VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

CURSO 2022/23

Análisis estadístico para identificar el impacto de variables socioeconómicas en la distribución espacial del COVID-19 en la C.A. de Castilla y León

---

Autor

Raúl Eleazar Tizado Núñez

Tutor

José Luis Vicente Villardón

Salamanca, 2023

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

CURSO 2022/23

Análisis estadístico para identificar el impacto de variables socioeconómicas en la distribución espacial del COVID-19 en la C.A. de Castilla y León

---

Autor
Raúl Eleazar Tizado Núñez

Tutor
José Luis Vicente Villardón

Firmado por Raúl Eleazar Tizado Núñez el día SIGNDATE=03/07/2023

D. José Luis Vicente Villardón, profesor del Departamento de Estadística de la Universidad de Salamanca,

HACE CONSTAR:

Que el trabajo titulado "Análisis estadístico para identificar el impacto de variables socioeconómicas en la distribución espacial del COVID-19 en la C.A. de Castilla y León", que se presenta, ha sido realizado por Raúl Eleazar Tizado Núñez con DNI ****4380J y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 3 de julio de 2023

VICENTE VILLARDON JOSE LUIS - 07839395Y

Firmado digitalmente por VICENTE VILLARDON JOSE LUIS - 07839395Y
Fecha: 2023.07.03 11:07:34 +02'00'

Fdo.: José Luis Vicente Villardón

RESUMEN

El Covid–19 es una enfermedad que afecta al sistema respiratorio y que se ha extendido por todo el mundo desde finales del 2019. Esta enfermedad, al igual que el resto de pandemias de la historia de la humanidad ha tenido una gran influencia no solo sanitaria sino también socioeconómica. Para la realización del estudio se usaron 247 zonas básicas de salud de Castilla y León pertenecientes a 11 gerencias sanitarias. El objetivo principal es estudiar si las variables socioeconómicas tienen algún efecto en las variables sociosanitarias. Para ello, se utilizaron el análisis de correlaciones, el análisis de componentes principales y el análisis de correlaciones canónicas, así como el biplot. Los resultados muestran que las variables estudiadas se podían agrupar en cuatro grupos, uno formado por las sociosanitarias y otros 3 para las socioeconómicas que destacando la edad media, la renta bruta y los índices de desigualdad. El análisis de componentes principales muestra dos dimensiones principales, una sociosanitaria dirigida por el número de enfermos y la edad media, y otra socioeconómica por la renta bruta y el índice Gini. El número de fallecidos es una variable dependiente tanto de la dimensión sociosanitaria como socioeconómica, lo que apunta a que el fallecimiento no sólo depende de factores sanitarios sino también socioeconómicos. Este resultado está soportado por el análisis de correlaciones canónicas donde la primera correlación canónica es claramente distinta de cero.

ABSTRACT

The Covid–19 is a disease that affects the respiratory system and has spread all over the world since the end of 2019. This disease, like the rest of pandemics in the history of mankind, has had a great influence not only on health but also on socioeconomics. To carry out the study, 247 basic health areas of Castilla y León belonging to 11 health management units were used. The main objective is to study whether socioeconomic variables have an effect on socio-health variables. For this purpose, correlation analysis, principal component analysis and canonical correlation analysis, as well as biplot analysis were used. The results show that the variables studied could be grouped into four groups, one formed by the socio-health variables and another three for the socio-economic variables, which highlight the average age, gross income and inequality indexes. The principal component analysis shows two principal dimensions, one socio-health-related based on the number of patients and average age, and the other socio-economic based on gross income and inequality indexes, and a socio-economic one driven by gross income and the Gini index. The number of deaths is a dependent variable of both the socio-health and socio-economic dimensions, which suggests that death is not only dependent on health but also on socio-economic factors. This result is supported by the analysis of canonical correlations where the first canonical correlation is clearly non-zero.

CONTENIDO

1
INTRODUCCIÓN

La enfermedad conocida como Covid–19 es una enfermedad causada por el virus Sars-cov-2 que afecta sobre todo al sistema respiratorio, siendo esta una enfermedad contagiosa de origen zoonótico, es decir, está originada por un patógeno que se transmite desde los animales a los humanos. Sus principales síntomas son dificultad a la hora de respirar, dolores musculares, cansancio, fiebre y tos (McArthur et al., 2020).

La Organización Mundial de la Salud (OMS) notifica a finales de diciembre de 2019 los primeros casos de esta enfermedad en Wuhan, una población de la zona central de China donde originalmente se aisló (Zaim et al., 2020), y desde donde se extiende al resto del mundo generando una pandemia. Esta pandemia origina un impacto en la humanidad a nivel político, sanitario, económico y social. Comparando con otras grandes pandemias de la historia, como la gripe española o las peste negra, se puede observar que los factores sociales y económicos siempre han sido significativos a la hora de evaluar la gravedad y afectan a la velocidad de recuperación de la misma (Sánchez-González, 2021). Diversos investigadores ya han hecho referencia a que nos encontramos en una crisis económica en la humanidad a causa de esta enfermedad (Esparza-Rodríguez et al., 2021).

En España se detecta por primera vez el 31 de enero de 2020, y no llega a ser diagnosticada en Castilla y León hasta el 27 de febrero del mismo año (Riquelme, 2020). En este trabajo analiza el Covid–19 en la Comunidad Autónoma de Castilla y León, para ello se estudian diversas variables sociosanitarias que están relacionadas con el Covid–19 y la posible influencia de ciertas variables socioeconómicas en la afección de esta enfermedad. Para ello se ha tenido en cuenta la organización espacial del sistema de salud de C.A. de Castilla y León, más concretamente en las áreas de salud como unidad territorial básica en este sistema.

Según el Ministerio de Sanidad (2023), se entiende por área de salud: "una zona administrativa la cual tiene un grupo de centros y de profesionales sanitarios de atención primaria bajo su dependencia tanto funcional como organizariva". Son la forma principal de organización dentro del sistema sanitario de España, siendo este quien lleva a cabo las prestaciones y los programas sanitarios de cada una de las Comunidades Autónomas. Dentro de las zonas de salud se realiza una división en zonas básicas formadas por centros de especialidades, centros de salud y hospitales.

La división de las áreas sanitarias de la Junta de Castilla y León (2007) se desarrolla en el Decreto 32/1988 (BOCyl nº 41, 1988) y sus posteriores modificaciones. En este decreto se establece la delimitación territorial de las zonas básicas de salud y los municipios que la integran en la C.A. de Castilla y León. En total existen 247 zonas básicas de salud organizadas en 11 Gerencias de Atención Sanitaria (Figura 1.1) generalmente uniprovinciales excepto en León y Valladolid donde existen 2 gerencias en cada una de ellas.

1.1 Antecedentes

Varios estudios han analizado el efecto de los factores socioeconómicos en el Covid–19 , a continuación se aportan algunos de ellos que sirvieron de base para este trabajo.

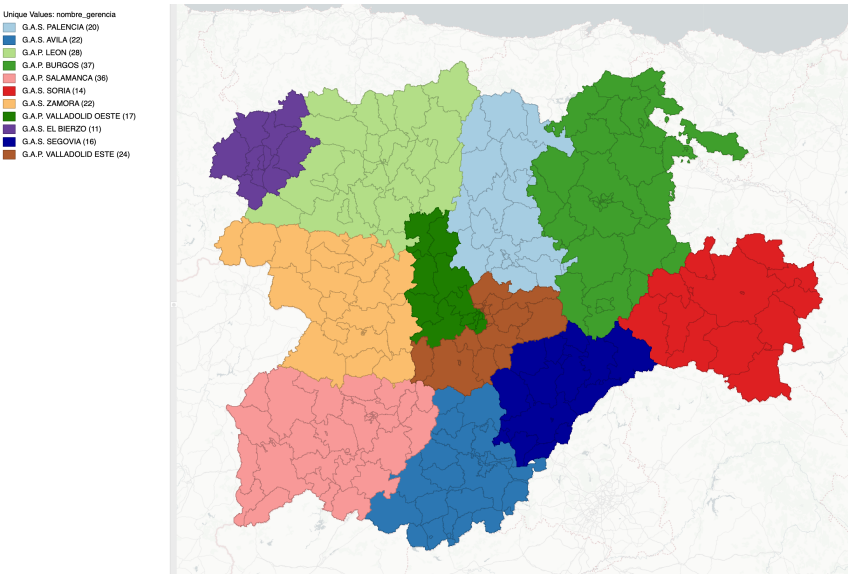


Figura 1.1: En la imágen podemos observar las distintas Gerencias de Atención Sanitaria en la C.A. de Castilla y León y su división en Zonas Básicas de Salud. En la leyenda se indica el nombre de las gerencias y entre paréntesis el número de zonas básicas de salud incluidas en cada una de ellas.

Wachtler et al. (2020) es un trabajo sobre una investigación epidemiológica social, en ella se describen las distintas correlaciones que existen entre lo afín a lo socioeconómico y el riesgo al que puede enfrentarse la población tanto de enfermarse como de morir. Fue un pionero de esto en cuanto a Covid–19 se refiere En el trabajo se realiza una revisión de la literatura de la investigación internacinal disponible en ese momento, para ello se encontraron 138 artículos distintos, de los cuales se incluyeron 46 de ellos en el análisis que se realizó. Se obtuvieron diversos resultados, donde cabe destacar que tanto en Reino Unido como en Estados Unidos se observaron desigualdades económicas en cuanto al riesgo de infección se refiere, al igual que en la gravedad del curso de la enfermedad, también cabe destacar que se vio como las poblaciones socioeconómicas de un carácter menos privilegiado se vieron más afectados. Para la fecha de publicación de este trabajo, no se contaba con la suficiente información para obtener resultados significativos tanto para Alemania como para el resto de países europeos. A pesar de ello, y con la escasez de datos, ya se pudieron obtener indicativos de que las desigualdades sociales eran un factor a tener en cuanto a la hora de hablar del Covid–19 . Finalmente se habla de que en la mayor parte del análisis de los datos han sido de estudios ecológicos, siendo unos pocos de los que se obtuvieron diferencias socioeconómicas a la hora de hablar a nivel individual, donde dicho nivel individual nos da una mayor comprensión del tema.

Brandily et al. (2021) indica que generalmente ya se pueden encontrar desigualdades en la mayor parte de países en cuanto a mortalidad se refiere, donde estas diferencias pueden verse aumentandas por las pandemias. También se comenta que la gran diferencia de casos confirmados por Covid–19 puede deberse a ésto, aunque en el momento de realización del estudio se contaban con una escasez de evidencias para poder afirmar que esto ocurría debido a las desigualdades existentes. El artículo se centra en Francia, uno de los países más afectados por el Covid–19 en el mundo, en él se toman datos a nivel municipal para buscar estudiar una posible relación entre los ingresos y la mortalidad de las zonas urbanas en las distintas olas de Covid–19

. Se obtuvieron resultados tales como que en las zonas más pobres había un 30% más de mortalidad. Finalmente se obtuvo que factores como la exposición laboral o las condiciones de vivienda tenían un efecto inducido en la enfermedad y las desigualdades en cuanto a la mortalidad, pero cabe destacar que dependían del estado de la epidemia.

Rudi Rocha Rifat Atun (2021) es un estudio de Brasil, donde se proporciona una muy importante evidencia de la relación existente entre distintas variables socioeconómicas y los resultados de la atención hospitalaria referentes al Covid–19 . En el estudio se habla de las inequidades preexistentes, donde las define como situaciones donde se presenta una desigualdad que viola la justicia social, en cuanto a ellas, se observó que estas se vieron aumentadas por la pandemia, especialmente en cuanto a la atención y las tecnologías del sector sanitario. Cabe destacar que los resultados también tienen que ver con la crisis política y las medidas tomadas desde 2015 en Brasil. En el estudio se habla de la precaria situación en Brasil, la cual se ha visto empeorada por el Covid–19 en todos los ámbitos.

Laajaj et al. (2022) muestra como a lo largo de distintos países del mundo, el Covid–19 ha tenido un efecto desproporcionado, afectando de una forma mayor a los grupos económicos más desaventajados. Esto puede deberse a distintos motivos, donde cada uno se debe interpretar de forma distinta y teniendo en cuenta también las situaciones políticas de cada lugar. Se tomaron datos de un país de nivel bajo - medio económico, donde se observó un mayor impacto de la pandemia en los más desfavorecidos. Posteriormente se combinó un modelo epidemiológico con datos ricos de Bogotá, en Colombia, donde se pudo demostrar que la desigualdad de las infecciones se deben en gran medida a la desigualdad de poder trabajar de forma no presencial y en las tasas de ataque en los hogares. Se obtuvieron también resultados donde se muestra que las desigualdades de aislamiento son algo menores, pero para nada despreciables, en cambio lo referido a los contratos no guarda prácticamente relación, principalmente por su lentitud frente al virus. Finalmente se puede hablar de las intervenciones en orden a la mitigación de la transmisión, la cual es mucho más eficaz cuando se dirigía grupos de un perfil socioeconómico más bajo.

1.2 Objetivos

Este trabajo tiene como objetivo general conocer si alguna de las variables socioeconómicas seleccionadas han tenido o no una influencia en la enfermedad Covid–19 a lo largo de las zonas básicas de salud de la Comunidad Autónoma de Castilla y León.

Para alcanzar este objetivo se ha decidido tener en cuenta una serie de objetivos específicos:

1. Describir las variables sociosanitarias y socioeconómicas más importantes que se han utilizado en este estudio.

2. Estudiar la posible relación entre las principales variables sanitarias del Covid–19 y las variables socioeconómicas en los parámetros usados para describir el desarrollo del Covid–19 en la C.A. de Castilla y León.

3. Conocer la relación de las diversas variables socioeconómicas y las variables sanitarias del Covid–19 (p.ej. número de enfermos, número de fallecidos, prevalencia) o saber si se puede calcular un indicador general de dichas variables.

1.3 Estructura del documento

Según las normas de estilo, presentación y estructura del Trabajo Fin de Grado de Estadística de la Universidad de Salamanca (2023) el contenido de la memoria se ha organizado en Introducción, Desarrollo y Bibliografía y referencias.

El Desarrollo de este trabajo está dividido en diversas secciones:

1. Se comienza con los distintos métodos estadísticos empleados para la realización del trabajo, primero se habla del Análisis de Correlaciones, posteriormente del Análisis de Componentes Principales, el siguiente método se trata de un Análisis de Correlaciones Canónicas, finalmente se habla del Biplot siendo este último método el utilizado para la representación gráfica.

2. La siguiente parte del desarrollo trata de las bases de datos y los programas utilizados. Se empieza indicando cómo se obtuvieron las bases de datos empleadas en el trabajo, dónde también se apunta el principal programa utilizado para realizar los análisis y los mapas empleados en el trabajo para realizar el análisis descriptivo. Finalmente se habla de las distintas bases de datos con una mayor profundidad, comenzando por las que contienen datos sociosanitarios y finalizando con las que contienen datos de variables socioeconómicas.

3. La parte final del trabajo está dedicada a los resultados y su posterior discusión. Se inicia realizando un análisis descriptivo de los datos y su distribución espacial, posteriormente se explican los resultados obtenidos mediante el análisis de componentes principales, análisis de correlaciones, y sus Biplot. En esta parte se comentan y discuten los resultados, siendo las conclusiones del trabajo la parte final del trabajo.

2
DESARROLLO

El desarrollo de este trabajo se ha llevado a cabo a partir de la utilización de diversas bases de datos tanto socieconómicas como sociosanitarias. Estas bases de datos son de acceso público y están relacionadas con el Covid–19 . Una vez realizada la descarga de todas las bases de datos necesarias para la realización del trabajo, se han ordenado y modificado para extraer la información. Una vez se tuvieron las bases de datos preparadas y listas, se utilizó el programa de R para poder analizarlas, también se utiliza el programa conocido como Geoda para realizar los distintos mapas empleados en el trabajo.

R es un entorno que cuenta con una licencia GNU GLP, es decir, estamos hablando de un software libre y lenguaje de programación interpretado. Esto quiere decir que es capaz de ejecutar las instrucciones que recibe sin necesidad de haber realizado una compilación previa entre el lenguaje máquina y las instrucciones que ha recibido. Se presentó en 1993 y está orientado a realizar cálculos Estadísticos.

Geoda es un programa empleado para el análisis de datos espaciales. Fue lanzado en el 2003, y es utilizado por su gran utilidad para analizar datos, el cual es realizado principalmente mediante el modelado de patrones espaciales. Se pueden utilizar una gran variedad de datos vectoriales en una gran cantidad de formatos, pudiendose trabajar con GeoJson, geodatabases y otros formatos de datos vectoriales.

2.1 Métodos estadísticos

2.1.1 Análisis de Correlaciones

El análisis de correlaciones es una técnica estadística que nos ofrece información de cómo se relacionan dos variables que estamos estudiando. El concepto de correlación fue desarrollado por Francis Galton en 1888, pero la fórmula matemática que utilizamos a día de hoy sería introducida por Karl Pearson en 1905.

La correlación viene determinada por el coeficiente de correlación, el cual se limita a valores entre -1 y $+1$. Dependiendo del valor numérico obtenido en el coeficiente de correlación, se puede hablar de la fuerza y de la dirección de la correlación.

