ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS
DEPARTAMENTO DE POSTGRADO

PROYECTO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

“MAGÍSTER EN ESTADISTICA CON MENCIÓN EN GESTIÓN DE LA CALIDAD Y PRODUCTIVIDAD”

TEMA:

ANÁLISIS ESTADÍSTICO MULTIVARIANTE PARA IDENTIFICAR FACTORES QUE INFLUYERON EN EL DESARROLLO DE LA COVID-19 EN ECUADOR

AUTOR:

ADOLFO DANIEL SALCEDO MALDONADO

Guayaquil - Ecuador

2021

AGRADECIMIENTO

Agradezco principalmente a mi tutor, el Doctor Omar Ruiz, por su constante motivación y valiosa guía para la ejecución de este proyecto de titulación.

Quisiera agradecer también a los doctores Sergio Bauz y Johny Pambabay, quienes evaluaron este proyecto. Sus recomendaciones y observaciones ayudaron a elevar la calidad de este documento considerablemente.

Finalmente, agradecer a los profesores y directivos de la ESPOL por ayudarme a incrementar mis conocimientos y crecer como profesional.

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría. El patrimonio intelectual del mismo corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

ADOLFO DANIEL SALCEDO MALDONADO

Nombre del Autor

TRIBUNAL DE GRADUACIÓN

Francisco Vera Alcívar, Ph.D.
PRESIDENTE

Omar Ruiz Barzola, Ph.D.
DIRECTOR

Sergio Bauz Olvera, Ph.D.
VOCAL 1

Johny Pambabay Calero, Ph.D.
VOCAL 2

TABLA DE CONTENIDO

Listado de tablas

Listado de figuras

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Antecedentes

Las pandemias no son algo nuevo para la humanidad. Como afirman (Arbós & Belles, 2020), los grandes asesinos de la historia son las bacterias y los virus, y en concreto los que han provocado las grandes epidemias de la historia. Algunas de las más representativas son: el Variola virus, causante de la viruela y causante de alrededor de 300 millones de muertes; el sarampión, que acabó con más de 200 millones de personas; la pandemia mundial de influenza de 1918 (también conocida como la gripe española), que afectó a un tercio de la población mundial y mató entre 50 y 100 millones de personas; y el virus del sida o VIH, que ha matado a más de 35 millones (Infosalus, 2020). Sin embargo, esta ha sido la primera pandemia provocada por un coronavirus.

Los coronavirus se pueden contagiar de los animales a las personas (transmisión zoonótica) (OMS, 2020a). De acuerdo con estudios exhaustivos al respecto, sabemos que el síndrome respiratorio agudo grave más conocido como SARS-CoV se transmitió de la civeta al ser humano, y que se ha producido transmisión del síndrome respiratorio de Oriente Medio - MERS-CoV del dromedario al ser humano. Además, se sabe que hay otros coronavirus circulando entre animales, que todavía no han infectado al ser humano. En los humanos, se caracterizan por provocar infecciones respiratorias que pueden variar desde un resfriado común hasta enfermedades más graves como el MERS-CoV y el SARS-CoV. (OMS, 2020b).

La enfermedad COVID-19 -acrónimo del idioma ingles coronavirus disease 2019 (OMS, 2020c), es una enfermedad generada por el virus SARS-CoV-2 (BBC News, 2020). Su primer caso fue detectado en la ciudad de Wuhan, China, en diciembre de 2019 y hasta los primeros días de mayo ya había afectado a más de 200 países y territorios, con más de 3,8 millones de casos confirmados y más de 260 mil muertes.

