudio, según lo que interese visualizar y analizar, se hace un tratamiento y reajuste previo de los datos para su mejor exploración. Por ejemplo, en nuestro caso no nos interesaba desgranar en tanto detalle cada fila del conjunto de datos, con el objetivo de realizar un análisis más fácilmente, por lo que se obvia la columna de provincias, englobando los valores de cada provincia únicamente en la CCAA a la que pertenece, como suma total de dichos valores numéricos, y según el periodo de tiempo que se quiera investigar, se agrupan de la misma manera los datos correspondientes únicamente al rango de fechas escogido.

Inicialmente, se escogen datos con información sobre el COVID-19 durante todos los meses desde enero de 2020, organizados por Comunidades Autónomas, provincias, sexo, fecha y edad.

Caso global. Pre-vacunación: (ε = 0.6)

En este estudio, el conjunto de datos recoge información sobre el número de casos confirmados, hospitalizados, que tuvieron que entrar en la UCI y fallecidos, categorizados por la comunidad autónoma a la que pertenecen, el sexo y la edad.

En este caso se seleccionan los datos comprendidos en el rango de 01/02/2020 hasta 30/04/2020. En total, nos quedaría un conjunto de datos con 324 filas, de ahí no poder proporcionar una imagen general de dichos datos. Sin embargo, como al inicio del ***Estudio de la seroprevalencia****,* haremos un previo análisis descriptivo de las distintas columnas:

	confirmados	*hospitalizados*	*uci*	*fallecidos*
Media	676.845	317.163	27.259	80.043
Mediana	190.5	61	5	2
Moda	2	2	0	0
Rango	9070	4791	658	2411

Tabla 12. Estadísticas descriptivas para el conjunto de datos del estudio por comunidades

Al aplicar nuestro algoritmo al conjunto de datos mencionado anteriormente de la misma manera que para el estudio anterior, surgen ahora, 4 gráficas:

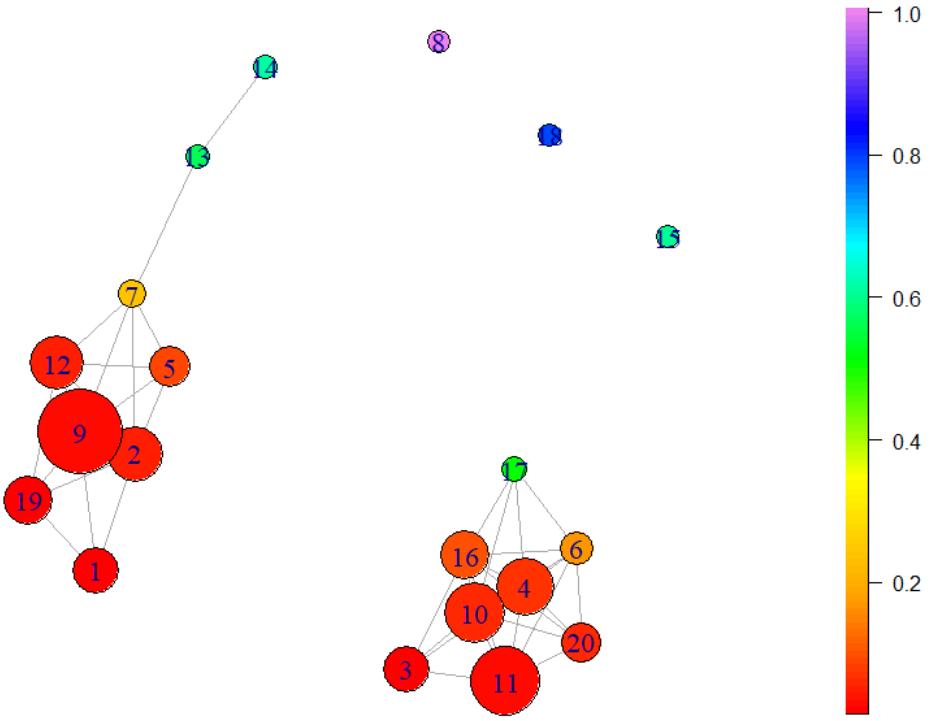


Ilustración 28. Columna referida al número de casos confirmados. Estudio por comunidades. caso global. Pre-vacunación

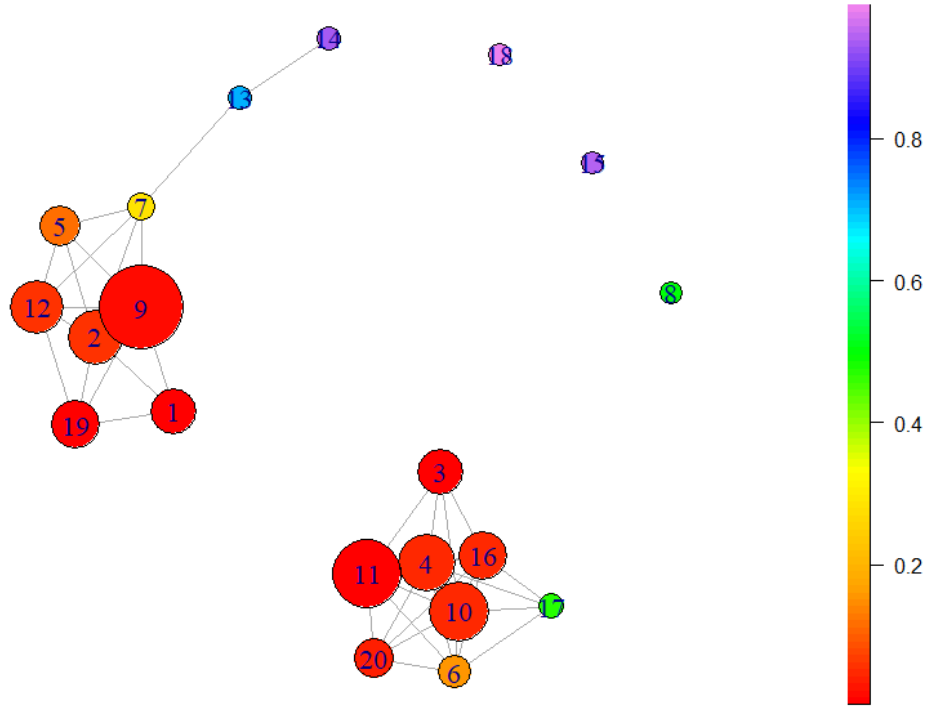


Ilustración 29. Columna referida al número de casos hospitalizados. Estudio por comunidades. Caso global. Pre-vacunación

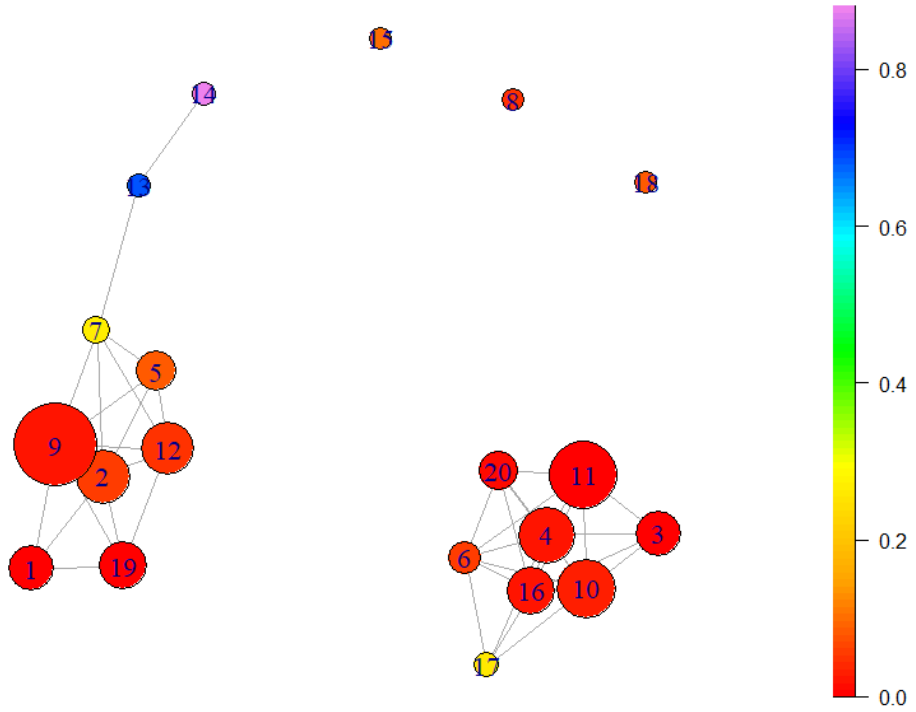


Ilustración 30. Columna referida al número de casos en la uci. estudio por comunidades. Caso global. Pre-vacunación

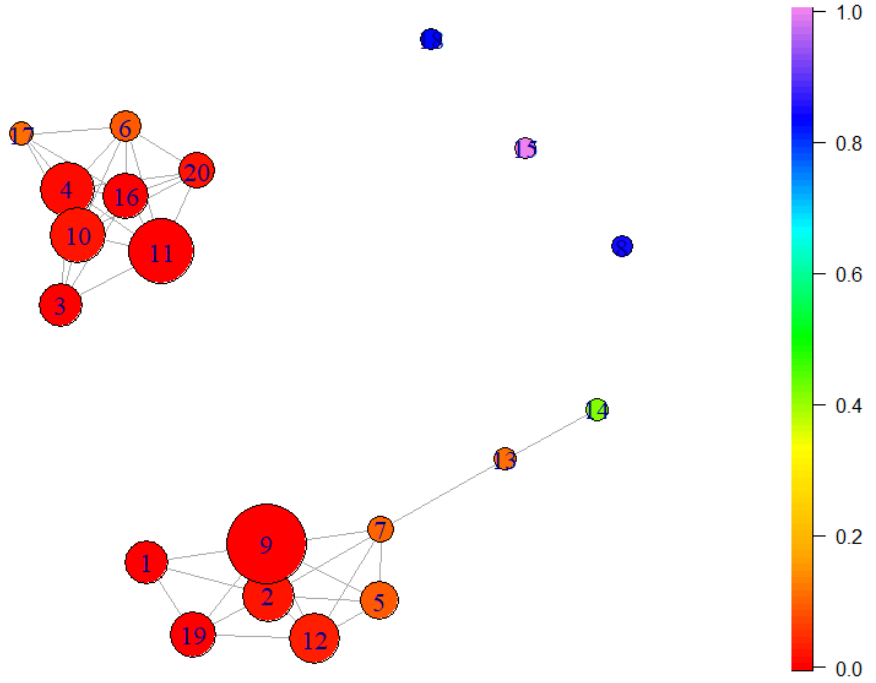


Ilustración 31. Columna referida al número de casos fallecidos. Estudio por comunidades. Caso global. Pre-vacunación

Análisis global de las bolas aisladas (8, 15 y 18)

Para comenzar, podemos ver que la mayor parte de las bolas que forman los clústeres conformados se comportan de manera similar en cada caso, demostrando que están relacionados estrechamente según las características estudiadas. Se forman dos clústeres principales y quedan aisladas las bolas 8, 15 y 18. Nos centraremos, por tanto, en analizar estas bolas inicialmente.