La dirección de la correlación viene determinada por el signo del coeficiente de correlación. Así, hablamos de correlación positiva cuando obtenemos valores positivos, es decir, entre 0 y $+1$; que aparece cuando los valores mayores de la primera variable se relacionan con los valores mayores de la segunda variable. Por el contrario, existe una correlación negativa cuando obtenemos valores entre -1 y 0, lo que surge cuando los valores mayores de la primera variable se relacionan con valores menores de la segunda variable.

También podemos hablar de la fuerza de la correlación que viene determinmada por el valor absoluto del coeficiente de correlación. Esta fuerza indica el grado de relación, correlación, entre las dos variables de la siguiente forma:

- ⋆ 0.0 a 0.1 : no hay correlación entre nuestras variables (incorrelacionadas o independientes).
- ⋆ 0.1 a 0.3 : hay poca correlación entre nuestras variables.
- ⋆ 0.3 a 0.5 : hay correlación entre nuestras variables.
- ⋆ 0.5 a 0.7 : hay una correlación alta entre nuestras variables.
- ⋆ 0.7 a 1.0 : hay una correlación muy alta entre nuestras variables.

Si obtenemos que dos variables están relacionadas entre sí, podemos comprobar posteriormente si pueden ser utilizadas para predecir la una a la otra. Pero siempre hay que tener en cuenta que no siempre van a ser relaciones causales. Esto quiere decir que todas las correlaciones que sean descubiertas se tienen que investigar más y que nunca deben ser interpretadas en términos de contenido (debemos tener en cuenta que es lo que estamos estudiando), sea o no sea evidente.

En el caso de que estemos trabajando con una muestra, y obtenemos que hay correlación en dicha muestra, será necesario comprobar si también hay correlación en la población. Para ello comprobamos si es lo suficientemente significativo mediante la realización de una prueba t. La hipótesis nula más habitual es comprobar si el coeficiente de correlación es diferente de 0, es decir, lo que estamos realizando es comprobar si hay independencia lineal. Para el contraste de hipótesis, la hipótesis alternativa es que sí exista correlación entre las variables estudiadas.

Para la prueba de hipótesis primero se fija el nivel de significación, el cual generalmente es del 5%. Si el valor que calculamos (p-valor) es menor del 5%, rechazaremos la hipótesis nula y aceptaremos por tanto la alternativa. En el caso de que podamos rechazar la hipótesis nula, podremos afirmar la existencia de correlación entre las dos variables en la población.

Para realizar el análisis de correlaciones generalmente se pueden utilizar dos métodos: Pearson o Spearman.

En el análisis de correlación de Pearson, obtenemos resultados sobre la correlación lineal de dos variables x e y de escala métrica. Para efectuar su cálculo se utiliza la covarianza entre ambas variables. Esta covarianza nos dará un valor positivo si existe una correlación positiva entre las variables y un valor negativo si existiese una correlación negativa. Esta covarianza se calcula con la siguiente fórmula:

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2.1}$$

Sin embargo, esta covarianza nos genera un problema, ya que no está normalizada y esto hace que los valores que obtiene pueden ser más y menos infinitos. Esto hace que comparar la fuerza de las relaciones pueda llegar a ser difícil. Para ello se calcula el conocido como coeficiente de correlación, el cual también recibe el nombre de correlación producto–momento. El coeficiente de correlación se obtiene corrigiendo el problema de la covarianza normalizándolo con las varianzas de las variables implicadas. Este coeficiente de correlación se calcula con la siguiente fórmula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.2}$$

Su interpretación es como se indicó anteriormente, cuanto mas cerca de $+1$ mayor correlación positiva, cuanto más cerce de -1 mayor correlación negativa y con un valor de 0 cuando no existe correlación entre las variables.

El análisis de correlación de Spearman se utiliza para calcular la relación de dos variables de medición ordinal. Es el equivalente no paramétrico del análisis de correlaciones de Pearson. Por ello este método se emplea cuando nuestros datos no son métricos o no siguen una distribución normal. Este método trata prácticamente las mismas cuestiones que las del coeficiente de correlación de Pearson pero se denomina correlación de rangos de Spearman.

Este cálculo de correlaciones de rango se basa en el sistema de clasificar series de datos. Esto quiere decir que los valores que medimos no van a ser usados para el cálculo, sino que serán transformados en rangos. El valor calculado nuevamente se encuentra entre -1 y $+1$, y se mantiene la misma interpretación que en el coeficiente de correlación de Pearson.

Cuando se manejan conjuntos grandes de variables y se realiza un análisis de correlación por pares, una forma eficiente de visualización son los mapas de calor (*heatmap*) de correlaciones. Este tipo de gráficos se utiliza para poder visualizar los datos en forma de matriz, donde cada elemento de la matriz (celda) es coloreado según el valor de la correlación. Los mapas de calor tienen una gran utilidad para poder visualizar patrones y tendencias, por lo que es perfecto para visualizar las correlaciones. No obstante, a la hora de explorar grupos de variables con correlaciones similares, la ordenación de las variables es una parte fundamental de los mapas de calor.

En este trabajo, para realizar la ordenación de la matriz de correlaciones y poder así detectar grupos de variables con una elevada correlación se ha utilizado la ordenación derivada de un análisis de clúster jerárquicos con el método de agregación de Ward.

El método de agregación de Ward o de varianza mínima de Ward emplea las distancias existentes entre dos clases A y B sin elementos comunes, es decir, no vacías y disjuntas, con el objetivo de ir agrupando de dos en dos las clases que incrementen menos la inercia dentro de las clases. La distancia de Ward función objetivo a minimizar entre los dos grupos sería:

$$W(A,B) = \frac{p_A p_b}{p_A + p_B} d^2(g_A - g_B) \tag{2.3}$$

siendo p_A y p_B los pesos de cada clase; g_ay g_b los centros de gravedad de cada clase; y d la distancia euclídea.

En el caso del mapa de calor de correlaciones, se emplea una modificación de esta ecuación ya que utiliza el coeficiente de correlación (Fórmula 2.2) para calcular la distancia entre las variables en vez de usar la distancia euclídea. Para transformar los valores de correlación en valores de distancias que puedan ser utilizados por los métodos de clúster jerárquicos, la distancia entre dos individuos i y j se calcula como:

$$d(i,j) = 1 - r_{ij}$$

Una de las condiciones de una matriz de distancias es que la distancia de un elemento sobre sí mismo es 0. Con este cálculo se cumple esta condición ya que la correlación de una variable sobre sí misma es $r_{i,i} = 1$ y la distancia $d(i,i) = 1 - r_{i,i} = 0$. Además, la máxima distancia entre variables se encuentra cuando las variables tienen una correlación negativa perfecta $r_{i,j} = -1$, con lo que la distancia entre ambas variables sería $d(i,j) = 1 - r_{i,j} = 2$.

2.1.2 Análisis de Componentes Principales

El análisis de componentes principales (ACP) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos multidimensionales y revelar la estructura subyacente en los datos.

La mayoría de los investigadores considera la publicación de Pearson (1901) la primera relacionada con el ACP. En ella nos proponía una definición basada en el Análisis de Regresión Lineal, esto es debido a que su tratamiento se buscaba la línea de mejor ajuste de los datos. Posteriormente, Hotelling (1933) empieza a hablar de las componentes principales como combinaciones lineales de las variables originales, donde también mostró la variabilidad absorvida por ellas, además de mostrar la relación entre las cargas de estas componentes principales y los valores propios de la matriz de covarianzas. Finalmente se demuestra que esas cargas se corresponde con los vectores propios de la matriz de covarianzas. Eckart y Young (1936) publicarían un documento que hablaba sobre la Descomposición en Valores Singulares. El objetivo de este procedimiento era permitir descomponer una matriz de datos inicial en tres matrices distintas. La primera de ellas contendría los valores propios ortogonalesa las filas de la matriz inicial. La segunda de ellas sería una matriz diagonal, esta contendría los valores singulares de la matriz inicial. La tercera de ellas contendría los valores propios ortogonales a las columnas de la matriz inicial. Con la utilización de esta descomposición en valores singulares se abrieron las puertas a poder continuar con distintos estudios del ACP, ya que el empleo de esta técnica significó un gran avance a la hora de hablar de algoritmos de solución.

Actualmente se entiende que el ACP clásico es una técnica de reducción de la dimensionalidad de los datos que se suele emplear en los estudios multivariantes. Para su aplicación se utiliza una representación espectral de la matriz de varianzas–covarianzas de los datos que se estén analizando. Con ello se busca crear nuevas variables denominadas componentes principales, las cuales son combinanciones lineales de las variables originales de nuestro conjunto de datos, éstas corresponden a los vectores propios normalizados de la matriz de varianzas–covarianzas.

A continuación se comentará esta técnica utilizando una matriz X y de la descomposición en valores singulares (González-García & Taborda-Londoño, 2015; Ramírez-Figueroa, 2021).

Dentro de la estadística multivariante hay un núcleo al que sin duda alguna pertenecen los valores y vectores propios. Allí solemos encontrar el siguiente problema con expresión $Av = \lambda$, esto significa que estamos buscando encontrar un vector v que al ser multiplicado por una matriz cuadrada A, no alteren su dirección y además que su sentido sea maximizado.

Si dispusiesemos de una muestra de n individuos u objetos y midiésemos p variables aleatorias centradas tal que $X_1, X_2, ..., X_p$. El objetivo será el encontrar $k < p$ variables $Z_1, Z_2, ..., Z_k$ que no estén correlacionadas entre sí y que sean combinaciones lineales de las variables iniciales (X), de esta forma lo que estamos buscando es que expliquen la mayor parte de la variabilidad perteneciente a las variables iniciales.

Las k componentes principales se expresan de la misma forma, pudiendose expresar la primera como combinación lineal de X de la siguiente forma:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + ... + u_{1p}X_{pi}$$

con $i = 1, 2, ..., n$ individuos (filas).

En forma matricial sería así:

$$\begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \\ z_{1n} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{pmatrix}$$

equivalente a $Z_1 = Xu_1$.

El valor esperado y la varianza de Z_1 vendrían dados por:

$$\mathrm{E}[Z_1] = \mathrm{E}[Xu_1] \ = \mathrm{E}[X]u_1 = 0$$
$$\mathrm{Var}[Z_1] = \frac{1}{n}\sum_{i=1}^{n} Z_{1i}^2 = \frac{1}{n} Z_i' Z_1 \ = u_1' \frac{X'X}{n} u_i$$

La primera componente principal sería aquella combinación lineal que tenga varianza máxima y que cumpla $\|u_1\| = 1$. Aquí surge un problema de optimación, el cual trata de determinar la combinación lineal de Z_1 que posea la varianza máxima y que este sujeta a la anterior restricción.

Para poder resolverlo se utiliza el siguiente lagrangiano:

$$L = u_1' V u_1 - \lambda(u'u - 1)$$

Lo siguiente a realizar será derivar e igualar a cero la expresión del lagrangiano:

$$\frac{\delta L}{\delta u_1} = 2Vu_1 - 2\lambda u_1 = 0$$

obteniendose que $$Vu_1 = \lambda u_1$$

Con todo esto logramos saber que u_1 puede ser considerado como el vector propio que esta asociado al valor propio de λ. Para poder resolverlo, la matriz V debe ser una matriz semidefinida positiva y simétrica, esto significa que tiene p valores propios que podemos ordenar de forma decreciente tal que:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$$

Finalmente, si multiplicasemos a la ecuación anterior por u_1' obtendríamos que:

$$\underbrace{u_1' V u_1}_{\mathrm{Var}[Z_1]} = \lambda u_1' u_1$$

Sabemos que u_1 es unitario, por ello tenemos que $\mathrm{Var}[Z_1] = \lambda$, como nuestro objetivo es que la varianza de Z_1 sea máxima tomaremos que $\lambda = \lambda_1$ como el mayor valor propio de V.

Por todo ello, podemos decir que los coeficientes de la combinación lineal de Z_1 están dados por dos componentes: vector propio u_1 asociado al valor propio λ_1:

$$Z_1 = Xu_1$$

En resumen, el ACP se suele emplear con el objetivo de revelar como es la estructura subyacente de datos multivariantes. Para alcanzar este objetivo se suelen seguir los siguientes pasos:

⋆ Se utiliza como matriz V la matriz de varianzas–covarianzas obtenida de los valores de X centrados en cero.

Para ello es necesario restar la media de cada variable a cada observación para eliminar cualquier tendencia en los datos y asegurar que el origen esté en el centro del conjunto de datos. También es posible utilizar la matriz de correlaciones como matriz V.

⋆ Se procede a la descomposición espectral de la matriz.

Esto implica calcular los autovalores y autovectores de la matriz de covarianza. Los autovalores representan la varianza explicada de cada componente principal, mientras que los autovectores representan la dirección de cada componente principal.

⋆ Se procede a la interpretación de los resultados a partir de los autovalores y los autovectores obtenidos.

Los autovalores más grandes indican las componentes principales más importantes y, por lo tanto, la cantidad de varianza explicada por cada componente. Se puede calcular el porcentaje de varianza explicada por cada componente principal dividiendo el autovalor de esa componente entre la suma de todos los autovalores.

Los autovectores indican las combinaciones lineales de las variables originales que definen cada componente principal. Los coeficientes en los autovectores representan las ponderaciones de cada variable en la combinación lineal. Si un coeficiente es cercano a cero, significa que la variable tiene poca influencia en esa componente principal.

2.1.3 Análisis de Correlación Canónica

El análisis de correlación canónica o análisis canónico de correlaciones (ACC) es un método de análisis estadístico linear multivariante que pretende analizar las relaciones entre dos grupos de múltiples variables (datos multivariantes).

Descrito inicialmente por Hotelling (1936) para estudiar un conjunto de test psicológicos y un conjunto de medidas biométricas, se diferencia de la regresión múltiple en que se intenta conocer la relación de múltiples variables dependientes, no sólo de una única variable dependiente continua, en función de múltiples variables independientes. Por ello, puede considerarse una generalización del análisis de regresión múltiple y, en general, un modelo general en el que se basan otras técnicas multivariantes.

Así, siendo N el número de individuos, n el número de variables del primer conjunto de datos y m el número de variables del segundo conjunto de datos; se definen dos matrices de datos multivariantes:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nn} \end{pmatrix}$$

considerada habitualmente como variables dependientes, y

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nm} \end{pmatrix}$$

considerada como variables independientes.

Para obtener el ACC, es necesario calcular las funciones canónicas U y V que permitan transformar los datos de la forma:

$$X_u = X \cdot U$$

$$Y_v = Y \cdot V$$

con las siguientes características (Cuadras, 1981):

- La correlación entre las columnas de X_u o entre las columnas de Y_v es nula.
- La correlación entre las i-ésimas columnas de X_u y Y_v es la correlación canónica r_i. Teniendo en cuenta que al menos únicamente las primeras $p = min(m, n)$ variables canónicas no seran nulas, se ha de cumplir la condición de que

$$r_1 \geq r_2 \geq \cdots \geq r_p$$

Las ecuaciones generales para realizar el cálculo de las correlaciones canónicas a partir de las matrices de correlaciones entre las variables dependientes R_{yy}, las variables independientes R_{xx} y la cruzada R_{yx}, sería:

$$R = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R'_{yx}$$

Para el ACC se calculan los eigenvalores y eigenvectores de la matriz R que permiten la reducción de dimensionalidad, resolviendo:

$$\left| R_{yx} R_{xx}^{-1} R'_{yx} - \lambda R_{yy} \right| = 0$$

Los eigenvalores se corresponden con al cuadrado de las correlaciones canónicas

$$\lambda_i = r_i^2$$

Los eigenvectores U y V son utilizados con las variables originales para obtener las variables transformadas X_u y Y_v. Se calculan recolviendo para cada eigenvalor λ_i:

$$(R_{yx} R_{xx}^{-1} R'_{yx} - \lambda_k R_{yy}) u_i = 0$$

$$(R'_{yx} R_{yy}^{-1} R_{yx} - \lambda_k R_{xx}) v_i = 0$$

Hay que tener en cuenta que con el uso de las matrices de correlaciones R se obtienen variables canónicas relacionadas con las variables originales estandarizadas. Si se realiza el cálculo a partir de las matrices de varianzas-covarianzas S el resultado estaría relacionado directamente con las variables originales.

Mediante este procedimiento de cálculo, el ACC identifica la estructura óptima a través de una reducción de dimensionalidad que maximiza la correlación entre los dos conjuntos de variables. Además, se obtienen una serie de funciones o variables canónicas que maximizan la correlación entre las combinaciones lineales de las variables dependientes e independientes, es decir, la correlación entre dos funciones canónicas, una para las variables dependientes y otra para las variables independientes. Con este método, se puede evitar el análisis individual de correlación de cada una de las variables dependientes con el conjunto de variables dependientes.

La reducción de dimensionalidad derivada de la obtención de las variables canónicas lo hace similar al ACP. Sin embargo, en vez de obtener una combinación linear de variables que explican la máxima varianza dentro de un conjunto de variables (ACP), se obtinen varias funciones lineares que explican la máxima correlación entre dos grupos de variables (ACC).

En resumen, el objetivo principal del ACC es revelar la estructura común de los dos grupos de datos multivariantes y conocer si los dos grupos de variables son independientes y, en su caso, calcular el grado de correlación entre los dos grupos de variables. Para alcanzar este objetivo se suelen seguir los siguientes pasos:

1. Obtener las combinaciones lineales (funciones canónicas) de cada grupo de variables, independientes y dependientes, que maximicen la correlación entre ellas.