Los síntomas pueden aparecer entre 2 y catorce días después del contagio, con un promedio de 5 días, y varían ampliamente. Ciertas personas no presentan síntomas, pero generalmente se presentan síntomas similares a los de la gripe: fiebre, tos seca, disnea, mialgia y fatiga. La mayoría de los casos (alrededor del 80%) (OMS, 2020b) se recuperan de la enfermedad sin necesidad de tratamiento hospitalario, mientras el 20% (OMS, 2020b) restante acaba presentando un cuadro grave y experimenta dificultades para respirar. Las personas mayores y/o aquellas que padecen afecciones médicas previas tienen más probabilidades de presentar cuadros graves (OMS, 2020b). Algunas de estas afecciones previas pueden ser: hipertensión arterial, problemas cardiacos o pulmonares, diabetes o cáncer. Sin embargo, esto no quiere decir que una persona sana no pueda caer gravemente enferma. En casos graves se caracteriza por producir neumonía, síndrome de dificultad respiratoria aguda, sepsis y choque séptico. Hasta el mes de septiembre de 2020, no existe tratamiento específico; lo que se busca principalmente es aliviar los síntomas y mantener las funciones vitales.

La transmisión del virus se produce mediante pequeñas gotas (microgotas de Flügge) que una persona portadora emite al hablar, estornudar, toser o espirar (OMS, 2020b). Estas gotículas son relativamente pesadas, no llegan muy lejos y caen rápidamente al suelo, pero pueden ser inhaladas por otra persona, por lo que se recomienda mantener una distancia de al menos un metro. Otra forma de contagio se da cuando las gotas caen sobre una superficie, la cual luego sea tocada por otra persona que a su vez toca sus ojos, nariz o boca. Para prevenir este tipo de contagio se recomienda el lavado

frecuente de manos o su desinfección con alcohol, así como la utilización de mascarillas (OMS, 2020b). Otra recomendación es la utilización de la hendidura entre el antebrazo y brazo, opuesta al codo, para estornudar o toser (OMS, 2020c).

La expansión del coronavirus SARS-CoV-2 (causante de la COVID-19) fue sumamente rápida. Para el 30 de enero de 2020 se reportaban casos en 15 países además de China, fecha en la cual la Organización Mundial de la Salud la declaró una emergencia sanitaria de preocupación internacional, basándose no solamente en su velocidad de expansión, sino en el impacto que el virus podría tener en países subdesarrollados con menos infraestructuras sanitarias.

El 11 de marzo de 2020 la Organización Mundial de la Salud la declaró pandemia (OMS, 2020d), fecha para la cual los casos confirmados ascendían a más de 118 mil en el mundo, con más de 4 mil muertos (Departamento de Seguridad Nacional de España, 2020).

Ecuador fue el tercer país latinoamericano en confirmar un caso de COVID-19, luego de Brasil y México. El caso fue confirmado el 29 de febrero y se trataba de una ciudadana ecuatoriana que llegó desde España el 14 de febrero por el aeropuerto de Guayaquil.

Al 30 de junio de 2020, según cifras oficiales, en Ecuador se han confirmado 56.342 casos y 4.527 fallecidos a causa del virus a nivel nacional, además de 3.071 fallecidos probables a causa del virus (Ministerio de Salud Pública, 2020). Las provincias más afectadas son Guayas (31,7% de los casos), Pichincha (15,2% de los casos) y Manabí (9,4% de los casos).

1.2. Descripción del problema

El riesgo de aparición de pandemias se incrementa a medida que pasa el tiempo, debido al constante incremento poblacional y la capacidad y velocidad con la que las personas podemos viajar alrededor del mundo. Tal es la preocupación al respecto que, durante el 2019, meses antes de la aparición de la COVID-19, la Organización Mundial de la Salud (OMS) y el Banco Mundial comisionaron un informe para evaluar la posibilidad de una emergencia sanitaria a nivel mundial. El informe señala que entre 2011 y 2018, la OMS registró 1483 brotes epidémicos en 172 países.

Teniendo en cuenta la variabilidad en los síntomas de estas nuevas enfermedades, su propensión a ser cada vez más frecuentes y rápidas en su propagación, así como su impacto en los sistemas de salud, se hace vital conocer los diferentes factores de riesgo que pudieran hacer más vulnerable a la población.

1.3. Objetivo

Analizar variables sociodemográficas y de salud, y su relación con la enfermedad COVID-19 en el Ecuador; identificando factores o variables que pudieran ser considerados de riesgo, a través de metodologías multivariantes.