La bola 8 corresponde únicamente al grupo formado por mujeres de Cataluña mayores de 80 años. A su vez, la bola 15 engloba al grupo de hombres y la bola 18 al grupo de mujeres, ambas correspondientes a las categorías: Madrid, mayores de 80 años. Nos damos cuenta de que estas agrupaciones corresponden a los grupos de edad más afectados por el virus, concretamente en unas de las comunidades con más densidad de población de España, siendo Cataluña la segunda más poblada y Madrid la tercera. Estando en primer lugar Andalucía como la comunidad más densidad de población en España, pero debemos tener en cuenta la diferencia de territorio entre ésta y las dos ya mencionadas.

Observando cómo se comportan estas bolas respecto a las 4 variables numéricas estudiadas, podemos advertir que las bolas 8 y 18, en la ***Ilustración 28****,* entran en el intervalo de máximos valores, teniendo un color representativo de los valores más altos dentro de dicho intervalo la primera y un color representativo de los valores más bajos dentro de éste la segunda. Por otra parte, la bola 15 cuenta con un color categórico del intervalo medio [0.6, 0.8], representativo de los valores más bajos dentro de este rango.

Pasando a la comparación entre las diferentes visualizaciones, observamos que, siendo la bola 15 la que cuenta con el valor más pequeño de estas tres en cuanto al número de casos confirmados, cuenta con un valor más alto en la tasa de casos hospitalizados. De manera inversa, la bola 8, que cuenta con el máximo valor de estas tres agrupaciones en la ***Ilustración 28****,* toma el menor valor en la ***Ilustración 29****.* La bola 18 se mantiene bastante estable en esta comparación, contando con valores altos, comprendidos en el último intervalo, en ambas gráficas.

Por tanto, se podría observar cómo las mujeres ancianas residentes en Cataluña han contado con una incidencia muy alta de casos confirmados del virus, pero, sin embargo, solo alrededor de un 50% han sido hospitalizadas, mientras que, en el caso de las mujeres mayores de 80 años residentes en la Comunidad de Madrid, ante la alta tasa de casos confirmados, aunque menor que en Cataluña, prácticamente todos estos casos han sido hospitalizados.

Ambos grupos cuentan con un valor correspondiente, de nuevo, al máximo intervalo en la ***Ilustración 31***, por tanto, como ya sabíamos, la incidencia del virus en personas mayores es habitualmente mortífera. Sin embargo, si nos fijamos ahora en el grupo englobado por la bola 15, recordamos que contaba con el valor más pequeño de estos tres grupos respecto al número de casos confirmados, pero si centramos nuestra atención en las ilustraciones: ***Ilustración 29, Ilustración 31***, podemos observar cómo en estos casos, este grupo cuenta con un valor muy alto en la tasa de hospitalización y el valor más alto en el número de fallecidos. Por tanto, los hombres mayores de 80 años, al menos en la Comunidad de Madrid según este estudio, han tenido un porcentaje de muertes del 100% prácticamente, y la mayoría han tenido que ser hospitalizados. Además, el valor del número de casos que ingresan en la UCI no es nulo, como es el caso de los otros dos grupos analizados.

Podría decirse que los hombres mayores se han visto más afectados por el virus en Madrid, según los datos utilizados para este estudio.

Análisis global de las bolas poco coincidentes dentro de los clústeres (7, 13, 14 y 17)

Seguidamente, pasamos a la observación de las bolas 7, 13, 14 y 17 que, aunque estén incluidas en los clústeres, siempre presentan un color diferenciable con respecto al resto de bolas correspondientes a cada clúster. Pasamos, por tanto, a estudiar qué grupos engloba cada una de estas bolas. Por su parte, la bola 7 engloba a los grupos referidos a hombres entre 60-79 años en la comunidad de Castilla y León, hombres entre 60-79 años en Castilla La Mancha y hombres entre 40-59 años en Cataluña. Por otro lado, la bola 13 engloba a los hombres entre 40-69 años de Madrid y la bola 14 a los hombres 60-79 años también en Madrid. Finalmente, la bola 17 agrupa a las mujeres entre 40-79 años de Madrid y las mujeres mayores de 70 años en Cataluña.

Si comenzamos la comparación por la bola 17, que es la única de estas 4 que se encuentra sola en un clúster, podemos observar que tiene un valor más elevado que las demás bolas pertenecientes a dicho clúster en las ilustraciones correspondientes al número de casos confirmados (***Ilustración 28***), al número de casos hospitalizados (***Ilustración 29***) y al número de casos en la UCI (***Ilustración 30***). En las dos primeras representaciones, dicha bola toma un valor comprendido en el rango entre [0.4, 0.6] y en la tercera entre [0.2, 0.4], mientras que las demás bolas que comprenden el clúster cuentan con valores comprendidos entre el intervalo [0, 0.2] en las tres imágenes mencionadas. Sin embargo, en la última representación del algoritmo (***Ilustración 31***) podemos ver que el valor de dicha bola se asemeja a los de las demás,

comprendiéndose entre el rango [0, 0.2], por lo que podemos entender que la relación que tiene nuestro grupo de estudio con los demás del clúster podría venir por la baja tasa de fallecimientos.

Por tanto, podemos asegurar que, con respecto a nuestro grupo completo de individuos analizados, las mujeres madrileñas entre 40-79 años y las mujeres catalanas mayores de 70 años tienen una tasa de infección del virus esperada para dichos grupos de edad, pero no cuentan con una alta tasa de fallecimientos. A diferencia de grupos como los que estudiamos en el apartado anterior, *Análisis de las bolas aisladas (8, 15 y 18),* en concreto los grupos englobados en las bolas 8 y 18 (Mujeres +80 de Cataluña y Madrid), haciendo referencia a la ***Ilustración 29****,* podemos ver que los grupos estudiados anteriormente cuentan con unos valores bajos (entre [0, 0.2]), mientras que la bola 17 en esta cuenta con un valor más alto, comprendido entre [0.2, 0.4], pero en la ***Ilustración 31*** pasa al contrario, el valor de nuestra bola de estudio es muy bajo respecto a las otras dos mencionadas.

Pasando a la evaluación de las otras tres bolas que quedan, la 7, 13 y 14, debemos decir, inicialmente, que se espera un comportamiento similar para la 13 y la 14, ya que se produce solapamiento del grupo de hombres de Madrid entre 60-69 años en estas.

Si profundizamos más, podemos ver que pasa algo parecido al caso anterior. Las tres bolas cuentan con valores muy diferenciados de los demás grupos que forman el clúster en las tres primeras ilustraciones, pero, sin embargo, en la correspondiente al número de fallecidos, sobre todo las bolas 7 y 13, cuentan con valores similares, y, aunque la bola 14 no cuente con valores similares, podemos intuir que su correlación viene inducida por el solapamiento mencionado entre la 13 y la 14.

Conclusiones generales del estudio por comunidades

Una vez realizado un estudio global, con la gran cantidad de datos con la que contamos ahora, salta a la vista que **no podremos sacar análisis muy concretos** y, por tanto, se decide **reducir dicho conjunto** a una comunidad en concreto para cada caso posterior.

Simplemente por el hecho de ser la capital de España y, además**, una de las zonas más afectadas por el virus durante la pandemia**, hemos decidido analizar la **Comunidad de Madrid**, en concreto, los meses más decisivos en la pandemia ocasionada por el virus.

Comunidad de Madrid. Pre-vacunación. Caso 1: (ε = 0.6, marzo - abril 2020)

Para este caso, se realiza una filtración sobre el conjunto de datos del caso anterior, de manera que solo quedan las filas referidas a la Comunidad de Madrid entre las fechas 01/03/2020 – 30/04/2020.

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	15	1	1	124	77	20	1
2	15	1	2	149	53	9	0
3	15	1	3	1313	350	23	4
4	15	1	4	2577	1011	67	12
5	15	1	5	4362	2364	223	67
6	15	1	6	5575	3636	468	219
7	15	1	7	5435	4119	658	560
8	15	1	8	5602	4791	495	1434
9	15	1	9	5486	4505	67	2411
10	15	2	1	93	55	14	0
11	15	2	2	187	61	7	2
12	15	2	3	2632	477	17	5
13	15	2	4	4133	1047	40	9
14	15	2	5	5690	1765	92	44
15	15	2	6	6732	2720	172	84
16	15	2	7	4783	2818	290	218
17	15	2	8	4154	3273	209	631
18	15	2	9	7200	4709	59	1991

Ilustración 32. Conjunto de datos para el estudio de la comunidad de Madrid. Pre-vacunación

	confirmados	*hospitalizados*	*uci*	*fallecidos*
Media	3679.278	2101.722	162.778	427.333
Mediana	4258	2064.5	67	55.500
Moda	124	77	67	0
Rango	7107	4738	651	2411

Tabla 13. Estadísticas descriptivas para el conjunto de datos del estudio de Madrid. Pre-vacunación

Normalizando nuestro conjunto de datos final, quedaría de la siguiente manera:

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	0.5	0	0.000	0.004361897	0.005065428	0.019969278	0.0004147657
2	0.5	0	0.125	0.007879555	0.000000000	0.003072197	0.0000000000
3	0.5	0	0.250	0.171661742	0.062051499	0.024577573	0.0016590626
4	0.5	0	0.375	0.349514563	0.201983959	0.092165899	0.0049771879
5	0.5	0	0.500	0.600675390	0.487758548	0.331797235	0.0277892990
6	0.5	0	0.625	0.771352188	0.756226256	0.708141321	0.0908336790
7	0.5	0	0.750	0.751653300	0.858168003	1.000000000	0.2322687681
8	0.5	0	0.875	0.775151259	1.000000000	0.749615975	0.5947739527
9	0.5	0	1.000	0.758829323	0.939003799	0.092165899	1.0000000000
10	0.5	1	0.000	0.000000000	0.000422119	0.010752688	0.0000000000
11	0.5	1	0.125	0.013226397	0.001477417	0.000000000	0.0008295313
12	0.5	1	0.250	0.357253412	0.089278176	0.015360983	0.0020738283
13	0.5	1	0.375	0.568453637	0.209582102	0.050691244	0.0037328909
14	0.5	1	0.500	0.787533418	0.361333896	0.130568356	0.0182496889
15	0.5	1	0.625	0.934149430	0.562684677	0.253456221	0.0348403152
16	0.5	1	0.750	0.659912762	0.583579569	0.434715822	0.0904189133
17	0.5	1	0.875	0.571408471	0.679400591	0.310291859	0.2617171298
18	0.5	1	1.000	1.000000000	0.982482060	0.079877112	0.8257984239