   Este objetivo está relacionado con la reducción de dimensionalidad de este método estadístico que permite condensar un número elevado de variables originales en un grupo menor de variables derivadas con una pérdida mínima de información.

2. Explicar las relaciones existentes entre ambos grupos de variables, generalmente a partir de la contribución relativa de cada variable en las funciones canónicas.

A la hora de evaluar el ACC es que la multicolinearidad, es decir, la dependencia linear o alta correlación entre las variables, es un problema que hay que tener en cuenta y se debe eliminar en los casos más graves. En este trabajo se ha utilizado el análisis de correlaciones y los mapas de calor para eliminar los casos más evidentes de colinearidad.

Finalmente indicar que la interpretación del ACC es más dificil que la de otros análisis multivariantes. Normalmente, la utilización de esta técnica es explorar la estructura multidimensional de los datos y revelar relaciones lineares entre los conjuntos de datos. Por ello, las representaciones gráficas son un elemento importante de los resultados, en particular, los Biplots.

2.1.4 Análisis Biplot

Los Biplots son un tipo de visualización de carácter exploratorio que se emplea en Estadística con matrices de datos multivariantes (Gower & Hand, 1996). Podemos decir que esta técnica es una generalización multivariante de los diagramas de dispersión de dos variables donde, al igual que un diagrama de dispersión, nos muestra como es la distribución conjunta de dos variables pero un Biplot es capaz de representar tres o más variables.

El Biplot fue introducido por el estadístico Rubén Gabriel (1971). Inicialmente el Biplot era simplemente una representación del Análisis de Componentes Principales basado en la descomposición en valores singulares pero rápidamente se extendería al Análisis Canónico basado en descomposición canónica para analizar grupos de variables (Gabriel, 1972). En 1985, se realizo otra versión simétrica que guardaba un parecido con el Análisis de Correspondencia, este era capaz de representar filas y columnas manteniendo la misma calidad (Galindo, 1986). Yang y Kang (2003) describieron diversos métodos que podían ser empleados para visualizar e interpretar los Biplots. Vicente-Villardón et al. (2006) proponen un Biplot lineal que se empleaba para datos binarios, este sería denominado como Biplot Logístico debido a que la relación que se observa entre las variables y las dimensiones del Biplot son modelados mediante una curva de respuesta logística.

El análisis Biplot representa la distribución de una muestra multivariante en un espacio de dimensión reducida, generalmente esta es de dimensión dos, así luego se superpone sobre la

representación de las variables sobre las que se ha medido la muestra. La representación de las variables se suele realizar mediante vectores, estos coinciden con direcciones, las cuales son las que mejor son capaces de mostrar el cambio individual de cada variable. "Bi" es un prefijo el cual hace referencia a la superposición de la misma representación, incluyendo también a los individuos y a las variables. La principal utilidad de los Biplots es representar y describir de una forma gráfica los datos que se están estudiando o para poder mostrar los resultados que son proporcionados por algunos modelos multivariantes.

Si tuvieramos que describir cual es la forma más sencilla de los Biplots, diriamos que es un diagrama de dispersión donde los individuos vienen representados por puntos y las variables vienen representadas por los vectores. En el caso de que tengamos tres o más variables la representación se complica ya que para su visualización se precisan tantas dimensiones espaciales como variables, es por ello que generalmente las distribuciones multivariantes tienen una mayor dificultad de visualización.

Para emplearlos debemos comprender que los Biplots son muy importantes debido a que su interpretación se basa en diversos conceptos geométricos de relativa sencillez, ya que estos suelen formar parte de los conocimientos matemáticos básicos de los posibles usuarios de esta técnica. Así, en la visualización Biplot:

- Podemos definir la similaridad entre los individuos como una función inversa de la distancia entre los propios individuos.

- Podemos decir también que en algunos tipos, tanto los ángulos como las longitudes de los vectores que están representando a las variables, pueden ser interpretados tanto en términos de variabilidad como en términos de covariabilidad respectivamente.

- Las relaciones que haya entre las variables y los individuos son interpretádos en términos del producto escalar, es decir, son interpretados en términos de las proyecciones de los puntos (los individuos) sobre los vectores (las variables).

Si hablamos del análisis Biplot, también debemos hablar de su definición matemática. Sea $X_{n\times p}$ la matriz que representa los datos multivariantes que queremos representar. Donde las n filas corresponden los individuos y las p columnas a las variables medidas sobre los individuos. Cabe destacar que también se puede realizar una representación de otro tipo de matrices, donde las filas y las columnas pueden corresponder a niveles de dos factores de clasificación, por ejemplo.

Podemos decir que un Biplot para una matriz que contenga los datos X, es una representación gráfica utilizando vectores (que llamaremos marcadores): $g_1, g_2, ..., g_n$ cuando estemos hablando de las filas y $h_1, h_2, ..., h_p$ para cuando estemos hablando de las columnas. El Análisis Biplot consiste en obtener estos vectores de forma que el producto interno $g_i' \cdot h_j$ se aproxime a x_{ij} de la matriz inicial de la mejor forma posible.

Si esta representación fuese en 2D, donde todos los marcadores tendrían dos coordenadas. Estas serían bien las del punto que este representando a la fila o a la columna en el Biplot.

Si considerasemos los marcadores $g_1, ..., g_n$ como las filas de una matriz que llamaremos G y los marcadores $h_1, ..., h_p$ como filas de una matriz que llamaremos H. Con esto podríamos decir que X es aproximable al producto de GH', y la estructura de la matriz X se puede visualizar mediante la representación de los marcadores en un espacio euclídeo, el cual generalmente es de dos o tres dimensiones.

También cabe destacar las propiedades que poseen tantos los marcadores filas como los marcadores columna a la hora de realizar la representación gráfica de la factorización que ha sido seleccionada. Esta tiene que ver con el tipo de métrica que haya sido introducido en los espacios de filas o de columnas (Cárdenas et al., 2007), pudiéndose definir varios tipos de biplots:

1. GH Biplot: se aproxima la matriz Y con la restricción $U'U = I$, donde U es una matriz de vectores singulares ortonormales e I es la matriz identidad. Con esta restricción, se mantiene la métrica existente entre las distintas columnas y con un nivel de calidad que pueda ser considerado como óptimo.

   La utilización de un GH Biplot será recomendable para la aproximación de las varianzas que se encuentran en la matriz $(Y'Y)$ a través de la matriz (BB'). En cambio, no lo será recomendable para las distancias euclídeas en la matriz $(Y'Y)$, debido a que a través de la matriz (AA') se reproducen las distancias de Mahalanobis que están contenidas en $Y(Y'Y)^{-1}Y'$.

2. JK Biplot: la aproximación de la matriz Y se realiza con una restricción distinta: $V'V = I$, siendo V una matriz de vectores singulares ortonormales. Con esta restricción, se mantiene la mátrica entre las filas, pudiéndose demostrar que tienen un nivel de representación que se puede considerar óptima.

   Por todo ello, el JK Biplot es bastante recomendable para realizar una aproximación entre los individuos de la matriz (YY') a través de (AA'), pero a diferencia del GH biplot, no lo es para las varianzas ya que a través de (BB') se obtendria una ponderación de las varianzas que están contenidas en $Y'(YY')^{-1}Y$.

3. SQ Biplot: la aproximación es única y se obtiene un resultado simétrico para las filas y para las columnas. En este caso las restricciones $A'A \neq I$ y que $B'B \neq I$.

   Este tipo de Biplot solo se emplea en los casos donde el objetivo del análisis sea principalmente en la aproximación de elementos y_{ij} de la matriz Y, como es el caso de las tablas de contingencia.

Estos Biplots se pueden denominar Biplots clásicos, pero se han desarrollados nuevos procedimientos para describir matrices rectangulares. Destacar entre ellos algunas de las contribuciones existentes:

⋆ HJ Biplot: Galindo (1986) propuso este tipo de Biplot alternativo para la obtención de altas calidades a la hora de representar filas y columnas, para ello llevo a cabo la siguiente factorización $Y = (U\Sigma)(\Sigma V') = AB'$.

   El principal problema de este Biplot es que no permite reproducir los datos iniciales debido a que $Y = U\Sigma^2V'$. Algunos autores no consideran al HJ Biplot como un biplot estrictamente hablando aunque es muy utilizado debido la alta calidad a la hora de representar tanto las filas como las columnas, lo que permite interpretar las relaciones entre ambas al igual que ocurre en el Análisis de Correspondencias.

⋆ Biplots Generalizados: Vicente-Villardón (1992) propuso este tipo de Biplots alternativo que permite considerar la importancia de las posibles diferencias existentes entre los individuos y las variables. Para ello se intruducen métricas definidas positivas (Ω, Φ) tanto en los espacios de filas como de columnas de la forma: $U'\Omega U = I$ y $V'\Phi V = I$ respectivamente. Ello es posible aproximando la matriz de datos Y mediante DVS Generalizada (DVSG) utilizando $X = \Omega^{1/2}T\Omega^{1/2} = P\Sigma Q'$ siendo $U = \Omega^{-1/2}P$ los vectores generalizados por la derecha,

$V = \Omega^{-1/2}Q$ los vectores generalizados por la izquierda y Σ la matriz diagonal con los valores singulares generalizados.

Los Biplots clásicos se pueden considerar casos particulares del Biplot Generalizado relacionados con Análisis de Componentes Principales, Análisis de Correspondencias, Análisis de Correlaciones Canónicas y Análisis Canónico de Poblaciones.

⋆ Biplot de Gower: Gower propuso con un enfoque distinto al resto una variedad de Biplots basándose en la obtención de marcadores columna utilizando la regresión multivariante para ello $E(Y) = AB'$, utilizando métodos de escalamiento multidimensional para la obtención de la matriz A. Gower y Harding en 1988 y Gower en 1992 propusieron los llamados como Biplots no Lineales, los cuales se emplean para obtener dentro del ajuste de trayectorias no lineales para la representación de variables que posteriormente serán proyectadas sobre representaciones que hayan sido obtenidas mediante coordenadas principales. Gower y Hand definirían en 1996 los Biplots de interpolación y Predicción. En cuanto a los de interpolación se pueden suponer nuevos individuos proyectándolos sobre el subespacio de la representación, en cambio para los de predicción existe la posibilidad de inferir valores de las variables originales dado un punto sobre representación con la dimensión reducida. Con todo esto se logró demostrar que la interpretación de los Biplots clásicos, a la hora de hablar en términos del producto escalar, se relacionan con los Biplots de Predición.

⋆ Biplot para Minería de Datos: en el 2003 Vairinhos entraría en el ámbito de la Minería de Datos, para ello propuso a los Biplots como una base muy buena para descubrir patrones a la hora de clasificar grandes conjuntos de datos. Nos propuso un procedimiento con base en la formulación matemática y que nos permite realizar aproximaciones en conjuntos tanto de individuos como de observaciones a través de su representación por gráficos de intersección.

⋆ Biplot para detectar multicolinealidad: Ramírez en 2005 consideró la posibilidad existente que nos brinda un Biplot a la hora de visualizar las relaciones entre las variables. Para ello propuso un método basado en la descomposición espectral de la inversa de la matriz de correlaciones, con esto buscaba demostrar que el coeficiente de inflación de varianza de cualquier variable es igual al producto de los distintos marcadores columna de un Biplot, además de que la correlación parcial para dos variables es la misma, pero hay que exceptuar el signo, al coseno entre los marcadores columna que le correspondan.

2.2 Bases de datos y preprocesamiento

Para la obtención de las distintas bases de datos utilizadas en este trabajo se utilizaron dos tipos de fuentes distintas, una para las variables sociosanitarias y otra para las variables socioeconómicas. La Junta de Castilla y León posee una página online de datos abiertos (https://datosabiertos.jcyl.es), desde la que se descargaron las bases de datos relacionadas con variables sociosanitarias. El resto de datos también se obtuvieron de forma online, pero esta vez desde el INE (Instituto Nacional de Estadística), cuya página online es la siguiente: (https://www.ine.es), desde ella se descargaron las bases de datos relacionadas con variables socieconómicas.

Además de estos dos conjuntos de bases, se ha descargado el mapa de centros de salud de Castilla y León en formato GEOJSON para su uso en programas de sistemas de información geográficas (SIG). Ha esta base de datos con información espaciales se le ha asociado la información

que fue obtenida de las bases de datos sociosanitarias y socioeconómicas, para así poder realizar el análisis descriptivo a partir de mapas de las zonas básicas de salud.

Para cada conjunto de bases de datos se indica la descripción general (metadata) que proporcionan las páginas web, la relación de variables que contiene cada una de ellas, una pequeña indicación descriptiva de la base de datos descargada y la información obtenida despues de su preprocesamiento.

2.2.1 Bases de datos con información espacial

Con objeto de realizar un análisis espacial se ha decidido analizar los datos sociosanitarios agrupados por las zonas básicas de salud como unidad fundamental del sistema sanitario de C.A. de Castilla y León. La información se ha extraído de la base de datos de Centros de Salud obtenida de la página de datos abiertos de la Junta de Castilla y León.

2.2.1.1 Centros de Salud

- *Relación de los Centros de Salud de Castilla y León y los municipios a los que se encuentran asociados. La información se presenta agrupada en varias columnas cuyos nombres son los siguientes:*

  NOMBRE GERENCIA, CÓDIGO ZONA, NOMBRE ZONA, NOMBRE CENTRO SALUD, MUNICIPIO, PAC

La base de datos descargada y utilizada consta de un total de 2317 registros, a partir de ella se han generado dos bases de datos:

– zonas básicas de salud de Castilla y León con el código (ID) y el nombre (NAME). 247 registros
– Municipios de Castilla y León con el código de la zonas básicas de salud (ID) y el nombre del municipio (MUNICIPIO). 2317 registros.

  Esta base de datos se utilizará posteriormente para relacionar las zonas básicas de salud con los datos socioeconómicos que se descargaron anteriormente del INE y los cuales están agrupados por municipio.

2.2.2 Bases de datos sociosanitarias

Para las variables sociosanitarias se pueden escoger muchas bases de datos que esten relacionadas con Covid–19 . Pero para la consecución de los objetivos de este estudio se ha seleccionado, descargado y preprocesado las siguientes (se dará una breve explicación de cada una, pero se explicará más en detalle en cada apartado):

⋆ Población de referencia: en estadística, la población es el nombre que recibe un grupo de individuos que poseen una característica en común, esta característica es aquella sobre la que se realiza las mediciones pertinentes. Por ello, podemos definir la población de referencia como el conjunto de individuos donde se incluya el objeto a estudio (se incluya al individuo) y este en relación a lo que se esté midiendo. Por ejemplo, si queremos conocer el rendimiento

de un alumno en la asignatura de estadística deberemos comporarlo con sus compañeros de esa misma asignatura.

⋆ Mortalidad: podemos definir la tasa de mortalidad como la proporción de muertes que hayan sido registradas, respecto al número total de individuos que haya en una población (ciudad, país, C.A., etc) en un periodo de tiempo, el cual suele ser de un año.

⋆ Enfermos: se puede definir la tasa de enfermos como una medida que se utiliza para medir la frecuencia de aparición de nuevos casos de un tipo de enfermedad en una población que haya sido definida previamente durante un periodo de tiempo específico. También se suele utilizar medidas para saber los enfermos acumulados en cierto periodo de tiempo, es decir, el número total de enfermos en el periodo de tiempo elegido, ya sean 7 días, 3 meses, etc.

⋆ Pruebas diagnósticas: según el Instituto Nacional del Cáncer es cualquier tipo de prueba que pueda ser utilizada para ayudar en el diagnóstico de una enfermedad o afección teniendo en cuenta los síntomas y signos que sean presentados por el individuo. Estas pruebas diagnóstico también pueden ser empleadas para el diseño de un tratamiento, determinar la eficacia del mismo y realizar el pronóstico. Hay una gran cantidad de pruebas diagnóstico como las que definiremos más tarde y que son conocidas como PCR, la endoscopia o la biopsia.

Según el Instituto Nacional del Cáncer, la conocida como PCR es una técnica de laboratorio que se emplea para realizar diversas copias de un trozo determinado de ADN a partir de alguna muestra que contenga cantidades de pequeño tamaño del mismo ADN. Mediante el uso de la PCR podemos lograr multiplicar ese trozo de ADN para que sea posible su detección. La PCR también puede ser usada para identificar algunos cambios en un gen o cromosoma que ayuden a identificar y diagnosticar una enfermedad o una afección genética.

⋆ Prevalencia de la enfermedad: según el Instituto Nacional del Cáncer de España, la prevalencia en el campo de Medicina se puede definir como una medida de la cantidad total de individuos en un conjunto específico que padecen o han padecido una cierta enfermedad, afección o factor de riesgo en un momento o periodo determinado.

2.2.2.1 Población de referencia

- *Todos los individuos incluidos en esta base de datos son aquellos ciudadanos con derecho a la asistencia sanitaria en el Servicio de Salud de Castilla y León, además de que deben poseer su residencia habitual en dicha Comunidad Autónoma. Cada registro representa una Tarjeta Sanitaria Individual. La información se presenta en columnas cuyos nombres son los siguientes:*

  PERIODO, PROVINCIA, AREA, AMBITO DE PROCEDENCIA, ZONA BASICA DE SALUD (CODIGO), ZONA BASICA DE SALUD, EDAD, SEXO

Se han descargado dos bases de datos una del año 2020 con 2.311.958 registros.