1.4. Objetivos específicos

- Obtener una base de datos depurada y completa para realizar el análisis planteado en el objetivo.
- Estudiar de manera univariante el comportamiento de cada variable que forma parte de este estudio, correspondientes a las agrupaciones: condiciones de vivienda, condiciones socioeconómicas, sistema de salud, población y de afectación
- Analizar las relaciones existentes entre las variables objeto de estudio.
- Identificar posibles factores de riesgo que favorecieron el impacto de la enfermedad.

1.5. Alcance

Este estudio está enfocado en el análisis de los factores que contribuyen a la vulnerabilidad de la población de cada provincia del Ecuador, frente a la pandemia de la COVID-19. Para ello se utilizarán datos oficiales provistos por INEC, Secretaría Nacional de Gestión de Riesgos, Ministerio de Salud pública; cada institución provee información con datos de diferentes fechas, las cuales son descritas en el apartado que corresponde. Sin embargo, para el presente estudio, con respecto a los datos COVID-19 en Ecuador, se hizo un corte al 31 de mayo de 2020.

CAPÍTULO 2

2. MARCO TEÓRICO

2.1. COVID-19

Poco se sabe sobre la COVID-19; actualmente se siguen realizando muchos estudios para saber cuáles pueden ser sus posibles efectos sobre la salud humana. Por ahora la OMS únicamente ha dado indicaciones para prevenir el contagio, ya que no se conoce de tratamientos efectivos contra la COVID-19. Sin embargo, hay varios tratamientos potenciales en estudio (Reuters, 2020).

Ante esta incertidumbre, lo recomendable es la prevención; esto ayudaría a esperar que científicos encuentren la vacuna que contrarreste los efectos de la COVID-19. Varios autores han trabajado sobre medidas de prevención (Firth, 2020).

Algunos de estos trabajos, como el de (Li, y otros, 2020a), han realizado un análisis y modelamiento predictivo sobre la propagación de la COVID-19. (Zhang, Ma, & Wang, 2020a), han realizado modelos para predecir el punto de inflexión, la duración y la tasa de ataque de los brotes de COVID-19 en los principales países occidentales.

(Chen, y otros, 2020) han realizado un estudio para integrar Big Data y análisis de datos con el fin de prevenir un posible brote hospitalario de COVID-19 en Taiwán. (Yang, Chen, & Chen, 2020) han estudiado estrategias preventivas del hospital comunitario en la batalla de lucha contra la pandemia COVID-19 en Taiwán.

Así mismo, (Yen, Schwartz, Chen, & King, 2020) han realizado un estudio para medir la interrupción de la transmisión COVID-19, mediante la implementación

de un paquete mejorado de control de tráfico. Ellos también analizan las implicaciones para los esfuerzos de prevención y control global.

En cuanto a factores de riesgo, un estudio realizado por (Zhang, y otros, 2020b) a 619 pacientes ingresados por COVID-19 en el hospital Renmin de la Universidad de Wuhan, entre el 11 de enero y el 6 de febrero 2020, determinó que, ser del sexo masculino, tener una condición severa de la COVID-19, expectoración, dolor muscular y una disminución de la albúmina, son factores de riesgo independientes con influencia en la mejoría de la enfermedad.

(Li, y otros, 2020b) realizaron un estudio similar a 548 pacientes en el hospital Tongji de Wuhan, el cual determinó como factores de riesgo el pertenecer al sexo masculino, tener una edad avanzada, presentar leucocitosis, niveles elevados de LDH, presencia de alguna lesión cardíaca e hiperglucemia.

En un estudio paralelo realizado por otro grupo de investigadores, (Li, y otros, 2020c) evaluaron si la existencia de una enfermedad cardiovascular contribuye a la progresión y podría adelantar un mal pronóstico para los pacientes. El estudio incluyó a 83 pacientes con COVID-19, de los cuales 42 tenían enfermedades cardiovasculares y 41 no. Su conclusión fue que efectivamente una enfermedad cardiovascular es un fuerte factor de riesgo para una rápida progresión y un mal pronóstico de la enfermedad.