Ilustración 33. Conjunto de datos para el estudio por individuos en la comunidad de madrid. Pre-vacunación. Caso 1

Finalmente, aplicando nuestro algoritmo con ε = 0.6, queda el reparto de grupos englobados de la siguiente manera:

Nº bola	*Grupo(s) cubierto(s)*
1	Hombres de 0-39 años
2	Hombres de 30-59 años
3	Hombres de 50-79 años
4	Hombres mayores de 80 años
5	Mujeres de 0-29 años
6	Mujeres de 20-49 años
7	Mujeres de 40-79 años
8	Mujeres mayores de 80 años

Tabla 14. Recubrimiento para el problema por CCAAs. Comunidad de Madrid. Pre-vacunación. Caso 1

Aplicando BM de manera similar a los estudios realizados previamente, en las cuatro columnas más relevantes para este caso, las cuatro últimas, obtenemos diferentes representaciones:

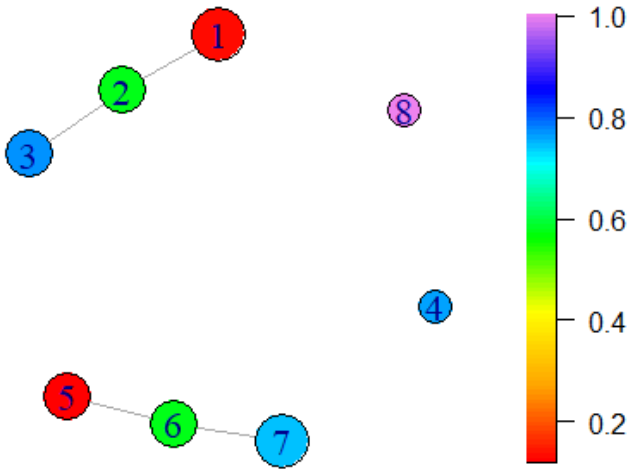


Ilustración 34. Columna referida al número de casos confirmados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 1

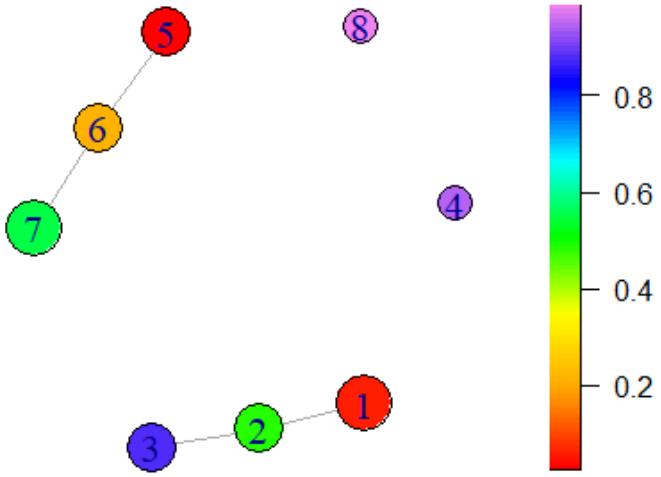


Ilustración 35. Columna referida al número de casos hospitalizados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 1

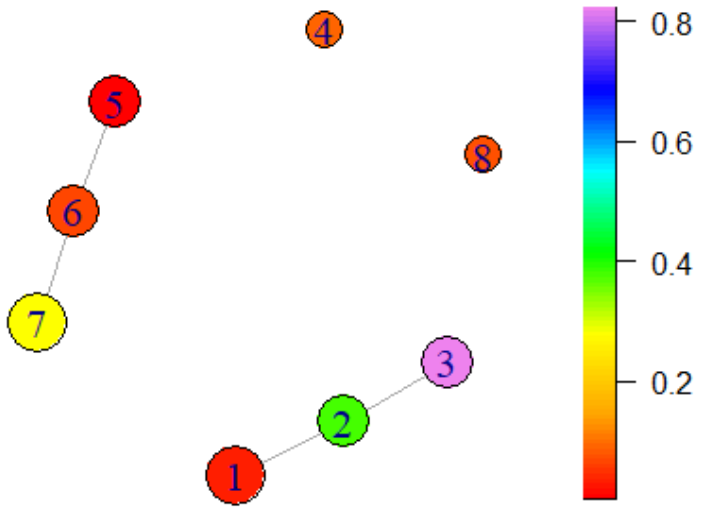


Ilustración 36. Columna referida al número de casos ingresados en la uci. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 1

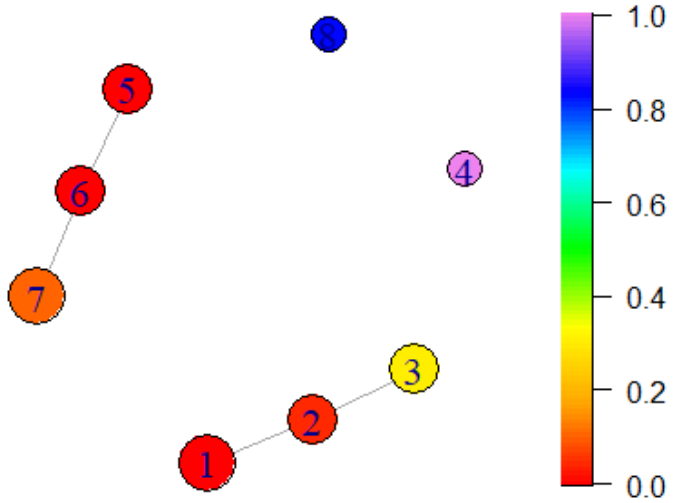


Ilustración 37. Columna referida al número de fallecimientos. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 1

Una vez descritos los clústeres que se forman para la aplicación de nuestro algoritmo sobre los datos respectivos a la población madrileña entre los meses de marzo y abril de 2020, pasamos ahora a analizar y estudiar los resultados obtenidos, en busca de nuevas conclusiones.

Como podemos observar a simple vista, las bolas formadas esta vez, se relacionan de manera similar diferenciando entre sexo, es decir, quedan unidas las bolas correspondientes a mujeres y, por otro lado, las correspondientes a hombres, a

excepción de las que engloban a personas mayores de 80 años, independientemente del sexo, que se representan individualmente en nuestras gráficas.

Inicialmente, la primera representación con la que nos encontramos se trataría de la correspondiente al número de confirmados, ***Ilustración 34***. Esta representación muestra los datos de manera muy visual, ya que podemos advertir que los grupos en los que se engloban individuos de rangos de edad similares cuentan con un color representativo parecido o igual, a excepción de los grupos de personas mayores. Por tanto, vemos, como sabíamos, que conforme la edad de los individuos es más avanzada, el número de casos confirmados es mayor. Sin embargo, para los grupos de personas mayores de 80 años, bolas 4 y 8, existe una gran diferencia de valor dependiendo de si se trata del grupo referido a hombres o a mujeres. Para la bola 8, mujeres mayores, este valor, aproximadamente el máximo en la escala es mucho más alto que para la bola 4, hombres mayores, el cual queda comprendido entre el rango de valores de [0.6, 0.8].

Advertimos, pues, que el número de casos confirmados durante estos meses contaba con un porcentaje más alto de mujeres que de hombres, algo de lo que se había hablado durante el ***Estudio de la seroprevalencia***, realizado con un conjunto de individuos aleatorio.

Si pasamos ahora a la gráfica referida al número de casos hospitalizados, ***Ilustración 35***, advertimos que, de nuevo, conforme la edad se incrementa, también lo hace esta variable. Aunque, esta vez, se producen más anomalías entre los grupos de edades medias diferenciando por sexo.

Si observamos, los grupos de mínima edad, bolas 1 y 5, y los grupos de edad avanzada, bolas 4 y 8, cuentan con valores muy parecidos, siendo muy bajos para los primeros y muy altos para los segundos. Sin embargo, esta vez, si comparamos, por ejemplo, las bolas 2 y 6, que son las que se corresponden a los grupos de hombres entre 30-59 años y mujeres entre 20-49 años respectivamente, ya no coinciden, tomando la bola 2 un valor más alto (en el intervalo [0.4, 0.6]) que la bola 6 (intervalo [0.1, 0.3]). Además, la bola 2 sí coincide en un color similar con la bola 7, correspondiente a las mujeres entre 40-79 años, que cuenta también con un valor entre [0.4, 0.6]. En cuanto al último grupo que queda, representado por la bola 3, hombres entre 50-79 años, cuenta con el valor más elevado de todas las bolas englobadas dentro de algún clúster (comprendido en el intervalo [0.8, 1]).

Comparando ahora los grupos según los rangos de edad englobados, la bola 3 se diferencia de la bola 7 solamente en que esta última engloba una década más, las mujeres entre 40-49 años. Simplemente por eso, no debería haber tanta diferencia entre los valores tomados por ambas bolas, lo que lleva a pensar que este contraste viene también ligado al sexo de los individuos.

Si seguimos observando la siguiente imagen, ***Ilustración 36***, representativa de los casos confirmados que tuvieron que ingresar en la UCI, todos los valores son prácticamente nulos, o muy pequeños. En el caso de las mujeres entre 40-79 años, bola 7, se puede ver que toma un valor ciertamente más grande, entre [0.2, 0.4]. Pero, para los grupos de hombres, estos valores se disparan. Tanto para la bola 2 como para la bola 3, se toman valores muy altos en comparación con los de las demás. Si volvemos al estudio de la gráfica anterior, el número de casos hospitalizados, vemos que las bolas 6 y 7 mantienen cierta relación entre los valores tomados en ambas representaciones, disminuyendo un poco sus valores en la ***Ilustración 36***. Sin embargo, las bolas 2 y 3, no solo mantienen los valores en esta imagen, sino que, además, la bola 3 cuenta con el valor máximo en ella, es decir, prácticamente todos los casos de hombres hospitalizados de la edad representada van a la UCI.

Resumidamente, y haciendo referencia a los estudios realizados previamente, como vimos en los diferentes casos del *Estudio de la seroprevalencia*, el virus afecta de manera más directa a los pocos hombres que se contagien, en comparación de las mujeres. Aunque, debemos seguir teniendo en cuenta que, aunque haya anomalías entre grupos según el sexo, el virus afecta a ambos grupos y se incrementa según lo va haciendo la edad.

Finalmente, se confirma todo lo descrito hace un momento, ya que, si nos fijamos en la última gráfica, ***Ilustración 37***, referida al número de fallecimientos a causa del virus, vemos como en los grupos de mujeres es prácticamente nulo, exceptuando a las mayores de 80 años, y para los hombres toma valores por encima del valor 0.2 a partir de la bola 3 (hombres entre 50-79 años). Además, en cuanto a los grupos de personas mayores, aunque ambos valores son altos, la bola 4 (hombres) cuenta con el máximo, mientras que en la bola 8 (mujeres) se comprende entre [0.7, 0.9].