Estas bases de datos se han procesado (Anexo A.1 pág. 2) con el objetivo de conseguir una información sintética de las zonas básicas de salud de Castilla y León, por ello se han descartado 30.167 registros sin información de zona básica de salud. La base de datos final contiene 247 registros y 10 variables:

- ID : código de la zonas básicas de salud
- NAME : nombre de la zonas básicas de salud
- AREA : nombre del área de salud donde se localiza la zonas básicas de salud
- PROVINCIA : provincia de localización de la zonas básicas de salud
- AMBITO : tipo de localización (Rural o Urbano)
- HOMBRE : número agregado de hombres
- MUJER : número agregado de mujeres
- E0-17 : número agregado de personas menores de 18 años
- E18-64 : número agregado de personas entre 18 y 64 años (ambos incluidos)
- E65-00 : número agregado de personas mayores de 64 años

2.2.2.2 Tasa de mortalidad

- *Los individuos que forman parte de esta base de datos son el número de personas fallecidas tras diagnóstico de* COVID–19 *(confirmados y compatibles con la enfermedad) por número de tarjetas sanitarias en cada zona básica de salud. De aquí obtenemos los datos acumulados desde el 1 de marzo de 2020. La información se presenta en las siguientes columnas:*

  FECHA, GERENCIA, NOMBREGERENCIA, CS, CENTRO, FALLECIDOS, TASAX100, X_GEO, Y_GEO, ZBS_GEO, PROVINCIA, POSICIÓN, MUNICIPIO

Debido a que la base de datos incluye todos los fallecimientos a partir de 1 de marzo de 2020 se ha descargado una única base de datos con 274.417 registros.

Estas bases de datos se han procesado (Anexo A.1 pág. 2) con el objetivo de extraer para cada una de las zonas básicas de salud durante el año 2020, las variables:

- ID : código de la zonas básicas de salud
- FALLECIDOS : número de fallecimientos

2.2.2.3 Tasa de enfermos

- *Datos diarios en cada zona básica de salud. Incidencia diaria de pacientes enfermos y de porcentaje de personas enfermas por número de tarjetas sanitarias en cada zona básica de salud. También se indican los enfermos y los porcentajes para los últimos 7 y 14 días. La información se presenta agrupada en las columnas:*

  FECHA, GERENCIA, NOMBREGERENCIA, CS, CENTRO, TOTALENFERMEDAD, TSI, TASAX100, X_GEO, Y_GEO, ZBS_GEO, TOTALENFERMEDAD_7DIAS, TASAX100_7DIAS, TOTALENFERMEDAD_14DIAS, TASAX100_14DIAS, PROVINCIA, TIPO_CENTRO, PCR_REALIZADOS, TASAX100_PCR_REALIZADOS, PCR_POSITIVOS, TASAX10000_PCR_POSITIVOS, PCR_POSITIVOS_SINTOMAS, TASAX10000_PCR_POSITIVOS_SINTOMAS, PCR_POSITIVOS_SINTOMAS_7DIAS, TASAPCR_POSITIVOS_SINTOMASX10000_7DIAS, PCR_POSITIVOS_SINTOMAS_14DIAS, TASAPCR_POSITIVOS_SINTOMASX10000_14DIAS, SOSPECHA_TRANSMISION_COMUNITARIA, POSICIÓN, MUNICIPIO

Debido a que la base de datos incluye todos los datos diarios de enfermos, datos acumulados en 7 y 14 días y de pruebas diagnósticas se ha descargado una única base de datos con 271.206 registros.

Estas bases de datos se han procesado (Anexo A.1 pág. 2) con el objetivo de extraer las siguientes variables para cada una de las zonas básicas de salud durante el año 2020:

- ID : código de la zonas básicas de salud
- ENFERMOS : número diario de enfermos
- ENFERMOS7 : número acumulado de enfermos en los últimos 7 días
- ENFERMOS14 : número acumulado de enfermos en los últimos 14 días
- PDIA : número de pruebas diagnósticas realizadas (PCR y otros)
- PDIA+ : número de pruebas diagnósticas realizadas con resultado positivo

2.2.2.4 Prevalencia de la enfermedad

- *El dato de prevalencia nos indica las personas que cada día siguen siendo compatibles con COVID-19. Datos diarios en cada zona básica de salud. Desde el 9 de junio de 2021 la actualización paso a ser semanal. La información se presenta agrupada en las columnas:*

  FECHA, GERENCIA, NOMBREGERENCIA, CS, CENTRO, PREVALENCIA, TASAX100, X_GEO, Y_GEO, ZBS_GEO, PROVINCIA

Debido a que la base de datos incluye los datos de prevalencia de la enfermedad se ha descargado una única base de datos con 271.206 registros.

Estas bases de datos se han procesado (Anexo A.1 pág. 3) con el objetivo de extraer las variables para cada una de las zonas básicas de salud durante el año 2020:

- ID : código de la zonas básicas de salud
- PREVALENCIA : personas que diariamente permanecen con COVID–19

2.2.3 Bases de datos socioeconómicas

Las conocidas como variables socioeconómicas son un grupo de medidas que pueden ser empleadas con el objetivo de describir y analizar distintos aspectos tanto de la sociedad como de la economía. Estas variables socioeconómicas generalmente guardan relación con características y situaciones del nivel socioeconómico de la materia que se esté estudiando. La información que suelen proporcionar suele estar relacionada con los siguientes: ingresos, educación, empleo, salud, vivienda y condiciones de vida, pero en este trabajo hablaremos sobre las que guardan relación con los ingresos, la renta y la desigualdad. Las variables socioeconómicas que se han tenido en cuenta son:

⋆ Ingresos: las que hacen referencia a este apartado son las que tienen que ver con la cantidad de dinero que es obtenido por una o un conjunto de personas bajo el mismo techo (que suele ser denominado como hogar) recibe de cualquier tipo de fuente, siendo el más común el que proviene de los salarios, pero también pueden venir de otros tipos de ingreso como los

beneficios sociales, beneficios obtenidos de inversiones, etc. Los ingresos generalmente son una de las medidas más importantes para evaluar el nivel económico y el bienestar material.

⋆ Renta: la renta es una medida que nos indica como es la falta de recursos económicos que serían necesarios para satisfacer las necesidades básicas de una o un conjunto de personas que vivan juntas (como dijimos anteriormente, es denominado hogar). Podríamos definir como una de sus principales utilidades el que se suele utilizar con el fin de analizar la distribución del nivel de riqueza y entre otras cosas, ser capaces de evaluar como son los programas de asistencia social que son utilizados para combatir la posible desigualdad económica que exista.

⋆ Desigualdad: la desigualdad tiene que ver a la disparidad de las distintas variables y factores, estan son algunas como pueden ser los ingresos y las oportunidades en nuestra sociedad. Para poder estudiarlas se suelen utilizar algunos índices y coeficientes como el de Gini, su principal objetivo es ayudar a evaluar el nivel de desigualdad en todos los casos posibles.

La información socioeconómica accesible en el INE se organiza por provincias por lo que se han descargado nueve bases de datos provinciales para obtener información de la C.A. de Castilla y León. Posteriormente se han unido en una única base de datos que se ha preprocesado para extraer la información socioeconómica a nivel municipal.

Las bases de datos socioeconómicas de Castilla y León que se han generado son las siguiente:

⋆ Indicadores demográficos: según el INE, los indicadores demográficos son aquellos que constituyen una colección de indicadores que son capaces de resumir la evolución en un periodo de tiempo del comportamiento de algún fenómeno demográfico (entre los que podemos incluir algunos como la mortalidad, natalidad, fecundidad, nupcialidad o cualquiera que tenga que ver con la estructura del lugar a estudiar entre otros) en un país, ciudad, C.A., etc.

⋆ Fuentes de ingresos: podemos definir las fuentes de ingresos como el lugar de procedencia de las ganancias de una persona o una empresa dentro de la actividad económica que estos realicen. Cabe destacar también el concepto de renta bruta, siendo esta el total de los ingresos de un individuo o empresa a lo que se le restará los costes directos en que se incurrió para su obtención.

⋆ Indicadores de renta: según el INE es quel indicador relativo capaz de medir desigualdad, cabe destacar que no mide pobreza de forma absoluta, ya que da un valor cuantificable de que número de individuos poseen un nivel de ingresos bajos en relación al conjunto de la población que se este estudiando.

⋆ Distribución de la renta: podemos definir la distribución de la renta como la manera en la que se realiza el reparto de los ingresos y rentas que hayan sido generados por los diferentes factores de producción de la econmía que se este estudiando. Uno de las formas más utilizadas para estudiar como es esta distribución es el índice de Gini, el cual es una medida de carácter económico que se emplea para poder calcular como es la desigualdad de los ingresos que hay entre los ciudadanos de un lugar o territorio, generalmente se utiliza para países, pero puede ser utilizado para cualquier unidad geográfica.

2.2.3.1 Indicadores demográficos

- *La información se presenta agrupada en las columnas:*

  MUNICIPIOS, DISTRITOS, SECCIONES, INDICADORES DEMOGRÁFICOS, PERIODO, TOTAL

Las 9 bases de datos provinciales descargadas se han agrupado en una autonómica de 343.392 registros.

Esta base de datos se ha procesado (Anexo A.1 pág. 3) para extraer los datos del 2020 (2.248 registros) y las variables:

– IDM : identificador de municipio
– POB_INE : número de habitantes
– EDAD_MEDIA : edad media de la población
– MENOR_18 : número de personas menores de 18 años
– ENTRE_18_64 : número de personas entre 18 y 64 años (ambos incluidos)
– MAYOR_64 : número de personas mayores de 64 años
– HOGAR_SIZE : tamaño medio del hogar
– HOGAR_UNIP : número de hogares unipersonales
– POB_SPAIN : número de habitantes españoles
– POB_NO_SP : número de habitantes no españoles

2.2.3.2 Distribución por fuente de ingresos

- *La información se presenta agrupada en las columnas:*

  MUNICIPIOS, DISTRITOS, SECCIONES, DISTRIBUCIÓN POR FUENTE DE INGRESOS, PERIODO, TOTAL

Las 9 bases de datos provinciales descargadas se han agrupado en una autonómica de 288.048 registros.

Esta base de datos se ha procesado (Anexo A.1 pág. 4) para extraer los datos del 2020 (1.521 registros) y las variables:

– IDM : identificador de municipio
– RENTA_BRUTA : renta bruta media por persona
– ING_SALARIO : ingresos por salario
– ING_PENSIONES : ingresos por pensiones
– ING_DESEMPLEO : ingresos por prestaciones de desempleo
– ING_OTRAS_PRESTACIONES : ingresos por otras prestaciones distantas a desempleo
– ING_OTROS : otras fuentes de ingreso distintas a salario, pensiones o prestaciones

2.2.3.3 Indicadores de renta

- *La información se presenta agrupada en las columnas:*

  MUNICIPIOS, DISTRITOS, SECCIONES, INDICADORES DE RENTA MEDIA Y MEDIANA, PERIODO, TOTAL

Las 9 bases de datos provinciales descargadas se han agrupado en una autonómica de 294.336 registros.

Esta base de datos se ha procesado (Anexo A.1 pág. 4) para extraer los datos del 2020 (2.248 registros) y las variables:

- IDM : identificador de municipio
- NETA_PERSONA : renta neta media por persona
- NETA_HOGAR : renta neta media por hogar
- MEDIA_UNIDAD : media de la renta por unidad de consumo
- MEDIANA_UNIDAD : mediana de la renta por unidad de consumo
- BRUTA_PERSONA : renta bruta media por persona
- BRUTA_HOGAR : renta bruta media por hogar

2.2.3.4 Distribución de la renta

- *La información se presenta agrupada en las columnas:*

  MUNICIPIOS, DISTRITOS, SECCIONES, ÍNDICE DE GINI Y DISTRIBUCIÓN DE LA RENTA P80/P20, PERIODO, TOTAL

Las 9 bases de datos provinciales descargadas se han agrupado en una autonómica de 98.112 registros.

Esta base de datos se ha procesado (Anexo A.1 pág. 5) para extraer los datos del 2020 (1.521 registros) y las variables:

- IDM : identificador de municipio
- P80.P20 : distribución de la renta como división entre percentil 80 y 20
- GINI : índice de Gini

2.3 Resultados y discusión

Una vez preprocesadas las bases de datos se seleccionaron 13 variables sociosanitarias en 247 zonas básicas de salud y 20 variables socioeconómicas en 2.248 municipios. Obtenidas ambas bases de datos se procede a asignar la información socioeconómica que está a nivel de municipios a zonas básicas de salud que es la unidad utilizada en las variables sociosanitarias.

La información de la base de datos Centros de Salud incluye un campo MUNICIPIO donde se relacionan los municipos que estan atendidos por la zona básica de salud. Los municipios con mucha población (como las capitales de provincia) se dividen en varias zonas básicas de salud. En este caso, los datos socioeconómicos del municipio se han incorporado a cada una de las zonas básicas de salud.

Sin embargo, lo más frecuente es que una zona básica de salud incluya varios municipios. En este caso se han agregado los datos socioeconómicos de varios municipios en la zona básica de salud siguiendo dos procedimientos:

- ⋆ Las variables del INE que representan valores absolutos como POB_INE, MENOR_18, ENTRE_18_64, MAYOR_64, POB_SPAIN, POB_NO_SP, RENTA_BRUTA, ING_SALARIO, ING_PENSIONES, ING_DESEMPLEO, ING_OTRAS_PRESTACIONES, ING_OTROS, se han sumado los valores de los municipios para obtener el valor de la zona básica de salud.

- ⋆ La variables del INE que representan valores relativos como EDAD_MEDIA, RENTA_BRUTA, NETA_PERSONA, NETA_HOGAR, MEDIA_UNIDAD, BRUTA_HOGAR, GINI, P80.P20, se ha calculado la media ponderada al valor de POB_INE para obtener el valor de la zona básica de salud.

En la agregación de estas dos bases de datos se han producido algunos desajustes, p.ej. algunos nombres de municipios de la base de datos del INE no coincidian con los de JCyL, hubo que buscar su equivalencia aunque unos municipos (Corrales y San Ildefonso) no fue posible enlazarla con datos sanitarios y una zona básica de salud no incluye datos municipales del INE.

Por último, indicar que hay un tratamiento distinto de algunas variables entre ambas bases de datos. En concreto con el tamaño de la población y su correspondiente división en clases de edad. Así, la población extraída de la C.A. de Castilla y León (POBLACION) incluye 2.311.958 personas de las cuales 30.167 no estan incluidas en ninguna zona básica de salud por lo que el valor final utilizado ha sido de 2.2881.791 personas a la que se denominará población sanitaria. Por el contrario, la población del INE (POB_INE y se denominará población total) incluye a 2.361.508 personas de las cuales 67.437 son de nacionalidad no española y, posiblemente, no todos tengan tarjeta sanitaria en C.A. de Castilla y León y sea el motivo de esta diferencia de población entre ambas bases de datos y los desajustes observados en las clases de edad (E0.17, E18.664, E65.00 de JCyL respecto a MENOR_18, ENTRE_18_64, MAYOR_64 del INE).

2.3.1 Análisis descriptivo y distribución espacial

Las estadísticas descriptivas básicas de las variables sociosanitarias y socioeconómicas por zonas básicas de salud se presentan en la Tabla 2.1.

Los mapas donde se muestra la distribución espacial de todas las variables en C.A. de Castilla y León y su descripción se presentan en el Anexo A.3. Para la visualización de la información se clasificaron los valores en cuatro cuartiles como se muestra en la leyenda, además, en la leyenda se indica el número de zonas básicas de salud incluidas en cada cuartil; indicar que una de las zonas básicas de salud no se pudo cruzar con la información socioeconómica por lo que aparece como 'undefined'.

Los mapas de algunas variables son importantes para el posterior análisis estadístico y se comentan a continuación.

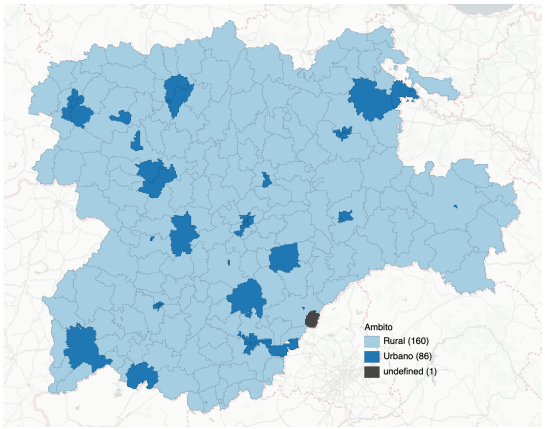
La Figura 2.1 muestra los mapas de tipo de ámbito de las zonas básicas de salud y la edad media de la población. En cuanto al ámbito de las zonas básicas de salud se observa que un total del 65% (160) se encuentran en áreas rurales mientras que el 35% restante (86) se encuentran en zonas urbanas. La edad media de la población de C.A. de Castilla y León es de 49.6 años pero irregularmente repartida. Aparecen un número mayor de zonas básicas de salud con edades superiores a 53 años en la parte oeste de la región mientras que en las zonas urbanas y en el este de la región, especialmente en el entorno de Burgos, Segovia y Valladolid se presentan más zonas básicas de salud con edades medias inferiores a 48 años.