(Pal & Bhadada, 2020) encontraron en los pacientes con COVID-19 y diabetes mellitus un círculo vicioso, en el cual la COVID-19 y algunas drogas utilizadas para tratarlo empeoran la disglicemia y la diabetes, lo cual a su vez exacerba la severidad de los problemas respiratorios e incrementa la mortalidad en estos pacientes.

La información recopilada para la elaboración de este trabajo será sometida a diferentes análisis como análisis univariante, bivariante, multivariante, de componentes principales, entre otros.

2.2. Indicadores epidemiológicos

Al realizar un proceso de investigación sobre epidemiología, tal como en las demás ciencias, se utilizan variables, proporciones, tazas, razones, etc. Para el presente trabajo se utilizarán como variables de respuesta a 3 indicadores de esta ciencia:

Tasa de mortalidad

La tasa de mortalidad es el volumen de muertes ocurridas en una población, por todas las causas de enfermedad, en todos los grupos de edad y para ambos sexos. Ésta puede ser cruda o ajustada. La diferencia radica en que la mortalidad ajustada considera posibles diferencias por edad, sexo, etc. (Moreno-Altamirano, López-Moreno, & Corcho-Berdugo, 2000)

Para este trabajo se utilizará la tasa de mortalidad cruda, a la cual nos referiremos de aquí en adelante simplemente como tasa de mortalidad. Esta tasa de mortalidad indica la relación entre la cantidad de fallecidos por COVID-19 y la cantidad de personas en una población y en un periodo de tiempo.

Tasa de letalidad

La letalidad es una medida de la gravedad de una enfermedad, en función de su capacidad para producir la muerte. Se define como la proporción de muertes a causa de una enfermedad con respecto al total de casos de dicha enfermedad, en un periodo especificado. En este trabajo, la tasa de letalidad indica la relación entre fallecidos a causa de COVID-19 y la cantidad de personas que contrajeron la enfermedad en la población, en el período estudiado.

Proporción de incidencia

En epidemiología, las proporciones son medidas que expresan la frecuencia con la que ocurre un evento en relación con la población total en la cual éste puede ocurrir (Moreno-Altamirano, López-Moreno, & Corcho-Berdugo, 2000).

La proporción de incidencia en este caso entonces nos indica la relación entre los casos confirmados de COVID-19 y la población. En otras palabras, nos indica qué tanto se ha esparcido la enfermedad.

2.3. Análisis univariante y bivariante

El análisis univariante, también conocido como estadística descriptiva, se utiliza para estudiar el comportamiento de una variable individualmente. Como indican (Walpole, Myers, Myers, & Ye, 2012): Hay ocasiones en las que no se desea llegar al detalle de la estadística inferencial, sino que únicamente se desea obtener la información resumen de un conjunto de datos representados a través de la muestra; para estos casos se utiliza un análisis univariante. Se utilizará el análisis univariante para realizar una evaluación inicial de los datos obtenidos.

Los análisis univariantes más comunes son:

- Calcular medidas de tendencia central y de dispersión
- Determinar distribuciones de frecuencias
- Obtener índices sobre la característica de la distribución de frecuencias (curtosis, simetría)
- Realizar inferencias estadísticas
- Graficas como: histogramas, de cajas, tallo y hojas

El análisis bivariante consiste en estudiar las relaciones que existen entre variables observadas o agrupadas de dos en dos, así como el comportamiento de una variable en función de otra. Se caracteriza por tener una variable predictora y otra de respuesta. Algunas de las metodologías utilizadas son:

- Tabla de doble entrada o de contingencia
- Distribución de frecuencias marginales
- Distribución conjunta de frecuencias marginales
- Distribución de frecuencias condicionadas

- Diagramas de dispersión

Una vez realizado los análisis univariantes y bivariante, el siguiente paso es realizar un análisis con más de dos variables, es decir, un análisis multivariante.