Comunidad de Madrid. Pre-vacunación. Caso 2: (ε = 0.5, marzo - abril 2020)

Con el objetivo de obtener información más detallada de cómo afecta el virus a los distintos grupos de edad y sacar conclusiones aún no encontradas, se reduce el valor de ε a 0.5 y se aplica de nuevo el algoritmo al conjunto de datos utilizado para el caso

anterior, ***Ilustración 33***, quedando las agrupaciones generadas por BM de la siguiente manera:

Nº bola	*Grupo(s) cubierto(s)*
1	Hombres de 0-29 años
2	Hombres de 10-49 años
3	Hombres de 40-69 años
4	Hombres de 60-79 años
5	Hombres mayores de 80 años
6	Mujeres de 0-29 años
7	Mujeres de 20-49 años
8	Mujeres de 40-79 años
9	Mujeres mayores de 80 años

Tabla 15. Recubrimiento para el problema por CCAAs. Comunidad de Madrid. Pre-vacunación. Caso 2

Igual que venimos haciendo desde la primera aplicación, se aplica el algoritmo a las columnas relevantes y se generan, de nuevo, cuatro gráficas:

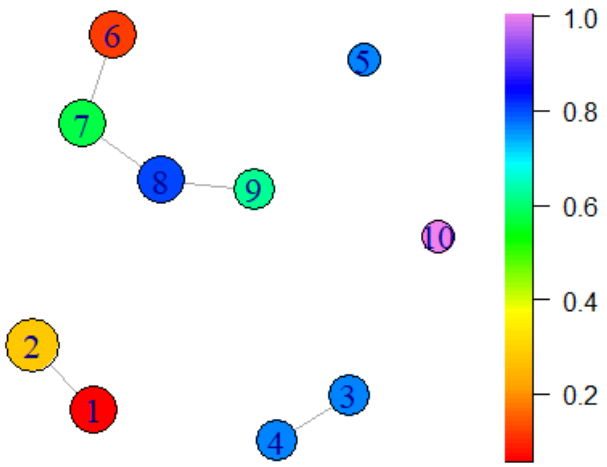


Ilustración 38. Columna referida al número de casos confirmados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 2

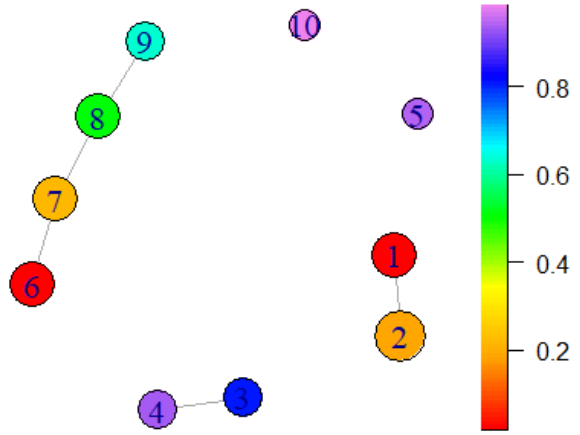


Ilustración 39. Columna referida al número de casos hospitalizados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 2

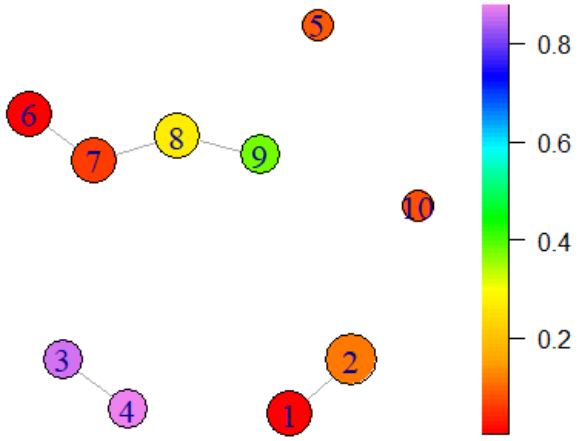


Ilustración 40. Columna referida al número de casos ingresados en la UCI. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 2

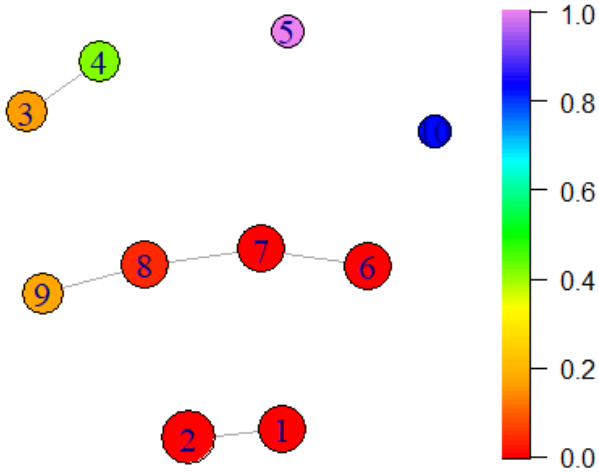


Ilustración 41. Columna referida al número de fallecimientos. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 2

Atendiendo al reparto de los grupos en las distintas bolas que se forman tras aplicar el algoritmo BM, podemos ver cómo, aun reduciendo el valor de ε para desgranar las relaciones un poco, los grupos de mujeres no han variado nada. Además, uno de los clústeres formados cuenta con todas las agrupaciones de mujeres entre 0-79 años, por lo que podemos advertir que el paralelismo entre los individuos de sexo femenino respecto a los efectos que el virus produce sobre ellos es muy alto, como pasaba ya, si recordamos, en casos anteriores.

Por otro lado, las agrupaciones de individuos masculinos sí se han separado en dos clústeres, uno formado por las bolas 1 y 2, hombres entre 0-49 años, y otro formado por las bolas 3 y 4, hombres entre 40-69 años. Las bolas correspondientes a las personas mayores de 80 años de ambos sexos, bolas 5 y 10, siguen estando aisladas del resto.

Siguiendo el caso anterior, podemos ver que, en cuanto al sexo masculino, los hombres de menor edad no se contagian tanto, como ya sabíamos, pero las bolas de los hombres a partir de 40 años tienen una tasa de contagio similar, siendo menor que en las de las mujeres. Sin embargo, podemos seguir apreciando que, tanto la bola 8 y 9, correspondientes a las mujeres mayores de 40 años, aunque toman valores más elevados para la medida de *Casos confirmados*, en la gráfica correspondiente al número de casos hospitalizados, ***Ilustración 39***, es menor que para las bolas representantes de esta edad en el sexo masculino. Pasa lo mismo realizando la comparación ahora en la representación del número de casos en la UCI, ***Ilustración 40***, en la cual se incrementa aún más la diferencia de valores entre los grupos masculinos y femeninos.

Además, si atendemos ahora a los grupos mayores de 80 años, bolas 5 y 10, para las mujeres tiene un valor muy bajo, pero destaca el hecho de que, para los hombres de esta edad, también. Siendo el virus más peligroso para las personas mayores, como hemos podido demostrar con los estudios anteriores, en esta gráfica podemos ver cómo los casos graves de contagios en hombres se dan en el rango de edades de entre 40-79 años.

En la última imagen, referida al número de muertes, confirmamos que el mayor número de muertes se produce en los grupos de edad avanzada, aunque más en los hombres que en las mujeres. Sin embargo, en cuanto a los clústeres, siguen tomando valores más elevados el clúster formado por las bolas 3 y 4, cuando los demás toman valores bajos o prácticamente nulos.

Comunidad de Madrid. Pre-vacunación. Caso 3: (ε = 0.35, marzo - abril 2020, reducción del conjunto de datos)

Una vez estudiado el alto nivel de vecindad entre los grupos femeninos, se decide aplicar nuestro algoritmo solamente a éstos. De manera que se reduce el conjunto de datos y se elige el valor de ε más adecuado, en concreto 0.35, queriendo evitar solapamiento y a la vez contar con algún clúster formado por más de una bola.

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	15	2	1	93	55	14	0
2	15	2	2	187	61	7	2
3	15	2	3	2632	477	17	5
4	15	2	4	4133	1047	40	9
5	15	2	5	5690	1765	92	44
6	15	2	6	6732	2720	172	84
7	15	2	7	4783	2818	290	218
8	15	2	8	4154	3273	209	631
9	15	2	9	7200	4709	59	1991

Ilustración 42. Conjunto de datos para el estudio de la Comunidad de Madrid. Pre-vacunación. Conjunto de mujeres

	confirmados	*hospitalizados*	*uci*	*fallecidos*
Media	3956	1880.556	100	331.556
Mediana	4154	1765	59	44
Moda	93	55	14	0
Rango	7107	4654	283	1991

Tabla 16. Estadísticas descriptivas para el conjunto de datos de la Comunidad de Madrid. Pre-vacunación. Conjunto de mujeres

Haciendo referencia a la ***Tabla 6***, podemos observar que los valores referentes a la columna de *confirmados* no varían en gran medida de los valores estadísticos para este caso. Esto puede indicarnos en una primera impresión de los datos, que la mayor parte de la información correspondiente a esta columna nos la aporta el grupo de individuos femeninos. Algo parecido pasa con la columna de hospitalizados, aunque en menor medida.

Sin embargo, para las columnas de *uci* y de *fallecidos*, pasa lo contrario. Esto puede respaldar las conclusiones extraídas hasta el momento referidas a la prevalencia del virus en mujeres y la gravedad con la que afecta al grupo masculino.