Tabla 2.1: Estadísticas de las variables sociosanitarias y socioeconómicas utilizadas en este trabajo

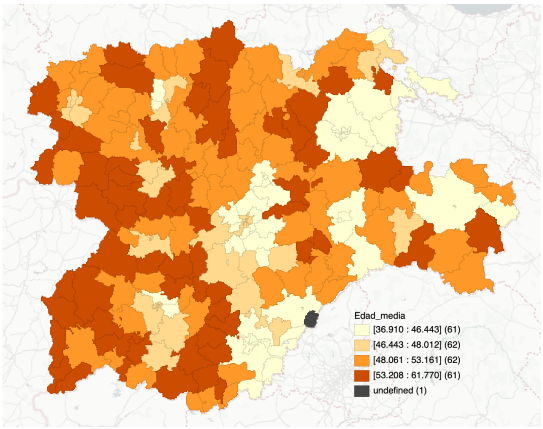
Nombre	Mínimo	Q1	Mediana	Media	Q3	Máximo
Sociosanitarias						
Poblacion	462	2906	7592	9255	14755	35471
Hombre	247	1532	3757	4534	7126	16574
Mujer	215	1393	3676	4721	7565	18897
E0.17	24	224	735.5	1210.3	2028.2	5469
E18.64	244	1537	4472	5697	8990	23291
E65.00	150	974	2024	2348	3495	7880
Fallecidos	0	10	26	32.57	44	167
Enfermos	12	210.8	626.5	836.7	1339	3411
Enfermos7	84	1447	4294	5811	9334	23766
Enfermos14	168	2880	8498	11550	18579	47305
PDIA	99	865.2	2188	2825.3	4067.8	13266
PDIA.	10	134.8	377.5	501	765.8	2086
Prevalencia	0.51	20.18	65.37	94.05	132.97	497.13
Socioeconómicas						
Pob_ine	661	3648	9641	56520	65757	304496
Menor_18	30	301.5	1270	8371.9	10343	44917
Entre_18_64	368	2037	5850	33646	40268	180085
Mayor_64	221	1170	2520	14502	15918	79494
Pob_spain	643	3613	9282	54180	65258	287681
Pob_no_sp	0	0	0	2340.3	530.2	16815
Renta_bruta	12942	17395	71122	94107	132106	621587
Ing_salario	7244	10149	32154	44981	59357	364933
Ing_pensiones	2628	4695	21918	28294	40024	132667
Ing_desempleo	526	653	2726	3474	4916	23806
Ing_otras_prestaciones	464	666	3079	4041	5908	17918
Ing_otros	729	1705	9343	13301	19858	83278
Edad_media	36.91	46.5	48.04	49.58	53.16	61.77
Renta_bruta_media	10470	13314	14761	14720	15949	18290
Neta_persona	9320	11666	12675	12611	13484	15135
Neta_hogar	20521	25428	28831	28482	31197	39504
Media_unidad	13348	16478	18057	18031	19538	22769
Bruta_hogar	22640	29311	33659	33410	37121	48823
Gini	24.4	28.09	29.28	29.45	30.9	34.62
P80.P20	1.96	2.37	2.49	2.49	2.60	2.98

La media de personas enfermas por zona básica de salud es de 836.7 personas, valor claramente superior la media de 32.6 personas fallecidas, lo que supone una tasa de mortatildad de 3.89%. Su distribución tampoco es uniforme a lo largo de la C.A. de Castilla y León.

La Figura 2.2 muestra dos mapas importantes relacionados con el Covid–19 en C.A. de Castilla y León, el número de enfermos y el número de fallecimientos por cada zona básica de salud. Se aprecia en estos mapas que no hay una coincidenca importante entre el número de enfermos



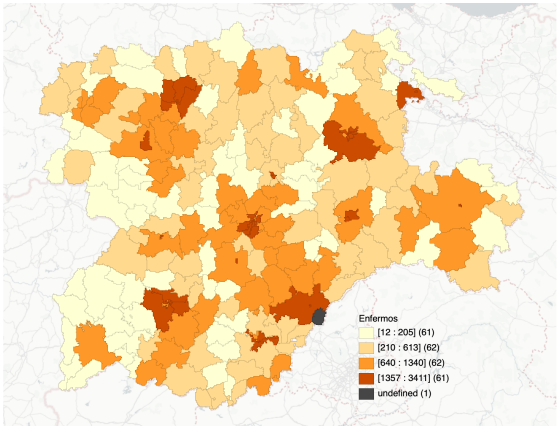
Ámbito de localización



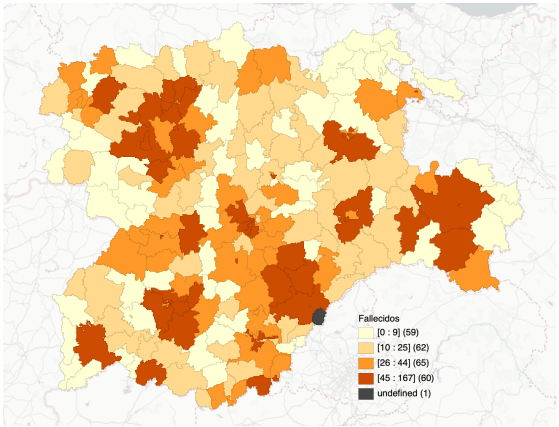
Edad media de la población

Figura 2.1: Distribución de la zonas básicas de salud en función del ámbito de ubicación (Rural o Urbano) y el número de fallecimientos (derecha)

y el número de fallecimientos. De forma general, el último cuartil del número de enfermos de Covid–19 se presenta alrededor de los grandes núcleos de población con una distribución parecida al mapa de la Figura 2.1 de ámbitos de las zonas básicas de salud. Sin embargo, en la distribución espacial de fallecidos no esta tan clara esta relación.



Número de enfermos



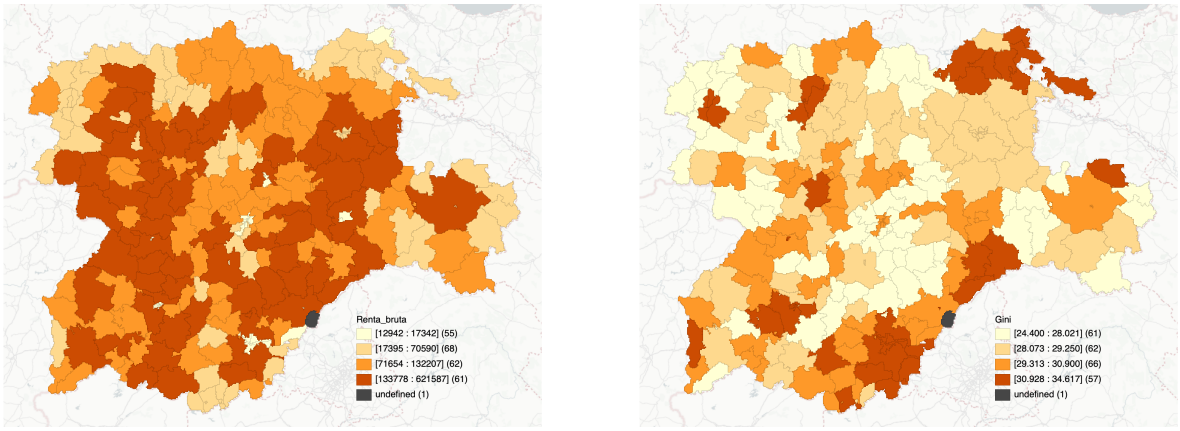
Número de fallecimientos

Figura 2.2: Distribución de la zonas básicas de salud en función de dos variables sanitarias relacionadas con el Covid–19

En la Figura 2.3 se presentan dos mapas relacionados con dos variables económicas, la renta bruta por persona y el índice de Gini. Sobre la renta bruta indicar que el valor medio en las zonas básicas de salud de C.A. de Castilla y León es de 94.107 euros (renta bruta media por persona es de 14.720 euros) con valores superiores a 133.778 euros (Q3) distribuidos por todas las provincias de la C.A. de Castilla y León. Lo destacable en este mapa es la presencia de algunas zonas básicas de salud con rentas brutas inferior a 70.590 euros (Q1 y Q2), sobre todo en zonas montañosas de la región, muy inferior al valor máximo de 621.587 euros.

En cuanto al índice Gini se observa más desigualdad en el entorno a los principales núcleos urbanos, especialmente en Ávila, Segovia y norte de Burgos. Por el contrario las zonas de salud

de las zonas rurales del centro de Castilla, noroeste de Zamora y El Bierzo presentan mayores valores de igualdad económica.



Renta bruta Índice de Gini

Figura 2.3: Distribución de la zonas básicas de salud en función de dos variables económicas

2.3.2 Análisis de Correlaciones

Para obtener los resultados del análisis de correlaciones se utiliza el programa R y el paquete corrplot (Wei & Simko, 2021). Para reordenar la matriz de correlación se ha utilizado la opción hclust con el método de agregación ward.D que permite un agrupamiento jerárquico y facilita la detección de grupos de variables fuertemente correlacionadas.

Empezaremos primero analizando las variables sociosanitarias, siguiendo con las socioeconómicas y acabando con el análisis de ambas a la vez.

En la matriz de correlaciones de las variables sociosanitarias (Figura 2.4) se observa que los fallecidos no guarda apenas relación con ninguna de las otras variables. En cuanto al resto de variables podemos decir que se encuentran bastante correlacionadas, siendo la prevalencia la que menos correlación guarda con el resto, pero a pesar de ello está bastante correlacionado.

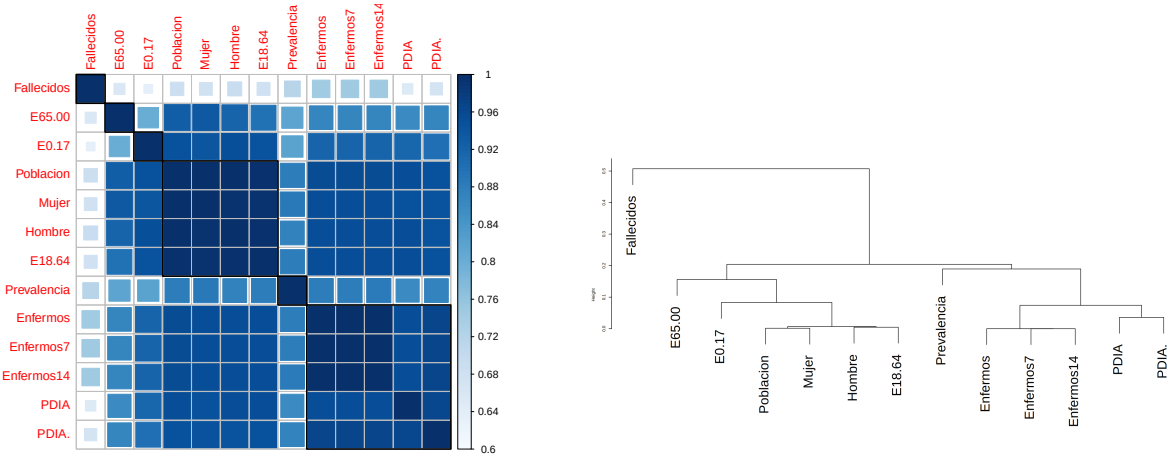


Figura 2.4: Matriz de correlaciones de las variables sociosanitarias y dendrograma utilizado para la ordenación de las variables

Se seleccionan con variables por formar parte de un mismo grupo de variables (muy alta correlación entre ellas)

- E18-64: grupo E18-64, Hombre, Mujer, Población
- Enfermos: grupo Enfermos, Enfermos7, Enfermos14

Siguiendo con el dendrograma observamos que hay tres grupos distinos:

- Primer grupo: formado únicamente por los fallecidos, además como ya vimos en la matriz de correlaciones, esta variable no estaba relacionada con ninguna más.

- Segundo grupo: conformado por las siguientes variables: E0.17, E65.00, población, mujer, hombre y E18.64. Si nos volvemos a fijar en la matriz de correlaciones observamos una zona de correlaciones más cercanas a 1 antes de la variable de prevalencia en la cual se encuentran todas estas variables.

- Tercer grupo: conformado por las siguientes variables: prevalencia, enfermos, enfermos7, enfermos14, PDIA y PDIA_. Observamos que aunque la prevalencia esté en este grupo, no está tan relacionada como el resto. Nuevamente en la matriz de correlaciones volvemos a observar una zona de correlaciones muy cercanas a 1 entre las variables de este grupo (excluyendo a la prevalencia), además se podría hablar también de 2 subgrupos, uno conformado por las 3 variables de los enfermos y otro por las 2 variables restantes.

Para los posteriores análisis valoraremos la eliminación de algunas variables debido a su alta correlación con otras variables para así conseguir un mejor resultado a la hora de analizarlos.

- E18.64: una variable que forma parte del segundo grupo, y que creaba un subgrupo con las siguientes variables: Hombre, Mujer y Población

- Enfermos: una variable que forma parte del tercer grupo y que creaba un subgrupo con las siguientes variables: Enfermos7 y Enfermos14.

En la matriz de correlaciones de las variables socioeconómicas (Figura 2.5) se observa que las variables crean grupos con las variables con las que más se correlacionan, donde observamos que en un principio parece que observamos 4 grupos distintos, donde los dos primeros correlacionan algo entre ellos y los 2 últimos se relacionan mejor entre ellos que con el resto. Cabe destacar que el primer grupo sería simplemente la edad media, lo veremos mejor con el dendrograma.

Para el dendrograma observamos que hay 4 grupos bien diferenciamos como bien predecimos, los cuales serían:

- El primer grupo estaría compuesto por simplemente la edad media, esta se relacionan medianamente bien con las variables del segundo grupo y se relaciona escasamente con el resto, aunque se relaciona inversamente bien con algunas de los del tercer grupo.

- El segundo grupo estaría constituido por las siguientes variables: ingresos de las pensiones, ingresos de otras prestaciones, otros ingresos, ingresos de desempleo, la renta bruta y los ingresos de los salarios. Todas ellas tienen muy buenas correlaciones y muy escasas con las de los grupos 3 y 4.

- El tercer grupo estaría constituido por las siguientes variables: neta del hogar, bruta del hogar, neta por persona, renta bruta media y media unidad. Se relacionan muy bien entre ellas y

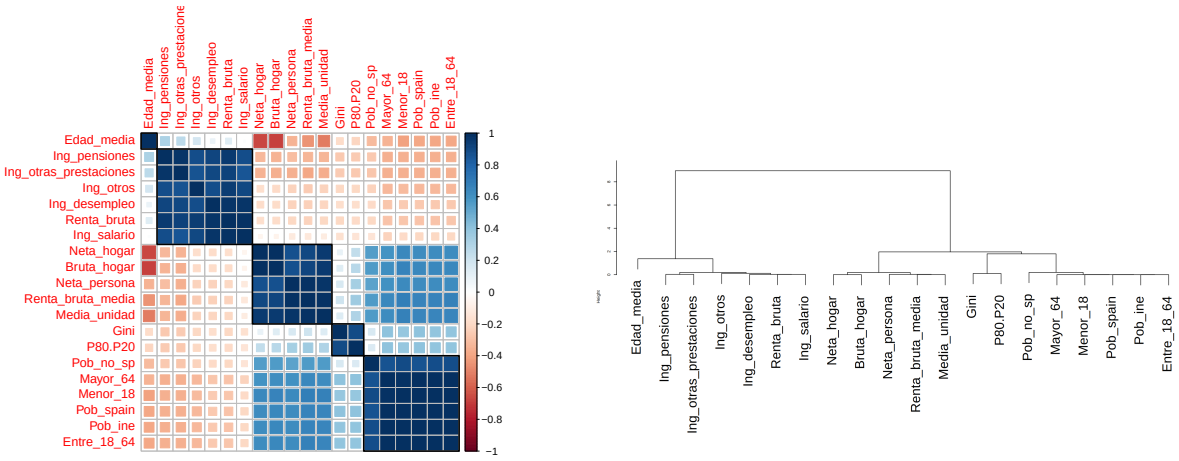


Figura 2.5: Matriz de correlaciones de las variables socioeconómicas y dendrograma utilizado para la ordenación de las variables

medianamente bien con las del cuarto grupo, con el resto tienen una correlación bastante pobre.

⋆ El cuarto grupo estaría constituido por las siguientes variables, aunque cabe destacar que podríamos haber hablado de 5 grupos ya que las dos primeras forman un subgrupo que se podría considerar como grupo aparte: índice de Gini, P80.P20, población no española, mayores de 64 años, mayores de 18 años, población española, población según el INE y individuos entre 18 y 64 años. Ocurre lo mismo descrito anteriormente, se relacionan muy bien entre ellas y medianamente bien con el grupo 3, pero tienen una relación escasa con los grupos 1 y 2.

En la matriz de correlaciones de las variables socioeconómicas y sociosanitarias (Figura 2.6) se observa que los dos primeros grupos que se habían creado en las socioeconómicas se mantienen, y con la adición de las sociosanitarias; éstas han creado un grupo propio, relacionándose de forma similar a la descrita anteriormente, se verá mejor a la hora de explicar el dendrograma.