2.4. Análisis Multivariante

Cuando se desea analizar las relaciones que existen entre más de 2 variables estadísticas estudiadas conjuntamente sobre una muestra de individuos, las cuales pueden ser cualitativas, cuantitativas o una mezcla de ambas, es necesario aplicar un análisis multivariante.

La información estadística en análisis multivariantes es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental (Cuadras, 2004). Debido a la cantidad de datos y variables que intervienen, este análisis se realiza mediante el uso de software.

Según (Peña, 2002), los objetivos del análisis multivariante son los siguientes:

- Crear nuevas variables a partir del conjunto original. Estas variables deben ser menos y corresponderán a transformaciones de las originales, perdiendo la menor cantidad de información posible.
- Evaluar si es posible crear grupos con los datos
- Utilizar agrupaciones definidas para clasificar nueva información.
- Relacionar dos conjuntos de variables.

Dado que los datos levantados contienen una gran cantidad de variables predictoras, se reducirá su cantidad utilizando una técnica conocida como análisis de componentes principales.

2.5. Análisis de componentes principales (ACP)

Una de las técnicas más utilizadas para el análisis multivariante es la de componentes principales. Fueron desarrollados por Hotelling en 1933, pero se originan de los ajustes ortogonales por mínimos cuadrados planteados por K. Pearson en 1901 (Peña, 2002). Su objetivo es reducir la dimensionalidad de la información, es decir, reducir la cantidad de variables mediante la cual se explica un problema, disminuyendo la precisión lo menos posible y perdiendo la menor cantidad posible de información. Adicionalmente, es el paso inicial para detectar variables que podrían estar generando variabilidad de los datos, pero que no han sido observadas. Si se está trabajando con variables altamente dependientes, con frecuencia uno encontrará que una pequeña cantidad de variables nuevas (menor al 20% original) explican la mayor cantidad de la variabilidad original (más del 80%) (Peña, 2002).

Geométricamente, lo que busca el ACP es encontrar un subespacio a partir del original, pero que sea de menor dimensión y en el cual al proyectar los puntos, éstos conserven su estructura, con la menor distorsión posible. Para plantear un ejemplo sencillo, se considerará un espacio de dos dimensiones. En la **Figura 1** se presenta un diagrama de dispersión y una recta colocada intuitivamente, buscando que ésta proporcione un buen resumen de los datos. Con este fin, se dibujó la recta tratando de que pase cerca de todos los puntos, de tal manera que, al analizar la distancia entre cada punto y su proyección sobre la recta, estas resulten lo más pequeñas posible. Considerando entonces un punto x_i y una dirección $a_1 = \left(a_{11}, \dots, a_{1p}\right)'$, definida por un vector a_1 de norma uno, el vector del punto sobre esta dirección será $z_i a_1$ y su proyección será el escalar:

$$z_i = a_{11}x_{i1} + \cdots + a_{1p}x_{ip} = a'_1 x_i$$

El objetivo entonces será minimizar la distancia r_i entre el punto x_i y su proyección sobre la dirección a_1, esto es:

$$minimizar \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |x_i - z_i a_1|^2$$

donde $|u|$ es la norma euclídea o módulo del vector *u*.