El conjunto de datos final para el que se aplica BM en este caso queda de manera:

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	0.5	0.5	0.000	0.0000000	0.000000000	0.02473498	0.000000000
2	0.5	0.5	0.125	0.0132264	0.001074576	0.00000000	0.001004520
3	0.5	0.5	0.250	0.3572534	0.090479261	0.03533569	0.002511301
4	0.5	0.5	0.375	0.5684536	0.212980873	0.11660777	0.004520342
5	0.5	0.5	0.500	0.7875334	0.367504836	0.30035336	0.022099448
6	0.5	0.5	0.625	0.9341494	0.572533849	0.58303887	0.042189854
7	0.5	0.5	0.750	0.6599128	0.593810445	1.00000000	0.109492717
8	0.5	0.5	0.875	0.5714085	0.691381904	0.71378092	0.316926168
9	0.5	0.5	1.000	1.0000000	1.000000000	0.18374558	1.000000000

Ilustración 43. Conjunto de datos para el estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 3

Se aplica de manera similar BM surgiendo las cuatro gráficas representativas de las columnas relevantes para el estudio y un recubrimiento como se describe a continuación:

Nº bola	***Grupo(s) cubierto(s)***
1	Mujeres de 0-19 años
2	Mujeres de 20-39 años
3	Mujeres de 30-49 años
4	Mujeres de 50-59 años
5	Mujeres de 60-69 años
6	Mujeres de 70-79 años
7	Mujeres mayores de 80 años

Tabla 17. Recubrimiento para el problema por CCAAs. Comunidad de Madrid. Pre-vacunación. Caso 3

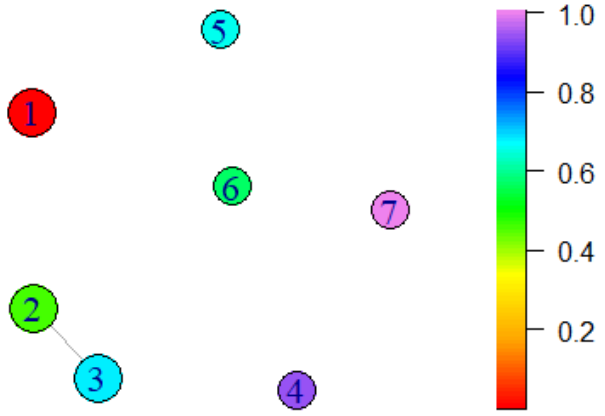


Ilustración 44. Columna referida al número de casos confirmados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 3

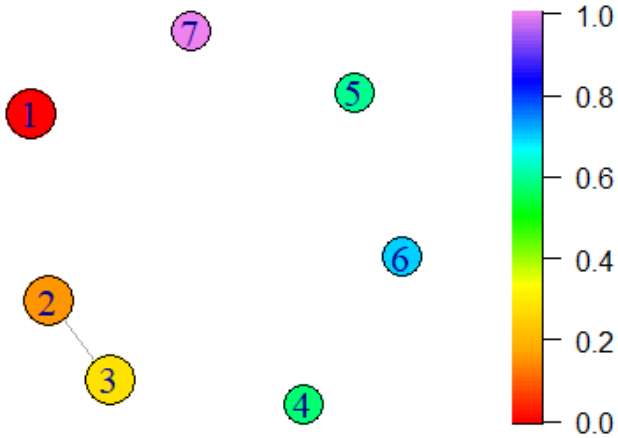


Ilustración 45. Columna referida al número de casos hospitalizados. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 3

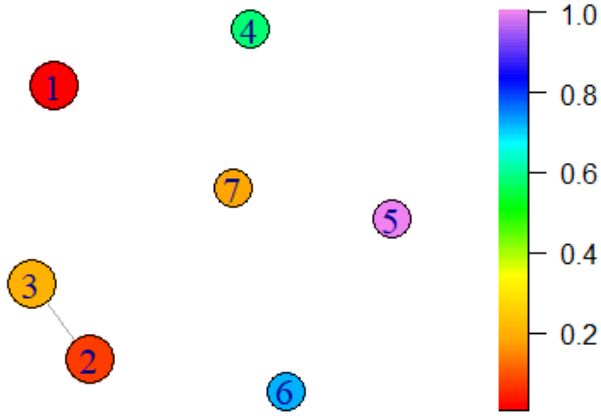


Ilustración 46. Columna referida al número de casos ingresados en la UCI. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 3

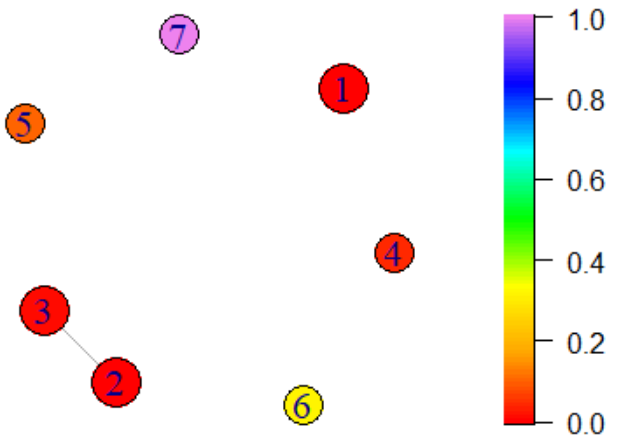


Ilustración 47. Columna referida al número de fallecimientos. Estudio por individuos en la Comunidad de Madrid. Pre-vacunación. Caso 3

Inicialmente, siguiendo el orden tomado para todos los casos, comenzamos a evaluar los resultados de la primera visualización, referente al número de casos confirmados.

Ahora que nuestro estudio se centra solo en un grupo de individuos reducido y, por tanto, casi todas las bolas engloban una única fila de nuestro conjunto de datos, podemos advertir cosas como que, además de las mujeres mayores de 80 años, las mujeres entre 50-59 años (bola 4) también cuentan con una tasa de contagio más alta que otras bolas como la 5 o la 6, que corresponden a los grupos de edades entre 60-69 años y 70-79 años, respectivamente.

Algo parecido sucede con la bola 5, mujeres entre 60-69 años, que cuenta con un valor comprendido en el intervalo [0.6, 0.8], mientras que la bola 6, mujeres entre 70-79 años, cuenta con un valor ciertamente más pequeño, entre [0.5, 0.6].

En cuanto al número de casos hospitalizados, ***Ilustración 45***, si nos fijamos, las bolas sí siguen un orden ascendente por la escala de colores, independientemente del porcentaje de casos confirmados que cubra cada una, de manera que los grupos correspondientes a las mujeres más jóvenes cuentan con valores bajos en dicha escala y éstos van creciendo hasta la bola 7, representativa de las mujeres mayores de 80 años, que cuenta con el máximo.

Pasamos ahora a analizar la ***Ilustración 46***, correspondiente al número de casos hospitalizados en la UCI. Aquí, si podemos encontrar ciertas anomalías respecto a las dos representaciones anteriores. En este caso, el máximo valor lo toma el grupo cubierto por la bola 5, seguido del cubierto por la bola 6, que toma un valor comprendido en el rango [0.6, 0.8].

Si recordamos, en cuanto al número de casos confirmados, la bola 4 contaba con una tasa de confirmación más alta que la bola 5 y la bola 3 con un valor similar. Por otra parte, en cuanto al número de casos hospitalizados, la bola 4 contaba con el mismo valor que la 5 y la 6 con uno superior. Sin embargo, aunque el número de casos confirmados de COVID o el número de casos hospitalizados por este virus sea mayor en mujeres de otras edades, podríamos afirmar que en las mujeres entre los 60-69 años (bola 5) tiene unos efectos considerablemente negativos y puede llegar a pasar más factura, puesto que tienen mucha posibilidad de entrar en la UCI.

Finalmente, fijándonos en la última gráfica podemos ver que, respecto al último grupo mencionado, aunque la agresividad del virus sea alta, la vida de estos individuos no corre tanto peligro como en los de mayor edad. Como cabía esperar, en esta gráfica destacan los grupos de edad avanzada, bolas 7 y 6, pero para el resto el valor tomado es mínimo o prácticamente nulo.

Conclusiones generales del estudio en la Comunidad de Madrid. Pre-vacunación

- Conforme la **edad de los individuos es más avanzada**, el número de **casos confirmados y hospitalizados aumenta**.
- El **número de casos confirmados** suele ser **más grande** en las bolas representativas de **grupos femeninos** que en las de grupos masculinos.
- El **número de hospitalizados en la UCI y de fallecimientos** suele ser **mayor** para los **individuos masculinos** que para los femeninos.
- Los **grupos femeninos** están **muy relacionados** en cuanto a la repercusión del virus independientemente de la edad.
- El **número de contagios** en las **mujeres** de entre **40-49 años** es **más elevado** que en las **mujeres entre 60-79 años**, pero el valor de los casos hospitalizados y en la UCI es superado por estas últimas.

Comunidad de Madrid. Pos-vacunación. Caso 1: ($\varepsilon = 0.6$, marzo – abril 2021)

Pasamos al estudio pos-vacunación seleccionando las filas correspondientes a los mismos meses estudiados en el caso anterior, pero un año después.

Para este periodo de tiempo, ya había comenzado el proceso de vacunación en la Comunidad de Madrid, además de en el resto de España.

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	15	1	1	3500	40	3	0
2	15	1	2	5925	74	8	0
3	15	1	3	7503	181	18	0
4	15	1	4	7246	340	21	3
5	15	1	5	8791	732	69	5
6	15	1	6	7251	1014	126	35
7	15	1	7	3933	969	194	115
8	15	1	8	2366	889	131	164
9	15	1	9	1104	584	18	270
10	15	2	1	3371	48	7	0
11	15	2	2	5750	69	4	0
12	15	2	3	7917	221	14	4
13	15	2	4	7578	329	25	2
14	15	2	5	9205	494	37	4
15	15	2	6	7446	690	57	17
16	15	2	7	4214	716	98	35
17	15	2	8	2734	758	87	91
18	15	2	9	1707	748	12	219

Ilustración 48. Conjunto de datos para el estudio de la Comunidad de Madrid. Pos-vacunación (mar-abr)

	confirmados	*hospitalizados*	*uci*	*fallecidos*
Media	5418.944	494.222	51.611	53.556
Mediana	5837.500	539	23	4.5
Moda	3500	40	18	0
Rango	8101	974	191	270

Tabla 18. Estadísticas descriptivas para el conjunto de datos de la Comunidad de Madrid. Pos-vacunación (mar-abr)

Haciendo referencia a la ***Tabla 6***, podemos ver que las medidas descriptivas para las columnas *hospitalizados*, *uci* y *fallecidos* han reducido considerablemente sus valores. Esto nos puede indicar previamente que los efectos esperados por la vacuna serán alcanzables y que, aunque los casos confirmados se mantengan o aumenten, como parece pasar entre estos estudios, las consecuencias graves provocadas por el virus se han mitigado.