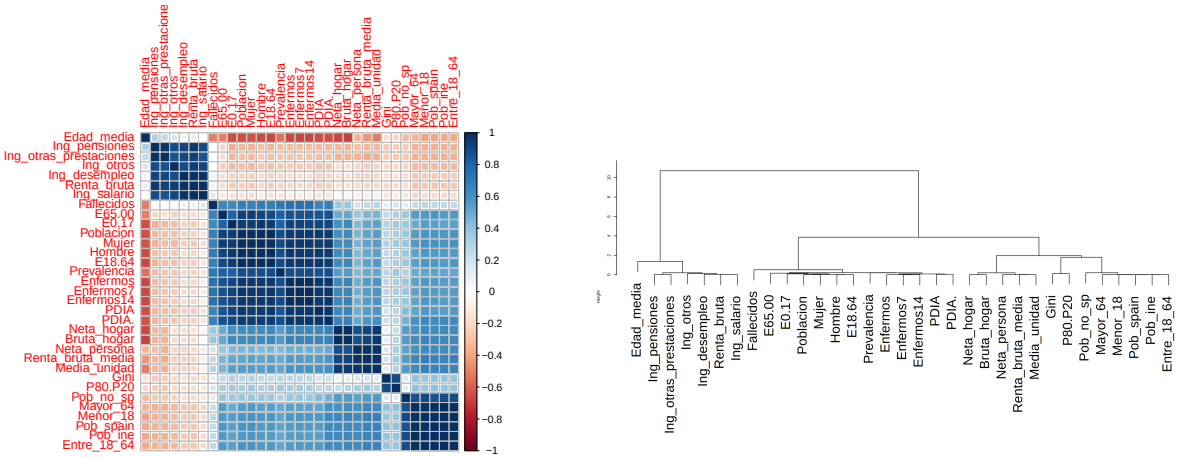


Figura 2.6: Matriz de correlaciones de las todas variables estudiadas y dendrograma utilizado para la ordenación de las variables

Hablando del dendrograma vemos que hay 4 grandes grupos:

- ⋆ El primer grupo nuevamente está formado solo por la edad media, donde esta se relaciona medianamente bien con el segundo grupo y de forma escasa con el resto, aunque se aprecian aún mas relaciones inversas bastante buenas con variables de los distintos grupos.

- ⋆ El segundo grupo seguiría formado por las mismas variables que en las socioeconómicas: ingresos de las pensiones, ingresos de otras prestaciones, otros ingresos, ingresos de desempleo, la renta bruta y los ingresos de los salarios. Se siguen relacionando bien entre ellas, pero con el resto sigue existiendo una relación mediocre.

- ⋆ El tercer grupo estaría formado por todas las variables sociosanitarias, donde se relacionan algo con las variables del grupo 4 y tienen unas relaciones malas con el resto de las variables.

- ⋆ El cuarto grupo estaría formado por las variables de los dos últimos grupos de las socioeconómicas, las cuales se siguen comportando de la misma forma, medianamente bien entre ellas, no del todo mal con el tercer grupo y de forma escasa con el resto de grupos.

2.3.3 Análisis de Componentes Principales

Para obtener los resultados del análisis de componentes principales se utiliza el programa R y los paquetes FactoMineR (Lê et al., 2008) y factoextra (Kassambara & Mundt, 2020).

Para el ACP primero se identificará el número de dimensiones con las que se analizará las variables mediante el Scree Plot, las varianzas explicadas acumuladas de las dimensiones y finalmente se hablará de a qué dimensión pertenecen las variables que han sido seleccionadas para el estudio según sus niveles de correlación con el resto de variables.

El Scree Plot (Figura 2.7) obtenido muestra que podríamos trabajar con 2 dimensiones, pero en el análisis se decide emplear 4 dimensiones para obtener unos resultados mejores según lo que buscamos con la realización de este trabajo.

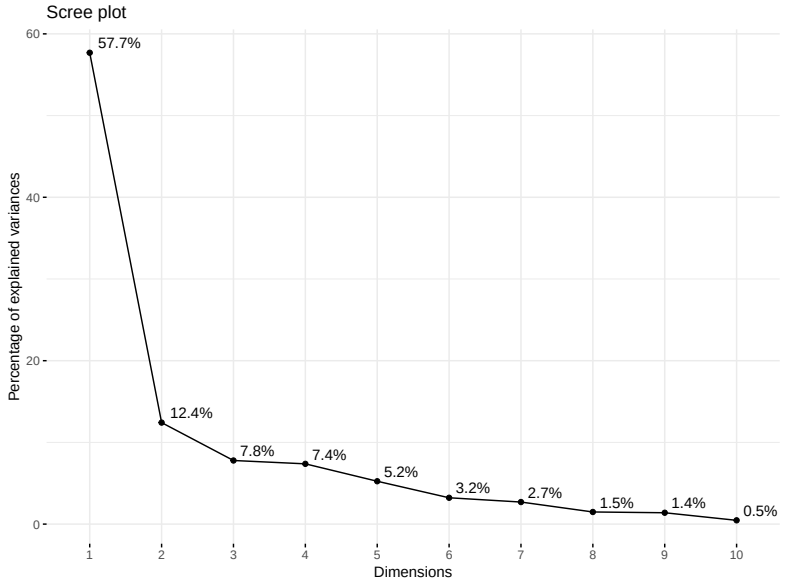


Figura 2.7: ACP: porcentaje de varianza explicada por cada una de las dimensiones

Los datos de varianza que explican las 4 primeras dimensiones se muestran en la Tabla 2.2, vemos que con las dos primeras dimensiones ya tenemos más de un 70% de la varianza explicada, con esto podríamos utilizar solo 2 dimensiones, pero se ha preferido utilizar 4 dimensiones (85.3% de la varianza) para el análisis de los datos.

Tabla 2.2: ACP: valores propios y porcentaje de la varianza explicado de las 4 primeras dimensiones

	Dimensión 1	**Dimensión 2**	**Dimensión 3**	**Dimensión 4**
Autovalor	6.346	1.367	0.856	0.811
Porcentaje de varianza	57.690	12.423	7.785	7.375
Porcentaje acumulado	57.690	70.113	77.898	85.274

Una vez decidido que se usarán 4 dimensiones, para estudiar las variables que pertenecen a cada dimensión tomaremos aquellas que tengan una correlación (Tabla 2.3) mayor de 0.4 en valor absoluto, con todo ello nuestras dimensiones estarían compuestas por:

Tabla 2.3: ACP: correlaciones de las variables con las 4 dimensiones

	Dimensión 1	**Dimensión 2**	**Dimensión 3**	**Dimensión 4**
Edad_media	-0.7001	-0.1213	-0.03857	0.06327
Renta_bruta	-0.2784	0.6427	0.18405	0.66868
Neta_persona	0.5592	-0.4105	0.60777	0.18726
Gini	0.3619	-0.4748	-0.62187	0.47773
Pob_ine	0.6715	-0.4856	0.21877	0.26574
E0.17	0.9294	0.0539	-0.06114	-0.08407
E18.64	0.9672	0.0922	-0.02221	-0.09471
E65.00	0.8929	0.1063	-0.04902	-0.02680
Fallecidos	0.7247	0.5088	-0.09046	0.07315
Prevalencia	0.9097	0.1060	-0.04645	-0.04789
Enfermos	0.9662	0.1261	-0.00786	-0.04280

- Primera dimensión: las variables que más se relacionan con la dimensión con una correlación positiva serían las siguientes (ordenadas de mayor carga a menor): enfermos, E18.64, E0.17, prevalencia, E65.00, fallecidos, población del INE y neta por persona. En cuanto a la variable con correlación negativa y que más se relaciona con la dimensión sería la siguiente: edad media.

- Segunda dimensión: las variables que más se relacionan con la dimensión y que tienen correlación positiva serían las siguientes (ordenadas de mayor carga a menor): renta bruta y fallecidos. En cuanto a las variables con correlación negativa y que más se relacionan con la dimensión serían las siguientes (ordenadas de mayor carga a menor): la población del INE, el índice de GINI y la neta por persona.

- Tercera dimensión: la variable que más se relaciona con la dimensión y que tiene correlación positiva sería la siguiente: neta por persona. En cuanto a la variable con correlación negativa y que más se relaciona con la dimensión sería la siguiente: índice de Gini

- Cuarta dimensión: las variables que más se relacionan con la dimensión y que tienen correlación positiva serían las siguientes (ordenadas de mayor carga a menor): renta bruta e

índice de Gini. En este caso no encontraríamos ninguna variable que se relacionase bien con la dimensión y aportase una correlación negativa significativa.

En la Figura 2.8 se observan de una forma gráfica como son las contribuciones de cada variable a las distintas dimensiones, donde el valor si todas contribuyesen de la misma forma viene representado por la línea roja (contribución media: $1/11$ variables $= 9.1\%$).

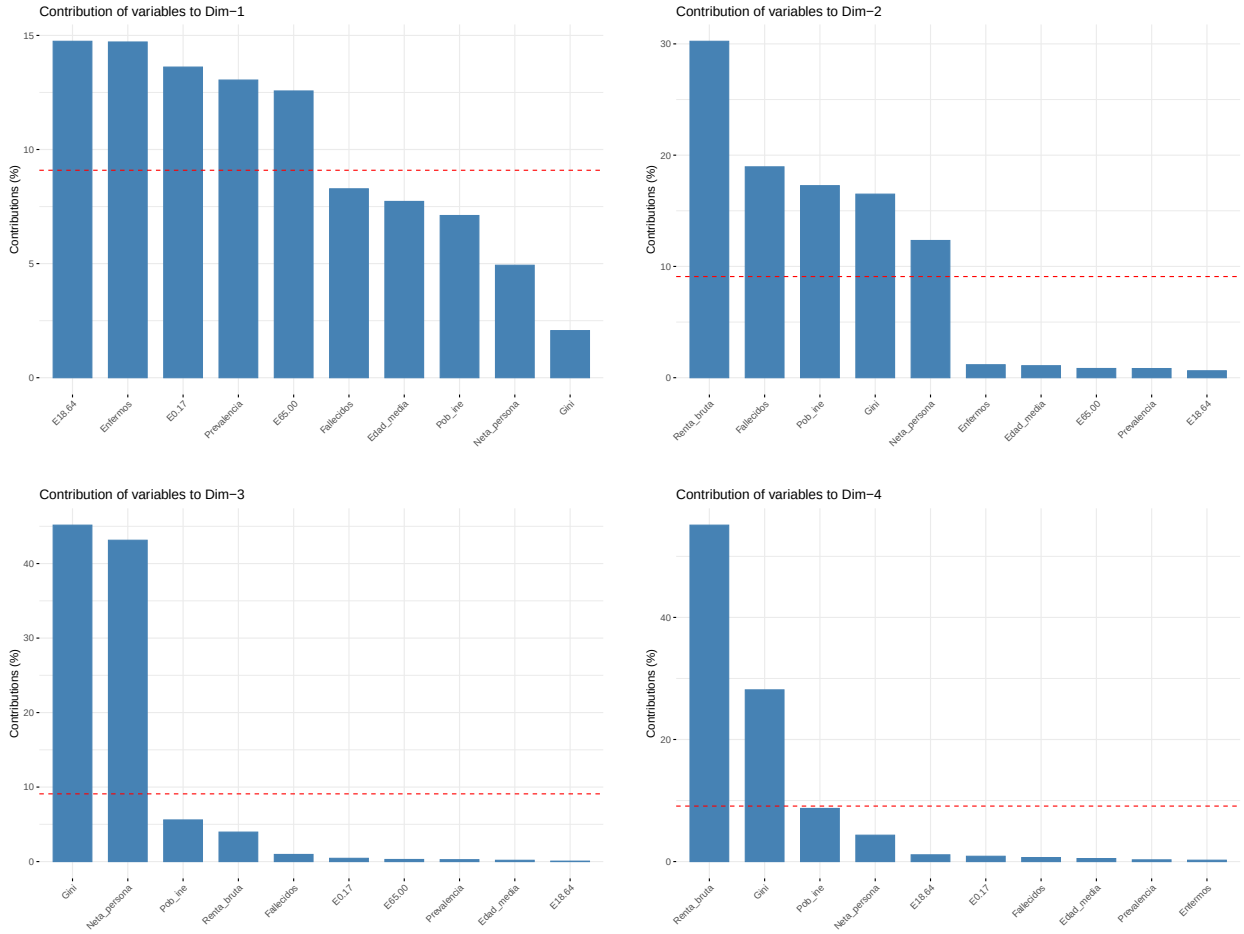


Figura 2.8: ACP: contribución de cada variable a cada una de las primeras 4 dimensiones

- ⋆ Primera dimensión: se observan 5 variables por encima de la contribución media, siendo estas las tres edades distintas, los enfermos y la prevalencia.
- ⋆ Segunda dimensión: se observan 5 variables nuevamente por encima de la contribución media, pero en este caso la diferencia con el resto de variables es mucho mayor. Las variables por encima de la media serían los fallecidos, la renta neta por persona, el índice de Gini, la renta bruta y la población del INE.
- ⋆ Tercera dimensión: se observan solo 2 variables por encima de la contribución media, donde vemos una clara diferencia con el resto, estas variables serían la renta neta por persona y el índice de Gini.
- ⋆ Cuarta dimensión: vemos que que hay 2 variables por encima de las contribuciones medias, aunque también podríamos llegar a considerar 3 si incluyesemos a la población del INE

que tiene valores muy cercanos, el resto se encuentran bastante alejados. Estas dos variables serían la renta bruta y el índice de Gini.

En la Figura 2.9 se presentan dos gráficos donde vienen represetandas las variables y como se relacionan con las dimensiones:

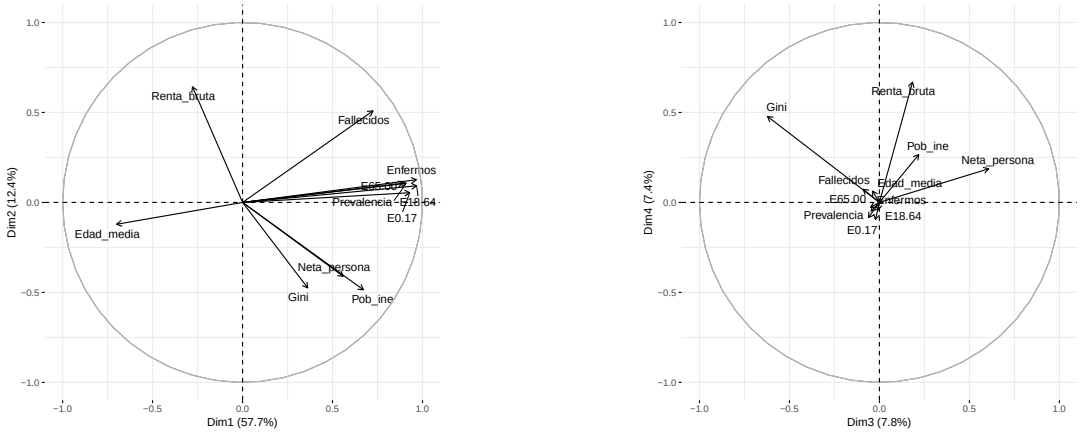


Figura 2.9: ACP: correlación de las variables con las 2 primeras dimensiones

⋆ Para el primer gráfico vemos que los que más se relacionan con la dimensión 2 son la renta bruta y los fallecidos, ambos de forma positiva, para la dimensión 1 vemos que hay un grupo de ellas formado por enfermos, las edades que se relacionan bien entre ellas y que además se relacionan bien de forma positiva con la dimensión 1, mientras que la población del INE, la renta neta por persona y el índice de Gini se relacionan medianamente bien entre ellas y se relacionan bien de forma negativa con la dimensión 1, aunque cabe destacar que el índice Gini es la que menos se relaciona en este caso.

⋆ Para el siguiente gráfico vemos que las únicas variables que se relacionan considerablemente bien con la dimensión 3 serían la renta neta por persona de forma postiva y el índice de Gini de forma negativa. El resto de ellas están agrupadas en el centro sin relacionarse bien ni con la dimensión 3 ni con la 4, aunque podríamos decir que la renta bruta si que se relaciona bien con la dimensión 4.

2.3.3.1 Biplot

En este apartado se trata sobre el Análisis Biplot del ACP, para ello se emplean 2 gráficos (Figura 2.10), el primero con las 2 primeras dimensiones (1 y 2) y el segundo con las otras 2 dimensiones (3 y 4).

En el biplot con las dimensiones 1 y 2 se observa la formación de dos grupos de zonas de salud, que se corresponde aproximadamente con el ámbito rural o urbano donde se encuentran.