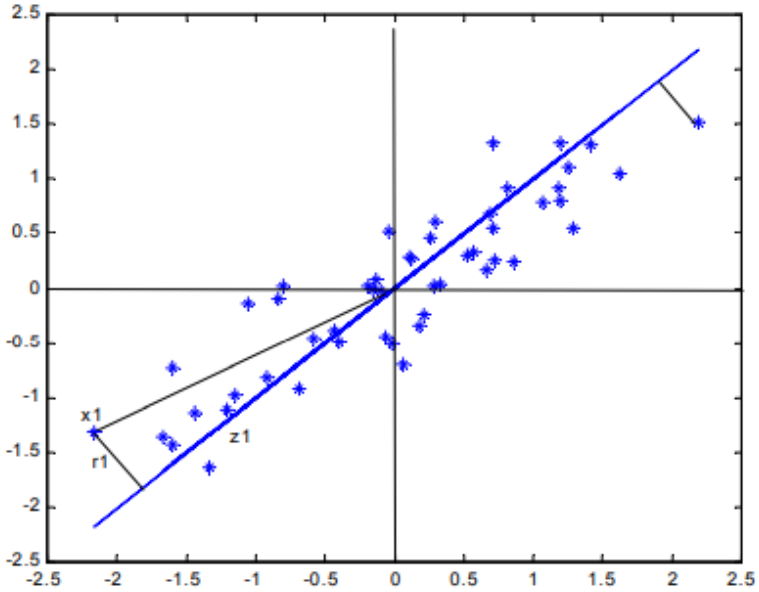


Figura 1. *Recta que minimiza distancias ortogonales (ACP)*

En la Figura 1 se puede ver que se forma un triángulo rectángulo entre el punto x_1, su proyección sobre la recta y el origen. Por el teorema de Pitágoras tenemos:

$$x_i' x_i = z_i^2 + r_i^2,$$

o para todos los puntos:

$$\sum_{i=1}^{n} x_i'\, x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2$$

Dado que la distancia de cualquier punto al origen es constante, minimizar la longitud de las proyecciones es equivalente a maximizar la distancia entre las proyecciones y el origen. Las proyecciones z_i son variables de media cero, por lo que maximizar la suma de sus cuadrados equivale a maximizar su varianza (Peña, 2002). Intuitivamente se puede llegar la conclusión de que esto es correcto, ya que viendo la recta de la Figura 1 se puede llegar a la conclusión de que es adecuada, ya que conserva tanto como es posible la variabilidad original de los puntos. Por el contrario, si se dibuja una recta en

sentido perpendicular, los puntos proyectados tendrían muy poca variabilidad y se perdería información de sus distancias.

Para explicar la forma de cálculo de los componentes principales, consideremos que se tiene una matriz de *n* individuos con *m* variables aleatorias, el ACP permite encontrar un número de factores subyacentes (cada uno denominado componente principal) $p < m$ que corresponde a una buena aproximación de los valores de las variables (*m*) para cada individuo (*n*). Para aplicar el ACP se pueden aplicar dos métodos básicos:

- Cuando las variables aleatorias tienen diferentes órdenes de magnitud o cuando las observaciones no tienen dimensiones homogéneas, se aplica el método basado en la matriz de correlación.
- Cuando los datos tienen valores promedio similares y los datos son homogéneos, se aplica el método basado en la matriz de covarianzas.

Método basado en la matriz de correlación

Partiendo de la matriz de correlaciones y teniendo los valores de las m variables aleatorias correspondientes a *n* individuos, tomaremos el cojnto de datos y los escribiremos en una matriz:

$$\left(F_j^{\beta}\right)_{j=1,\dots,m}^{\beta=1,\dots,n}.$$

Cada conjunto de datos $\mathcal{M}_j = \left\{F_j^{\beta} \middle| \beta = 1, \dots, n\right\}$ puede considerarse una muestra aleatoria para la variable F_j. La matriz de correlación puede elaborarse a partir de los *m* x *n* datos correspondientes a las *m* variables aleatorias, la cual está definida por:

$$R = [r_{ij}] \in M_{mxm}, \qquad \text{donde} \qquad r_{ij} = \frac{cov(F_i, F_j)}{\sqrt{var(F_i)var(F_j)}}.$$

Dado que la matriz de correlaciones es simétrica es posible diagonalizarla y sus valores propios λ_i verifican:

$$\sum_{i=1}^{m} \lambda_i = m$$

Dado que se cumple esta igualdad, a los *m* valores propios correspondientes a la misma cantidad de componentes principales, se les da la denominación de **pesos**. La base de vectores propios de la matriz *R* representa a los factores principales que fueron reconocidos matemáticamente. Finalmente, cabe recalcar que cada variable original puede ser planteada como una combinación lineal de los componentes principales o los vectores propios. (Wikipedia, 2020a)

Método basado en la matriz de covarianzas

Se utilizará la matriz de covarianza para obtener una matriz *Y* de dimensión *n* x *l* a partir de una matriz de datos *X* de dimensión *n* x *m*, perdiendo la menor cantidad de información útil.