En este caso, aplicamos el mismo valor de ε que se utiliza en el apartado ***Comunidad de Madrid. Pre-vacunacion. Caso 1 (ε = 0.6, marzo – abril 2020)***:

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	0.5	0	0.000	0.29406643	0.000000000	0.000000000	0.000000000
2	0.5	0	0.125	0.59410374	0.036649215	0.027472527	0.003921569
3	0.5	0	0.250	0.78840652	0.146596859	0.076923077	0.000000000
4	0.5	0	0.375	0.75805448	0.307853403	0.093406593	0.011764706
5	0.5	0	0.500	0.94800348	0.705759162	0.357142857	0.019607843
6	0.5	0	0.625	0.75668615	1.000000000	0.653846154	0.133333333
7	0.5	0	0.750	0.34892400	0.943455497	1.000000000	0.419607843
8	0.5	0	0.875	0.15648713	0.862827225	0.648351648	0.600000000
9	0.5	0	1.000	0.00000000	0.558115183	0.082417582	1.000000000
10	0.5	1	0.000	0.27876602	0.008376963	0.021978022	0.000000000
11	0.5	1	0.125	0.57121533	0.030366492	0.005494505	0.003921569
12	0.5	1	0.250	0.83716880	0.188481675	0.060439560	0.015686275
13	0.5	1	0.375	0.79947755	0.301570681	0.120879121	0.007843137
14	0.5	1	0.500	1.00000000	0.464921466	0.181318681	0.015686275
15	0.5	1	0.625	0.78293320	0.661780105	0.285714286	0.058823529
16	0.5	1	0.750	0.38338102	0.680628272	0.521978022	0.129411765
17	0.5	1	0.875	0.20039806	0.727748691	0.450549451	0.329411765
18	0.5	1	1.000	0.07500933	0.736125654	0.049450549	0.827450980

Ilustración 49. Conjunto de datos para el estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 1

Nº bola	*Grupo(s) cubierto(s)*
1	Hombres de 0-29 años
2	Hombres de 10-49 años
3	Hombres de 40-59 años
4	Hombres de 60-79 años
5	Hombres mayores de 80 años
6	Mujeres de 0-19 años
7	Mujeres de 10-49 años
8	Mujeres de 30-69 años
9	Mujeres de 60-79 años
10	Mujeres mayores de 80 años

Tabla 19. Recubrimiento para el problema por CCAAs. Comunidad de Madrid. Pos-vacunación. Caso 1

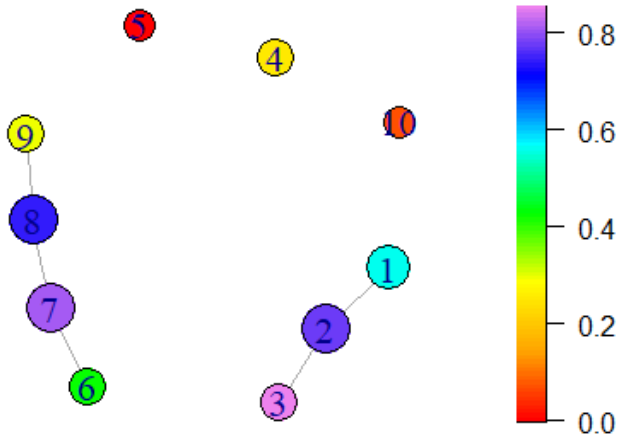


Ilustración 50. Columna referida al número de casos confirmados. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 1

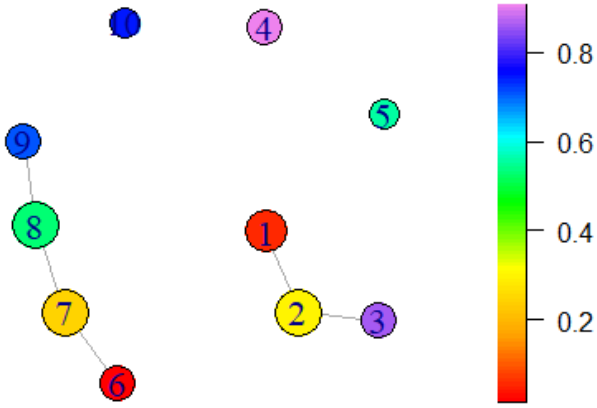


Ilustración 51. Columna referida al número de casos hospitalizados. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 1

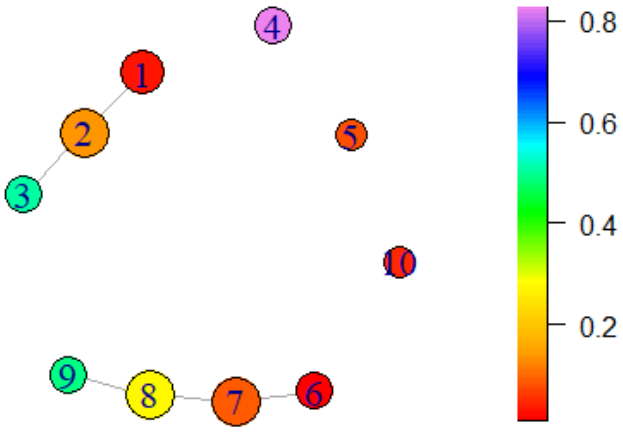


Ilustración 52. Columna referida al número de casos ingresados en la uci. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 1

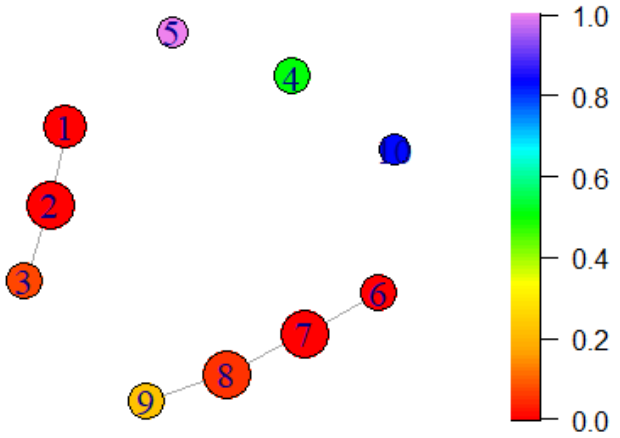


Ilustración 53. Columna referida al número de fallecimientos. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 1

Debido al objetivo de este segundo estudio en la Comunidad de Madrid, el de visualizar los efectos positivos de la vacunación contra el COVID-19, nuestra atención se centrará, sobre todo, en los grupos de mayor edad, ya que en el rango de fechas seleccionadas la vacunación se centraba únicamente en las personas de riesgo.

Como podemos observar inicialmente, en la ***Ilustración 50***, ya observamos ciertas diferencias con respecto a los casos anteriores. Como podemos ver, centrándonos primero en los grupos mencionados, las bolas 5 y 10, correspondientes a las personas mayores de 80 años, toman ahora los valores más pequeños de la escala. Si recordamos, en el apartado *Comunidad de Madrid. Caso 1 (ε = 0.6, marzo – abril 2020),* las bolas 4 y 8, referentes a esta edad, contaban con los valores más altos. Por tanto, vemos como uno de los objetivos de la vacuna se cumple, evitando el contagio masivo para estos individuos.

Algo parecido pasa con los grupos englobados en las bolas 4 y 9, representativas de las edades entre 60-79 años, que, aunque no tomen un valor mínimo, podemos ver cómo este se ha reducido con respecto al caso anterior.

Como consecuencia de estos cambios, pasan a tomar los valores más altos las bolas 2 y 3, hombres entre 10-59 años, y las bolas 7 y 8, correspondientes a los grupos de mujeres entre 10-69 años. Este cambio, no implica que el contagio entre las personas de estas edades haya aumentado, pero sí que ahora los casos positivos destacan más en ellas que en las personas de edad avanzada.

Seguidamente, haciendo alusión a la ***Ilustración 51***, que indica el número de casos hospitalizados, podemos observar que pasa algo parecido, pero en menor medida.

Las bolas 5 y 10, vuelven a reducir su valor. En el caso anterior con el que estamos comparando este estudio, los grupos de personas mayores contaban para esta columna con los máximos valores, sin embargo, aunque ahora los valores no son de los más bajos, sí que se han reducido en este caso.

Además, podemos observar cómo los máximos valores no incluyen a todos los individuos de edades medias, sino solo a los grupos de individuos masculinos, en concreto las bolas 3 y 4. Cabe destacar el hecho de que, para los hombres entre 60-79 años, bola 4, hemos visto que la tasa de casos confirmados era similar a la de las mujeres, bola 9. Sin embargo, en cuanto a casos hospitalizados, podemos ver que el primer grupo supera en valor al segundo.

Enlazando con las ilustraciones: ***Ilustración 52*** e ***Ilustración 53*****,** casos ingresados en la UCI y fallecimientos, respectivamente, podemos ver que este grupo vuelve a contar con el máximo valor en la escala, lo que acaba de confirmar la conclusión extraída anteriormente de que los hombres son más propensos a sufrir el contagio del virus.

Por último, debemos atender de nuevo a los objetivos de la vacunación del COVID-19. El principal objetivo de la vacunación es prevenir la enfermedad y disminuir su gravedad y mortalidad. Sin embargo, refiriéndonos a la ***Ilustración 53***, podemos ver cómo, aunque el número de casos confirmados en edades avanzadas era bajo, la tasa de fallecimiento en estos individuos vuelve a ocupar el rango de mayores valores.

Sin embargo, no debemos olvidar el hecho de que estamos estudiando los meses iniciales del proceso de vacunación en España, en concreto, para la Comunidad de Madrid, y puede que aún no se hubiera aplicado la pauta completa para todas las personas incluidas en estos grupos. (Dirección General de Salud Pública. Consejería de Sanidad): *Conforme se han ido autorizando nuevas vacunas frente a COVID-19 y ante la situación epidemiológica actual, se han ido ampliando los grupos a vacunar cuya priorización se ha realizado en función de criterios éticos y de la evidencia científica, comenzando con la vacunación de las personas más vulnerables y con mayor riesgo de exposición y de transmisión a otras personas.*

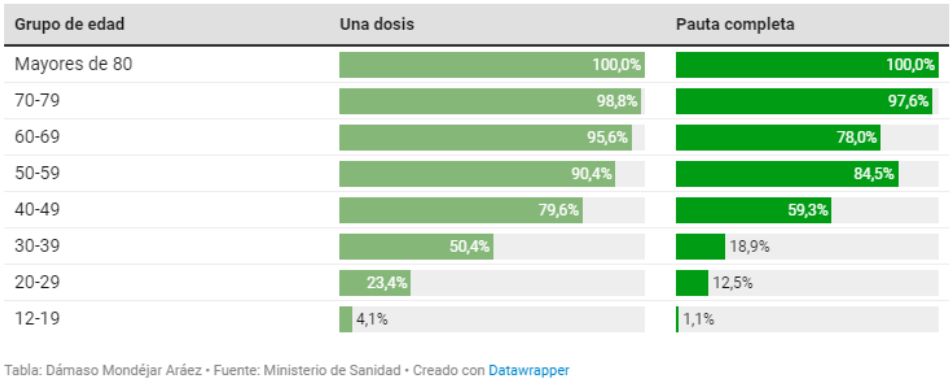
Comunidad de Madrid. Pos-vacunación. Caso 2: (ε = 0.6, junio – julio 2021)

Para sacar alguna conclusión acertada sobre si se alcanza o no el objetivo mencionado, deberemos realizar un análisis adicional de los meses posteriores, en los que el proceso de vacunación ya estará más avanzado.

(Dirección General de Salud Pública. Consejería de Sanidad) *31 de julio de 2021*: *Actualmente se está vacunando a los grupos 10, 11 y 12. De manera simultánea se está vacunando a las personas de los grupos de población que se priorizaron con anterioridad, como son personal sanitario o sociosanitario de nueva incorporación, personas de colectivos con una Dirección General de Salud Pública CONSEJERÍA DE SANIDAD 6 función esencial para la sociedad, personas de muy alto riesgo, colectivos vulnerables desde el punto de vista social, económico y/o laboral...*

Los grupos mencionados en la cita están explicados en el documento que se referencia, en concreto, corresponden a los grupos: Personas entre 40 y 49 años, Personas entre 30 y 39 años, y Personas entre 20 y 29 años, respectivamente.