En el primer gráfico observamos,

1. Un grupo de variables formado por las clases de edad, los enfermos y la prevalencia, relacionadas más con la dimensión 1 de forma positiva. Este agrupamiento sugiere que todas ellas están muy relacionadas entre si y se podría simplificar el análisis usando una de ellas.
2. Otro grupo formado por la renta neta por persona, el índice de Gini y la población total (INE), relacionadas más con la dimensión 2 de forma negativa pero con influencia en la dimensión

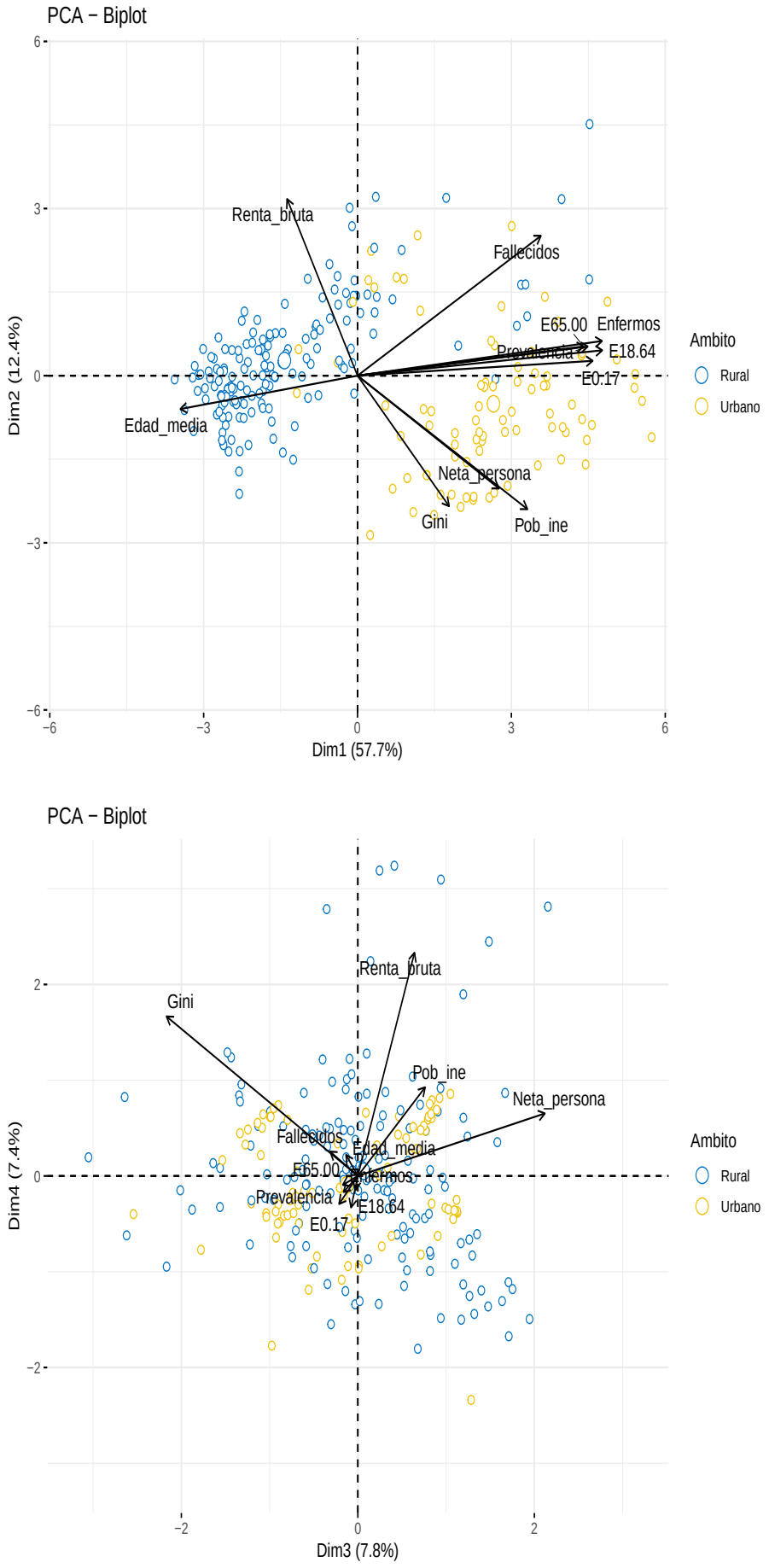


Figura 2.10: ACP: biplot usando las 4 primeras dimensiones

1. La relación entre renta neta y población es muy alta sugiriendo que se podría dejar una de ellas, mientras que el índice de Gini tiene relación pero con mayor influencia sobre la dimensión 2.
3. Variables sueltas como la renta bruta (con la dimensión 2), la edad media (con la dimensión 2) y fallecidos (con las dimensiones 1 y 2).

La dimensión 1 se relaciona de forma positiva con el primer grupo de variables lo que apunta a que esta dimensión es una dimensión sociosanitaria. La distribución de individuos muestra dos grupos que coinciden aproximadamente con el ámbito rural o urbano de su localización, los de zonas urbanas han presentado un mayor número de enfermos, de prevalencia y tienen una población sanitaria mayor en todas las clases de edad que los de zonas rurales. Las zonas rurales tienen una población con una edad media superior ya que esta dimensión se correlaciona de forma importante pero de forma negativa con ésta variable. El análisis biplot muestra que esta variable social es un buen indicador de la afección de Covid–19 en las zonas de salud de forma que ha afectado con mayor intensidad en las zonas con menor edad media que en las zonas de mayor edad media. Por lo tanto podemos afirmar que la edad tiene bastante relevancia junto con el nivel de afección del Covid–19 . Sin embargo este resultado contradice aparentemente a Herrer-Cartaya et al. (2022) que afirman que la edad constituye un factor de riesgo asociado a la gravedad en pacientes con la COVID-19 y podría explicarse por no haberse tenido en cuenta la situación socioeconómica de los pacientes.

La dimensión 2 se correlaciona con la renta bruta de forma positiva y con el segundo grupo de variables de forma negativa con lo que podemos decir que se trata de una dimensión socioeconómica. Donde tenemos que destacar que las desigualdades económicas, las rentas y la población tienen que ver con el nivel de afección del Covid–19 , sobre todo en zonas urbanas. Esta situación ha sido denunciada por diversas publicaciones como (Bohoslavsky et al., 2020).

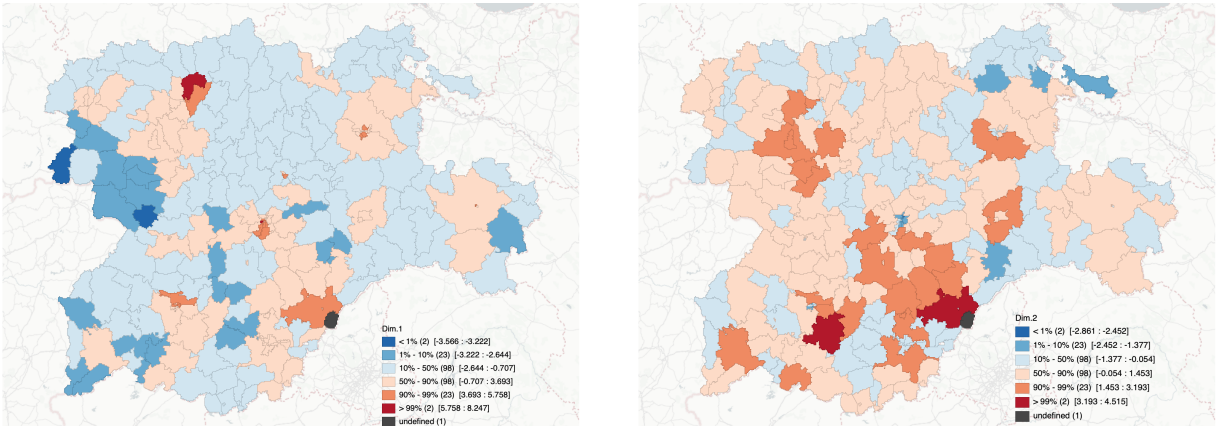
Destacar la variable fallecimientos que esta correlaciona de forma similar con las dimensiones 1 y 2. Además de relacionarse solo de forma negativa con la edad media.

En el segundo gráfico de la Figura 2.10 la mayor parte de las variables están en el centro y con una longitud del vector pequeña lo que apunta a que no tienen relación con las dimensiones 3 y 4. No obstante, el índice de Gini y la renta neta por persona si presentan una correlación importante con la dimensión 3 y la renta bruta con la dimensión 4. Ambas dimensiones son dimensiones socioeconómicas que no parece que tengan relación con variables sanitarias. Por lo tanto no serían de mayor interés para este trabajo que busca si existen variables socioeconómica que impacten en la distribución del Covid–19 .

2.3.3.2 Distribución espacial

Los biplot de la Figura 2.10 muestran la distribución en las dimensiones principales de las zonas básicas de salud, en este caso existe una variable (ámbito) que ayuda a la interpretación de los resultados pero cuando existe información espacial adicional, la elaboración de mapas es un complemento que mejora sensiblemente a la interpretación de los resulatdos del ACP.

En la Figura 2.11 se muestran los percentiles de los valores (scores) de las 2 primeras dimensiones.



Dimensión 1 Dimensión 2

Figura 2.11: ACP: distribución espacial de las 4 primeras dimensiones del ACP

La dimensión 1 es una dimensión sanitaria relacionada con la magnitud de la enfermedad. En el mapa con esta dimensión se observa que las zonas con un percentil superior al 90% se encuentran alrededor de las capitales de León, Segovia y un pequeño núcleo en Valladolid y Burgos; además de otra zona de salud del norte de Salamanca. No aparecen otras capitales como p.ej. Ávila, Palencia o Salamanca. Por el contrario, las zonas con percentiles inferiores al 10% se encuentran al noroeste de Zamora, zonas más despobladas y con mayor edad media.

La dimensión 2 es una dimensión económica. En el mapa correspondinte se observan dos núcleos alejados de los demás con un percentil superior al 99% y corresponde a las capitales de Segovia y Salamanca. Otros núcleos importantes con percentiles superiores al 90% se distribuyen por la C.A. de Castilla y León pero sin una relación clara con otras capitales de provincia. Por el contrario, las zonas con percentiles inferiores al 10% se presentan sobre todo en la provincia de Burgos.

Hay que tener en cuenta que ambas dimensiones son las que permiten explicar los fallecimientos por Covid–19 y los responsables de sanidad deberían tener en cuenta esta información para estudiar por qué se produce esta distribución espacial y no ha existido una relación, lógica de antemano, entre número de fallecimientos y de enfermos.

2.3.4 Análisis de Correlación Canónica

Para obtener los resultados del análisis de correlación canónica se utiliza el programa R y el paquete CCA (González & Déjean, 2022).

El análisis de correlación canónica se emplea para identificar la asociación entre dos grupos de datos multivariantes. Por este motivo, al tener un grupo de variables sociosanitarias y otras socioeconómicas, se decide en este trabajo explorar este tipo de análisis estadístico.

Cada dimensión de este análisis se refiere al conjunto de dos combinaciones lineales (variables canónicas), uno con las variables socisanitarias y otro con las variables socioeconómicas. En la Tabla 2.4 se muestran las correlaciones entre las dos variables canónicas en cada dimensión y vemos que para la primera dimensión tenemos una buena correlación de 0.78, pero la siguientes correlaciones son claramente inferiores, no tanto la segunda con un valor de 0.60, pero si la tercera y la cuarta dimensión. En cualquier caso se exploraran todas las dimensiones obtenidas.

Tabla 2.4: ACC: correlaciones canónicas de las primeras 4 dimensiones

	Dimensión 1	**Dimensión 2**	**Dimensión 3**	**Dimensión 4**
Correlaciones canónicas	0.781	0.509	0.310	0.214

Las correlaciones que obtenemos entre las variables de los dos conjuntos de datosº con las variables canónicas sociosanitarias y con las variables canónicas socioeconómicas se muestran en la Tabla 2.5.

En la dimensión 1, las variables sociosanitarias se relacionan de forma negativa y bastante bien con la primer variable canónica sociosanitaria, siendo los fallecidos la que peor se relaciona con un valor de -0.56 y la que mejor la clase de edad de entre 18 y 64 años con un valor de -0.96. Para las socioeonómicas vemos que las que mejor se relacionan son la edad media de forma positiva (0.68) y, en menor medida, pero de forma negativa la renta bruta y la población del INE. Como era de esperar, se relacionan mejor las sociosanitarias que las socioeconómicas con las variables canónicas sociosanitarias.

Tabla 2.5: ACC: correlaciones de las variables con las variables canónicas sociosanitarias

	Dimensión 1	**Dimensión 2**	**Dimensión 3**	**Dimensión 4**
Sociosanitarias				
E0.17	-0.9310	0.0349	0.1201	-0.1571
E18.64	-0.9601	0.0679	0.2244	0.0324
E65.00	-0.7944	-0.0885	0.5925	-0.0032
Fallecidos	-0.5593	0.5459	0.3937	-0.3880
Prevalencia	-0.8356	-0.0650	0.1694	-0.2191
Enfermos	-0.9531	0.0948	0.1934	-0.1932
Socioeconómicas				
Edad_media	0.6757	-0.2428	0.0123	-0.0081
Renta_bruta	0.3041	0.3016	0.1245	-0.0976
Neta_persona	-0.5171	-0.1557	-0.0699	-0.0521
Gini	-0.2815	-0.1876	0.0359	-0.1397
Pob_ine	-0.5740	-0.2234	0.1505	-0.0031

En la dimensión 1 de las variables canónicas socioeconómicas, las variables sociosanitarias se relacionan medianamente bien todas de forma negativa, aunque hay que tener en cuenta que los fallecidos presentan un valor de -0.43. Las que mejor se correlacionan serían nuevamente la clase de edad de entre 18 y 64 (-0.75) y los enfermos (-0.74). Por el contrario, la que presenta menor correlación es la clase de edad de mayores de 65 años con un valor de -0.62.

Para las propias variables socioeconómicas, vemos que la edad media se relaciona muy bien de forma positiva con un valor de 0.87, mientras que la renta bruta y la población del INE se relacionan bien también pero de forma negativa con valores de -0.66 y -0.73 respectivamente. El resto de las variables tienen una correlación baja con la variable canónica socioeconómica. En este caso, algunas variables sociosanitarias se relacionan mejor con la variable canónica socioeconómica que las propias variables socioeonómicas.

Tabla 2.6: ACC: correlaciones de las variables con las variables canónicas socioeconómicas

	Dimensión 1	**Dimensión 2**	**Dimensión 3**	**Dimensión 4**
Sociosanitarias				
E0.17	-0.7272	0.0177	0.0372	-0.0336
E18.64	-0.7500	0.0345	0.0695	0.0069
E65.00	-0.6205	-0.0450	0.1837	-0.0007
Fallecidos	-0.4369	0.2779	0.1221	-0.0830
Prevalencia	-0.6527	-0.0331	0.0525	-0.0468
Enfermos	-0.7445	0.0483	0.0600	-0.0413
Socioeconómicas				
Edad_media	0.8651	-0.4770	0.0398	-0.0379
Renta_bruta	0.3893	0.5924	0.4016	-0.4566
Neta_persona	-0.6620	-0.3059	-0.2254	-0.2438
Gini	-0.3604	-0.3686	0.1159	-0.6532
Pob_ine	-0.7349	-0.4389	0.4853	-0.0147

2.3.4.1 Biplot

El Análisis Biplot del ACC (Figura 2.12) se lleva a cabo teniendo en cuenta todas las dimensiones canónicas aunque, como ya se comentó anteriormente, las dos últimas dimensiones tienen muy poca correlación entre la parte sociosanitaria y la socioeconómica. En el paquete de R CCA utilizado solo es posible mostrar el biplot en dos gráficos separados, uno para variables y otro para individuos.

En el biplot de las dimensiones canónicas 1 y 2 se observa que las variables sociosanitarias y socioeconómicas se relacionan principalmente con la dimensión 1, excepto fallecidos que tiene una correlación aceptable con la dimensión 2. Todas las variables sociosanitarias tienen una correlación negativa con esta dimensión y relacionadas con las variables socioeconómicas: población del INE, índice Gini y renta neta. Mientras que en el lado positivo se presentan la edad media y la renta bruta.

De forma parecida al ACP, esta relación sugiere que las variables cosioeconómicas han tenido una influencia en la magnitud del Covid–19 , pero los falledidos se alejan de esta relación y la dimensión 2 tiene un efecto sobre esta variables.

La distribución de las zonas básicas de salud en el biplot muestra una aglomeración en valores algo positivos de la dimensión 1, relacionados con la edad media y la renta bruta. Solo unas zonas básicas de salud se acercan a las variables sociosanitarias en el lado negativo de la dimensión 1.