Para el día 31 de julio, específicamente el porcentaje de vacunación cubierto por edades sería el siguiente:



Grupo de edad	Una dosis	Pauta completa
Mayores de 80	100,0%	100,0%
70-79	98,8%	97,6%
60-69	95,6%	78,0%
50-59	90,4%	84,5%
40-49	79,6%	59,3%
30-39	50,4%	18,9%
20-29	23,4%	12,5%
12-19	4,1%	1,1%

Tabla: Dámaso Mondéjar Aráez • Fuente: Ministerio de Sanidad • Creado con Datawrapper

Ilustración 54. Porcentajes de personas vacunadas por grupos de edad. 31 julio de 2021

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	15	1	1	2395	53	2	0
2	15	1	2	8777	125	2	2
3	15	1	3	14225	314	11	0
4	15	1	4	9081	381	25	0
5	15	1	5	5829	406	36	4
6	15	1	6	2867	237	26	7
7	15	1	7	1793	209	35	24
8	15	1	8	533	91	12	31
9	15	1	9	490	194	1	21
10	15	2	1	2382	33	1	0
11	15	2	2	8880	134	4	0
12	15	2	3	12776	328	9	1
13	15	2	4	7919	338	12	0
14	15	2	5	5900	281	16	2
15	15	2	6	3489	195	11	2
16	15	2	7	1928	145	15	7
17	15	2	8	687	71	5	13
18	15	2	9	847	242	4	20

Ilustración 55. Conjunto de datos para el estudio de la Comunidad de Madrid. Pos-vacunación (jun-jul)

	confirmados	*hospitalizados*	*uci*	*fallecidos*
Media	5044.333	209.833	12.611	7.444
Mediana	3178	202	11	2
Moda	2395	53	2	0
Rango	13735	373	35	31

Tabla 20. Estadísticas descriptivas para el conjunto de datos de la Comunidad de Madrid. Pos-vacunación (jun-jul)

Esta vez, podemos ver cómo los valores descriptivos de las variables se reducen aún más, incluida la columna confirmados, una buena señal.

Por tanto, en este caso seleccionamos los datos correspondientes a los meses: junio y julio de 2021, y aplicamos el mismo valor de ε que se utiliza en el apartado ***Análisis de la Comunidad de Madrid. Pos-vacunación. Caso 1***, quedando los datos para el siguiente estudio de manera:

	iso	sexo	edad	confirmados	hospitalizados	uci	fallecidos
1	0.5	0	0.000	0.138696760	0.0536193	0.02857143	0.00000000
2	0.5	0	0.125	0.603349108	0.2466488	0.02857143	0.06451613
3	0.5	0	0.250	1.000000000	0.7533512	0.28571429	0.00000000
4	0.5	0	0.375	0.625482344	0.9329759	0.68571429	0.00000000
5	0.5	0	0.500	0.388714962	1.0000000	1.00000000	0.12903226
6	0.5	0	0.625	0.173061522	0.5469169	0.71428571	0.22580645
7	0.5	0	0.750	0.094867128	0.4718499	0.97142857	0.77419355
8	0.5	0	0.875	0.003130688	0.1554960	0.31428571	1.00000000
9	0.5	0	1.000	0.000000000	0.4316354	0.00000000	0.67741935
10	0.5	1	0.000	0.137750273	0.0000000	0.00000000	0.00000000
11	0.5	1	0.125	0.610848198	0.2707775	0.08571429	0.00000000
12	0.5	1	0.250	0.894503094	0.7908847	0.22857143	0.03225806
13	0.5	1	0.375	0.540880961	0.8176944	0.31428571	0.00000000
14	0.5	1	0.500	0.393884237	0.6648794	0.42857143	0.06451613
15	0.5	1	0.625	0.218347288	0.4343164	0.28571429	0.06451613
16	0.5	1	0.750	0.104696032	0.3002681	0.40000000	0.22580645
17	0.5	1	0.875	0.014342920	0.1018767	0.11428571	0.41935484
18	0.5	1	1.000	0.025991991	0.5603217	0.08571429	0.64516129

Ilustración 56. Conjunto de datos para el estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 2

Nº bola	*Grupo(s) cubierto(s)*
1	Hombres de 0-19 años
2	Hombres de 20-39 años
3	Hombres de 30-59 años
4	Hombres de 60-69 años
5	Hombres mayores de 70 años
6	Mujeres de 0-19 años
7	Mujeres de 20-39 años
8	Mujeres de 30-69 años
9	Mujeres mayores de 60

Tabla 21. Recubrimiento para el problema por CCAAs. Comunidad de Madrid. Pos-vacunación. Caso 2

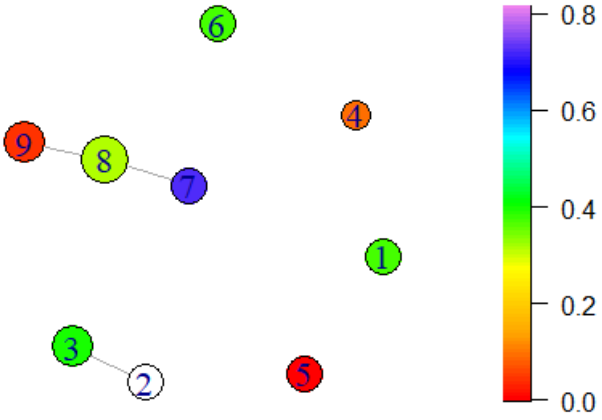


Ilustración 57. Columna referida al número de casos confirmados. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 2

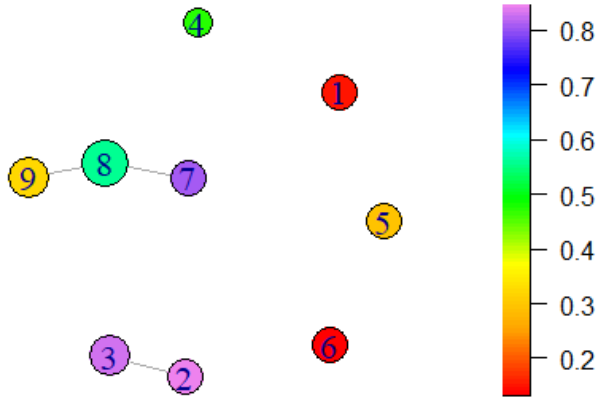


Ilustración 58. Columna referida al número de casos hospitalizados. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 2

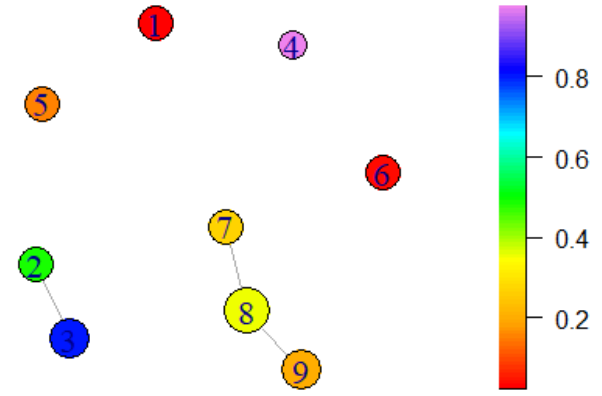


Ilustración 59. Columna referida al número de casos ingresados en la UCI. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 2

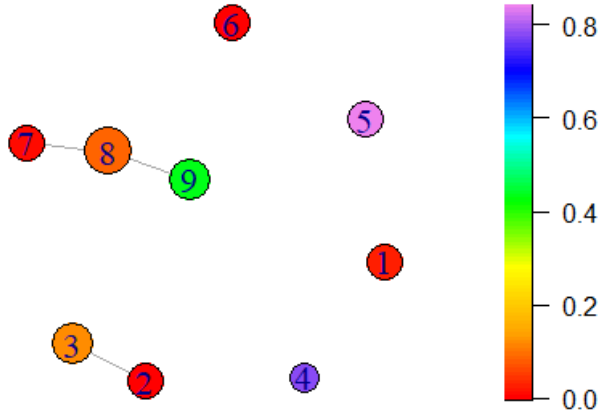


Ilustración 60. Columna referida al número de fallecimientos. Estudio por individuos en la Comunidad de Madrid. Pos-vacunación. Caso 2

Atendiendo a la primera gráfica que surge tras aplicar nuestro algoritmo a este caso, ***Ilustración 57***, advertimos, sobre todo, el hecho de que una de las bolas no cuenta con su correspondiente coloreado. En concreto, se trata de la bola 2, que engloba a los hombres entre 20-39 años.

Además, la bola que agrupa a los individuos comprendidos en estas edades, pero de sexo contrario, es la bola que cuenta con el valor máximo en esta representación, bola 7.

Si nos fijamos, también podemos notar que, en el caso anterior, para la gráfica correspondiente a esta misma columna, ***Ilustración 50***, la degradación de color sucede de forma progresiva, estando repartidos a lo largo de la escala todos los colores utilizados para colorear las bolas. Sin embargo, en este estudio podemos ver que la mayor parte de ellas se representan con un color comprendido en la mitad inferior de la escala de colores. Es decir, la diferencia entre valores ha aumentado entre los grupos que cuentan con un alto porcentaje de vacunación, casi completo, y las que comienzan a incrementar ese porcentaje en este periodo.

En cuanto al número de casos hospitalizados, los valores más altos corresponden a las bolas 2, 3 y 7, Hombres entre 30-39 años, Hombres entre 30-59 años y Mujeres entre 20-39 años, respectivamente. Las bolas 1 y 6 cuentan con los mínimos valores, como en los casos anteriores, ya que estos grupos son los menos afectados por el virus. Lo más interesante a destacar aquí es el hecho de que las bolas 5 y 9, que corresponden a las personas de mayor edad, cuentan con los segundos valores más bajos,

comprendidos en el intervalo [0.25, 0.35], y, si recordamos, en el caso anterior la tasa de personas hospitalizadas era mucho más elevada, tomando los grupos correspondientes valores entre [0.7, 1] en la ***Ilustración 51***.

Además de esto, pasando a la ***Ilustración 59***, gráfica que representa el número de casos ingresados en la UCI, podemos observar fenómenos como el de que los grupos que incluyen casos masculinos cuentan con valores más elevados que los que incluyen casos femeninos, que toman valores todos por debajo de la mitad de la escala.

Por tanto, podría observarse cómo, conforme ha subido el porcentaje de vacunación en los individuos mayores de 60 años, el número de casos hospitalizados e ingresados en la UCI para esta edad se ha reducido considerablemente, haciendo que se confirmen los objetivos que se querían alcanzar con la vacuna.