En el biplot de las dimensiones canónicas 3 y 4 se observa como todas las variables tienen poca correlación con las variables canónicas aunque la dimensión 3 tiene cierta correlación con E65.00 mientras que la variable fallecidos tiene cierta correlación tanto con la dimensión 3 como con la 4. A diferencia del ACP, estas dimensiones canónicas superiores podrían tener más relación con variables sociosanitarias que con variables socioeconómicas. Además, las variables sociosanitarias implicadas son variables especialmente importantes en el Covid–19 ya que se

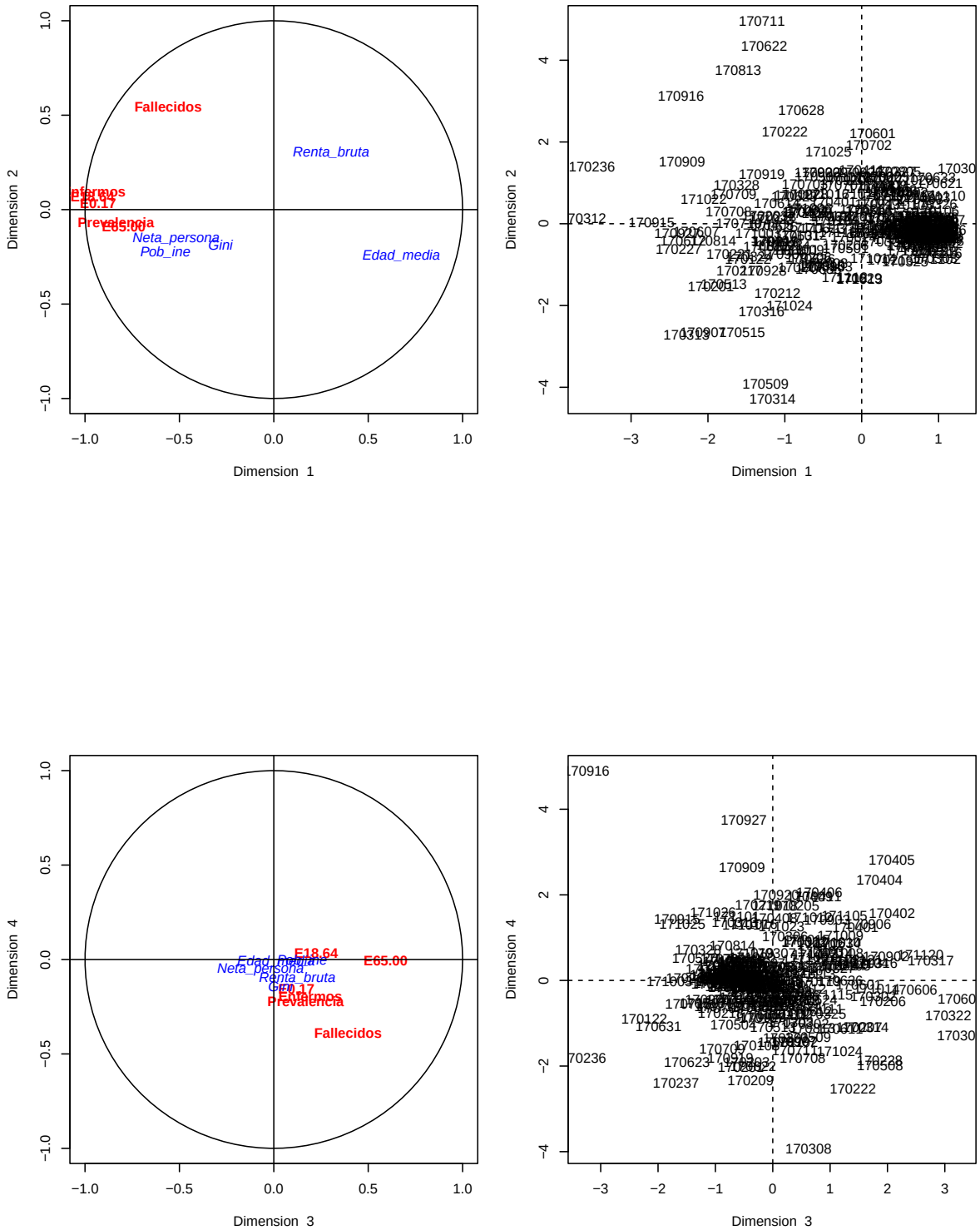


Figura 2.12: ACC: biplot usando las 2 primeras dimensiones canónicas

produjeron más fallecimientos en personas mayores de 65 años como se afirma en Bravo-Segal y Villar (2020).

2.3.4.2 Distribución espacial

En la Figura 2.13 se muestros los valores de las variables canónicas sociosanitarias y socioeconómicas de la primera dimensión canónicas. Se observa la existencia de una correlación positiva entre ambas con un $R^2 = 0.610$ que indica que el 61% de la variación de una de ellas se puede explicar con la otra. El valor no es muy elevado pero hay que tener en cuenta que el objetivo sería demostrar que no existe correlación entre ambas, es decir, que las variables socioeconómicas no permiten explicar la situación sociosanitaria del Covid–19 en la C.A. de Castilla y León.

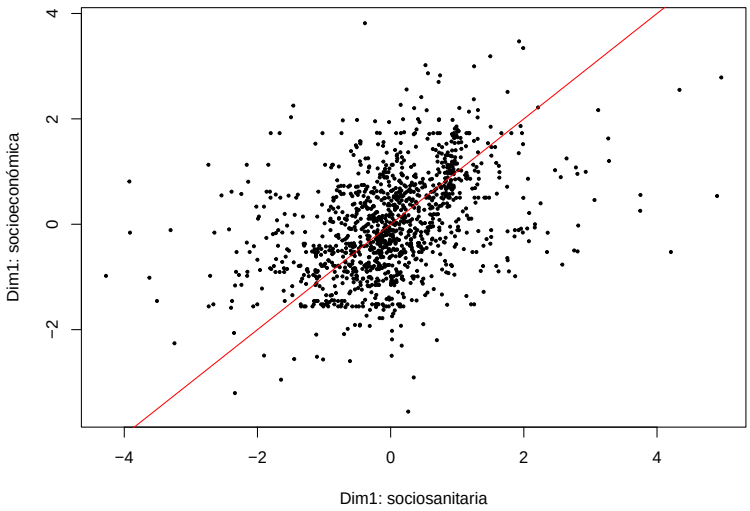
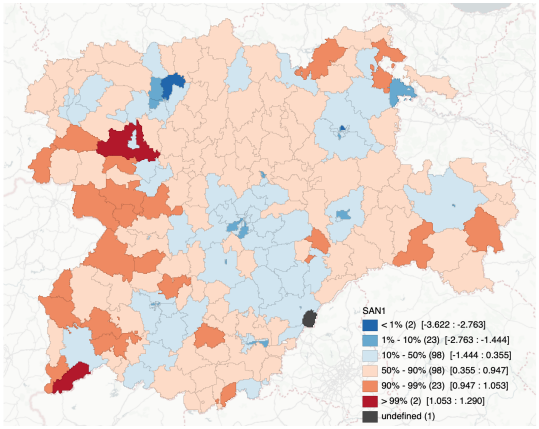


Figura 2.13: ACC: Diagrama de puntos de las dos variables canónicas de la primera dimensión ($r = 0.781$)

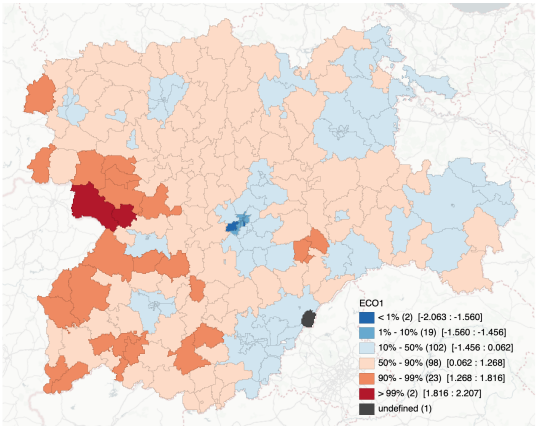
En la Figura 2.14 se muestros los valores de las variables canónicas sociosanitarias y socioeconómicas en las dos primeras dimensiones canónicas.

En la dimensión 1, fundamentalmente sanitaria relacionada con el número de enfermos se observa un parecido entre ambos mapas con las zonas básicas de salud con un percentil superior al 90% presentes en el oeste de la C.A. de Castilla y León. La relativa similitud de los mapas es debida a la correlación alta de la dimensión 1 (mayor de 0.78) y que describen el mismo efecto. Pero la información que se puede extraer es muy distintas de la observada en la Figura 2.13.

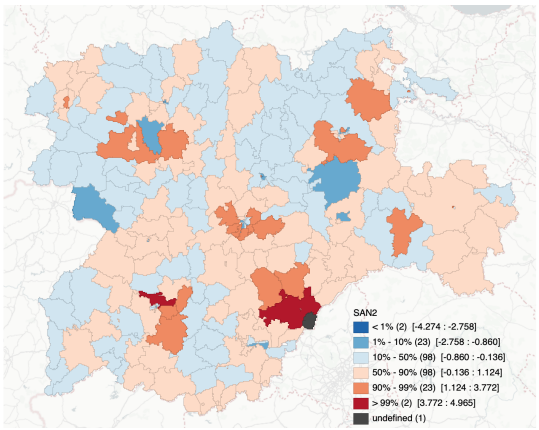
Por el contrario, la dimensión 2 tiene una menor correlación entre ambas variables canónicas (menor de 0.51), lo que se traduce en una mayor diferencia entre ambos mapas.



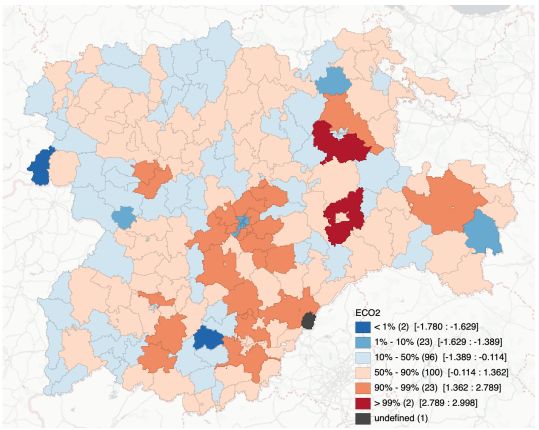
Dim.1: Sociosanitaria



Dim.1: Socioeconómica



Dim.2: Sociosanitaria



Dim.2: Socioeconómica

Figura 2.14: ACC: distribución espacial de las dos primeras variables canónicas derivadas de varibales sociosanitarias y socioeconómicas

2.4 Conclusiones

Las principales conclusiones de este trabajo son:

1. Las variables con mayor relación con el Covid–19 son: renta bruta, renta neta por persona, índice de Gini, población del INE, edad menor de 17 años, edad de entre 18 y 64 años, edad de más de 65 años, fallecidos, prevalencia y enfermos. El resto de variables obtenidas de las bases de datos muestran una elevada correlación con ellas.

2. Las variables sociosanitarias se relacionan principalmente entre ellas mientras que en las socioeconómicas aparecen tres grupos relativamente diferenciados. Estos grupos están caracterizados por las siguientes variables: edad media relacionada con los ingresos, rentas netas y renta bruta media, y los índices de desigualdad relacionados con los datos poblacionales.

3. En el estudio del Covid–19 se detectan dos dimensiones principales: una principalmente sociosanitaria y otra socioeconómica, aunque variables de ambos tipos participan en sus definiciones. Así, la dimensión socioeconómica viene definida por la renta bruta y el índice de Gini (similar renta neta o población total) aunque tienen influencia sobre el número de fallecidos. Mientras que en la dimensión sociosanitaria, la edad media es un buen indicar de forma que menor edad media en la zona de salud implica mayor nivel de Covid–19 .

3
BIBLIOGRAFÍA Y REFERENCIAS

BOCyl nº 41 (1988). *Decreto 32/1988 de 18 de febrero por el que se establece la delimitación territorial de las Zonas Básicas de Salud en el territorio de la Comunidad autónoma de Castilla y León*. BOCyL 1 de marzo de 1988.

Bohoslavsky, J. P., Bachelet, M., & Segato, R. L. (2020). *Covid-19 y derechos humanos: La pandemia de la desigualdad*. Editorial Biblos. Obtenido de https://books.google.es/books?id=XsgBEAAAQBAJ

Brandily, P., Brébion, C., Briole, S., & Khoury, L. (2021). A poorly understood disease? The impact of COVID-19 on the income gradient in mortality over the course of the pandemic. *European Economic Review*, *140*, 1-28.

Bravo-Segal, S. & Villar, F. (2020). La representación de los mayores en los medios durante la pandemia COVID-19: ¿hacia un refuerzo del edadismo?. *Revista Española de Geriatría y Gerontología*, *55*(5), 266-271.

Cárdenas, O., Galindo, P., & Vicente-Villardón, J. L. (2007). Los métodos Biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, *13*(1), 279-303.

Cuadras, C. M. (1981). *Métodos de Análisis Multivariante*. Barcelona: Editorial Universitaria de Barcelona.

DATAtab (2023). *Análisis de correlación*. Obtenido de https://datatab.es/tutorial/correlation (Acceso: 2023-06-21)

Eckart, C. & Young, G. (1936). The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, *1*, 211-218.

Esparza-Rodríguez, S. A., Martínez-Arroyo, J., García-Tapia, G., & Esquivel-Fernández, E. (2021). Retomando Estrategias Ante Crisis por Covid19: Zona Económica Especial de Lázaro Cárdenas. *Políticas Públicas*, *14*(1), 37-55.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

_____ (1972). Analysis of meteorological data by means of canonical descomposition and Biplots. *Journal of Applied Meteorology*, *11*, 1071-1077.

Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Questiio*, *10*(1), 13-23.

González, I. & Déjean, S. (2022). *Canonical Correlation Analysis*. Obtenido de https://CRAN.R-project.org/package=CCA (R package 'CCA' versión 1.2.1)

González-García, N. & Taborda-Londoño, A. (2015). *Análisis de Componentes Principales Sparse: Formulación, algoritmos e implicaciones en el análisis de datos*. (Trabajo Fin

de Máster en Análisis Avanzado de Datos Multivariantes). Departamento de Estadística, Universidad de Salamanca.

Gower, J. C. & Hand, D. J. (1996). *Monographs on Statistics and Applied Probability 54: Biplots*. London: Chapman and Hall.

Herrer-Cartaya, C. E., Lage-Dávila, A., Betancourt-Cervantes, J., Barreto-Flu, E., Sánchez-Valdés, L., & Hernández-Claro, L. (2022). La edad como variable asociada a la gravedad en pacientes con la COVID-19. *Revista Cubana de Medicina Militar*, *51*(1), e1766.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *24*, 417-441, 498-520.

Hotelling, H. (1936). Relations between two set of variates. *Biometrika*, *28*, 321-377.

Junta de Castilla y León (2007). *Guía de Ordenación Sanitaria de Castilla y León*. Valladolid: Junta de Castilla y León. Consejería de Sanidad. Obtenido de https://www.saludcastillayleon.es/institucion/es/organizacion/ordenacion-sistema-sanitario/guia-ordenacion-sanitaria-castilla-leon

Kassambara, A. & Mundt, F. (2020). *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Obtenido de https://CRAN.R-project.org/package=factoextra (R package 'factoextra' version 1.0.7)

Laajaj, R., Webb, D., Aristizaba, D., Behrentz, E., Bernal, R., Buitrago, G., … Vives, M. (2022). Understanding how socioeconomic inequalities drive inequalities in COVID-19 infections. *Scientific Reports*, *12*, 8269.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1-18.

McArthur, L., Sakthivel, D., Ataide, R., Chan, F., Richards, J. S., & Narh, C. A. (2020). Review of Burden, Clinical Definitions, and Management of COVID-19 Cases. *The American Journal of Tropical Medicine and Hygiene*, *103*(2), 625-638.

Middya, A. I. & Roy, S. (2021). Geographically varying relationships of COVID-19 mortality with different factors in India. *Scientific Reports*, *11*, 7890.

Ministerio de Sanidad (2023). *Centros y Servicios del Sistema Nacional de Salud: Introducción*. Obtenido de https://www.sanidad.gob.es/ciudadanos/prestaciones/centrosServiciosSNS/hospitales/introduccionCentro.htm (Acceso: 2023-06-21)

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 539-572.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Obtenido de https://www.R-project.org/

Ramírez-Figueroa, J. A. (2021). *Análisis de componentes principales disjuntas por medio de optimización por enjambre de partículas y sus aplicaciones*. (Tesis Doctoral). Departamento de Estadística, Universidad de Salamanca.

Riquelme, S. F. (2020). Primera Historia de la crisis del Coronavirus en España. *Revista hispanoamericana de Historia de las Ideas*, *46*, 12-22.

Rudi Rocha Rifat Atun Adriano Massuda, Paula. Spinola. (2021). Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. *Lancet Global Health*, *9*, e782-92.

Sánchez-González, M. A. (2021). HISTORIA Y FUTURO DE LAS PANDEMIAS. *Revista MÃ©dica ClÃnica Las Condes*, *32*(1), 7-13.

Vicente-Villardón, J. L. (1992). *Una alternativa a los métodos factoriales clásicos basada en una generalización de los metodos biplot*. (Tesis Doctoral). Departamento de Estadística, Universidad de Salamanca, Spain.

Vicente-Villardón, J. L., Galindo, P., & Blázquez, A. (2006). Logistic Biplots. En M. Greenacre & J. Blasius (Editores), *Multiple Correspondence Analysis and related methods*. (pp. 491-509). London: Chapman & Hall.

Wachtler, B., Michalski, N., Nowossadeck, E., Diercke, M., Wahrendorf, M., Santos-Hövener, C., … Hoebel, J. (2020). Socioeconomic inequalities and COVID-19 – A review of the current international literature. *Journal of Health Monitoring*, *5 (S7)*, 3-17.

Wei, T. & Simko, V. (2021). *Visualization of a Correlation Matrix*. Obtenido de https://github .com/taiyun/corrplot (R package 'corrplot' version 0.92)

Yang, W. & Kang, M. (2003). *GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists*. New York: CRC Press LLC.

Zaim, S., Chong, J. H., Sankaranarayanan, V., & Harky, A. (2020). COVID-19 and Multiorgan Response. *Current Problems in Cardiology*, *45*(8), 100618.