Finalmente, para el número de fallecimientos observamos que los mayores valores los toman los grupos en las bolas 4, 5 y 9, personas mayores de 60 años. Aunque el número de muertes se haya reducido, los individuos de edades avanzadas siguen siendo las personas con más riesgo ante el virus. Aunque, como podemos observar, se siguen manteniendo algunas de las conclusiones extraídas anteriormente, como la de que los hombres mueren mucho más por el virus. Si nos fijamos, las bolas 4 y 5, correspondientes a hombres, cuentan con valores mucho más altos que la bola 9, mujeres.

Conclusiones generales del estudio en la Comunidad de Madrid. Pos-vacunación

- El número de **casos confirmados disminuye** considerablemente en los **grupos de edad que cuentan con la vacuna** en cada caso.
- La **tasa de fallecimientos respecto el número de casos confirmados** no disminuye tanto como la de confirmados, como se esperaba que pasase tras la vacunación.
- Los valores para los gráficos de **casos hospitalizados descienden** también en los **grupos vacunados o semi-vacunados**.
- Se **mantienen conclusiones anteriores**, como la mayor afección del virus en los hombres.

Capítulo 7. Conclusiones

En este apartado se realiza el resumen de los resultados obtenidos y se comprueba si las hipótesis realizadas son correctas y se han cumplido los objetivos propuestos ante la situación de pandemia vivida durante estos dos últimos años.

Sintetizando todas las conclusiones extraídas en los apartados del capítulo anterior, podríamos resumir las conclusiones pre-vacunación en:

- La **prevalencia en las mujeres** suele ser **mayor** que en los hombres, independientemente de la edad.
- Cuanto **más nos acercamos a edades grandes**, la **tasa de mortalidad** comienza a **incrementarse**.
- Es **más probable contagiarse** siendo del **sexo femenino** que siendo del sexo masculino.
- El **número de hospitalizados en la UCI y de fallecimientos** suele ser **mayor** para los **individuos masculinos** que para los femeninos.
- Los **grupos femeninos** están **muy relacionados** en cuanto a la repercusión del virus independientemente de la edad.

Por tanto, hay resultados que ya esperábamos encontrar, como la agresividad del virus del COVID-19 en personas mayores. Sin embargo, gracias a la aplicación de BM a los distintos conjuntos de datos hemos descubierto matices que se escapaban a simple vista, como la afectividad del virus en los individuos de sexo masculino aun siendo más probable contagiarse siendo del sexo femenino.

En cuanto a los resultados obtenidos para los periodos pos-vacunación quedarían de la siguiente forma:

- El número de **casos confirmados disminuye** considerablemente en los **grupos de edad que cuentan con la vacuna** en cada caso.
- La **tasa de fallecimientos respecto el número de casos confirmados** no disminuye tanto como la de confirmados, como se esperaba que pasase tras la vacunación.
- Los valores para los gráficos de **casos hospitalizados descienden** también en los **grupos vacunados o semi-vacunados**.
- Se **mantienen conclusiones anteriores**, como la mayor afección del virus en los hombres.

Podemos ver que se consiguen los objetivos esperados tras la aplicación de la vacuna, ya que el número de casos confirmados es prácticamente nulo para los grupos vacunados o semi-vacunados en comparación con las estadísticas de 2020.

En conclusión, a la vista de los resultados obtenidos en ambas fases de la investigación, se puede decir con certeza que existen anomalías que no habían sido expresadas anteriormente en distintos estudios realizados. Además, las diferencias tras la aplicación de la vacuna son notables.

En cuanto a las tareas a desarrollar en el futuro, en caso de querer continuar con el estudio en cuestión, podrían centrarse en tratar de aplicar nuestro algoritmo a conjuntos de datos que recojan más características de la población estudiada, que aporten información adicional además de diferenciar en edades y sexo. El proyecto en sí ha contado con un conjunto de datos suficiente para obtener nuevos resultados, sin embargo, sería ideal ampliar dicho conjunto para afianzar las conclusiones obtenidas y descubrir algunas más que no ha sido posible destapar con este.

Además de todo esto, el proyecto y sus conclusiones han conllevado una satisfacción personal enorme, viendo cómo, poco a poco, con paciencia y sin bajar la guardia, entre todos podemos conseguir que todo vuelva a la normalidad y superar de una vez esta extraña situación a la que nos hemos enfrentado en los últimos meses.

Referencias bibliográficas

Chen, Y., & Volic, I. (13 de Agosto de 2021). *Topological data analysis model for the spread of the Coronavirus.* PLOS ONE 16(8): e0255584. Obtenido de https://doi.org/10.1371/journal.pone.0255584

Curran, J. (2011). 1.5 Why R? En J. Curran, *Introduction to Data Analysis with R for Forensic Scientists* (págs. 4-6). Nueva Zelanda: (1st ed.). CRC Press. Obtenido de https://doi.org/10.1201/9781420088274

Dirección General de Salud Pública. Consejería de Sanidad. (s.f.). *Documento informativo de vacunación frente a COVID-19 en la Comunidad de Madrid.* Obtenido de Portal informativo de la Comunidad de Madrid: https://www.comunidad.madrid/sites/default/files/doc/sanidad/prev/doc_tecnico_vacunacion_covid-19.pdf

DLOTKO, P. (2019). *Ball mapper: a shape summary for topological data analysis.* Swansea University, Department of Mathematics. Obtenido de https://arxiv.org/abs/1901.07410

Dlotko, P. (20 de 08 de 2019). *BallMapper*. Obtenido de The Ball Mapper Algorithm: https://cran.r-project.org/web/packages/BallMapper/index.html

Dlotko, P. (s.f.). Imagen aplicación de BallMapper sobre una nube de puntos. *Introduction to Ball Mapper*. Obtenido de https://www.youtube.com/watch?v=M9Dm1nl_zSQ

DŁOTKO, P., & RUDKIN, S. (21 de Abril de 2020). *Visualising the Evolution of English Covid-19 Cases with Topological Data Analysis Ball Mapper.* Swansea University, Bay Campus, Mathematics Department, Swansea, United Kingdom. Obtenido de https://arxiv.org/abs/2004.03282

Edelsbrunner, H., & Harer, J. (2010). Chapter IV, Homology. En *Computational topology - an Introduction.* (pág. 92). Durham, North Carolina: American Mathematical Society. ISBN: 978-0-8218-4925-5. Obtenido de https://www.maths.ed.ac.uk/~v1ranick/papers/edelcomp.pdf

FERNÁNDEZ CASAL, R. (2020). *Repositorio para datos de COVID-19*. Obtenido de https://github.com/rubenfcasal/COVID-19

Gobierno de España. Ministerio de Hacienda y Función Pública. (s.f.). *Tabla de coeficientes de amortización lineal*. Obtenido de https://www.agenciatributaria.es/AEAT.internet/Inicio/_Segmentos_/Empresas_y_profesionales/Empresas/Impuesto_sobre_Sociedades/Periodos_impositivos_a_partir_de_1_1_2015/Base_imponible/Amortizacion/Tabla_de_coeficientes_de_amortizacion_lineal_.shtml

Goldfarb, B. (21 de 05 de 2018). *The Mapper algorithm and its applications.* Obtenido de Topology research group. Departament of computer science and mathematics: http://topology.nipissingu.ca/workshop2018/slides/Goldfarb-Data-Beamer.pdf

Instituto de Salud Carlos III. (6 de Julio de 2020). *Estudio Nacional de sero-Epidemiología de la infección por SARS-CoV-2 en España. (ENECOVID)*. Obtenido de Estudio ENE-COVID: https://portalcne.isciii.es/enecovid19/informes/informe_final.pdf

MacFarland, T. (2014). Capítulo 1. Introduction: Biostatistics and R. En T. W. MacFarland, *Introduction to Data Analysis and Graphical Presentation in Biostatistics with R. Statistics in the Large.* (pág. 2). Fort Lauderdale, FL, USA: (1st ed.). Springer International Publishing. Obtenido de https://fama.us.es/discovery/fulldisplay?docid=alma991013068778104987&context=L&vid=34CBUA_US:VU1&lang=es&search_scope=all_data_not_idus&adaptor=Local%20Search%20Engine

Minervino, M. (s.f.). Illustration of the mapper algorithm. *Topological data analysis with Mapper*. Obtenido de https://www.quantmetry.com/blog/topological-data-analysis-with-mapper/

Palma, F., Burba, D., Tunstall, L., & Boys, T. (2020 - 2021). *Visualización de datos de alta dimensión con Giotto Mapper.* Obtenido de ICHI.PRO: https://ichi.pro/es/visualizacion-de-datos-de-alta-dimension-con-giotto-mapper-151182005073367

Portal oficial de la Administración de la Junta de Andalucía. (s.f.). *Protocolo de investigación de brote epidémico.* Obtenido de Portal oficial de la Administración de la Junta de Andalucía: https://www.juntadeandalucia.es/export/drupaljda/salud_5af95879cbfd0_protocolo_brote_epidemico.pdf

Servicio de Actualización Académica y Capitalización en Estadística. (SERAACE). (s.f.). *2. DESCRIPCION DE UN CONJUNTO DE DATOS.* Obtenido de 2.1 Descripción numérica de un conjunto de datos.: http://seraace.com/files/CAPITULO-2_3ce6oggt.pdf

Wikipedia. (s.f.). Gráfico de Reeb de la función altura en un toro. *Grafo de Reeb*. Obtenido de https://es.wikipedia.org/wiki/Grafo_de_Reeb

Wikipedia. (s.f.). *Homología persistente.* Obtenido de https://en.wikipedia.org/wiki/Persistent_homology

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DEL TRABAJO FIN DE GRADO

D/Dª MARIA CASASOLA CALZADILLA con DNI 31030209N

estudiante del Grado en

INGENIERÍA INFORMÁTICA DEL SOFTWARE

de la Universidad de Sevilla durante el curso académico

de acuerdo con el art. 11.4 de la Normativa sobre Trabajos Fin de Estudios de la ETSII, y como autor/a del Trabajo Fin de Grado titulado

VISUALIZANDO LA EVOLUCIÓN DE LOS CASOS DE COVID-19 CON TDA

DECLARA QUE

1. El documento entregado para evaluación es una obra original del autor/a.
2. El documento entregado para evaluación no contiene copias de textos, copias de imágenes, copias de gráficos, ni citas sin la debida indicación del autor/autores y/o la fuente o la debida autorización del propietario, si procede.
3. Es plenamente consciente de que el hecho de no respetar estos términos es objeto de sanciones universitarias y/o de otro orden legal, como plagio.

Y, para que conste a los efectos oportunos, firma la presente declaración, en

SEVILLA

Fdo.: MARIA CASASOLA CALZADILLA

(Pie de firma con nombre y apellidos del estudiante y firma manuscrita o con certificado digital)