Partiendo de un conjunto de datos con *n* muestras y sus correspondientes *m* variables descriptivas, se buscará describir cada muestra con una menor cantidad de variables (*l*) y que la cantidad de componentes principales sea menor que la menor de las dimensiones de *X*.

$$l \leq \min\{n, m\}$$

Antes de poder aplicar este método se debe centrar y/o autoescalar los datos. Lo primero se realiza restándoles la media de cada columna y lo segundo centrándolos y dividiendo cada columna por su desviación estándar.

$$X = \sum_{a=1}^{l} t_a \, p_a^T + E$$

Los vectores t_a son ortogonales y se los denomina scores. Éstos contienen la información sobre cómo las muestras se relacionan unas con otras. Los

vectores p_a son ortonormales y se los denomina loadings. Éstos contienen la información sobre la relación entre las variables. El error que se produce, debido a que se toman menos componentes principales que variables y por el mismo error de ajuste del modelo con respecto a los datos, se acumula en la matriz E.

La base de este método está en la descomposición de la matriz de covarianza en vectores propios, la cual se calcula de la siguiente manera:

$$cov(X) = \frac{X^T X}{n-1}$$

$$cov(X)p_a = \lambda_a p_a$$

$$\sum_{a=1}^{m} \lambda_a = 1$$

Donde λ_a es el valor propio asociado al vector propio p_a. Por último,

$$t_a = X p_a$$

Esta ecuación la podemos entender como que t_a son las proyecciones de *X* en p_a. La información representada por cada componente principal, es decir la cantidad de varianza explicada, se mide mediante los valores propios λ_a. La cantidad de varianza explicada disminuye a medida que se avanza por cada componente principal. Esto quiere decir que el primer componente dará más información que el segundo, el segundo más que el tercero, etc. (Wikipedia, 2020a)

Una vez obtenidos los componentes principales, se debe definir la cantidad que se utilizará para el estudio.

Selección del número de componentes principales

Según nos sugiere (Peña, 2002), para seleccionar el número de componentes principales a mantener se debe:

1. Realizar un gráfico donde se visualice cada componente (eje X) y su valor propio (eje Y). Se seleccionarán componentes hasta que en el

gráfico se visualice una especie de “codo”, es decir, hasta que los valores propios sean similares y bajos.

2. Seleccionar componentes hasta cubrir una proporción determinada de varianza (80% o 90%).
3. Definir una cota mínima para los valores propios, debajo de la cual se descartarán los componentes. Esta cota suele fijarse como la varianza media $\sum {}^{a_i}/_{p}$.

Ya seleccionado el número de componentes principales a estudiar, se suele utilizar gráficos denominados biplots, para la interpretación visual de los resultados.

Biplots

Los biplots son gráficos que nos permiten plasmar en una sola imagen tanto las observaciones como las variables estudiadas. Así como un diagrama de dispersión muestra la distribución conjunta de dos variables, los biplots representan a 3 o más variables. Dado que ya se cuenta con los componentes principales, para graficar el biplot se utilizarán los 2 primeros componentes como ejes X y Y. En el eje de las X estará el primer componente principal, es decir, aquel que explique la mayor cantidad de variación.

Las variables estudiadas se dibujan como vectores que parten del origen y las observaciones como puntos. Aquellos vectores que tengan direcciones similares tendrán mayor correlación, aquellos que tengan direcciones opuestas tendrán correlaciones altas negativas y aquellos que formen ángulos de 90° tendrán correlaciones cercanas a cero. Con respecto a los ejes, mientras mayor sea la longitud de un vector con respecto a un eje, mayor será su correlación, ya sea esta positiva o negativa. De contar con agrupaciones para los datos, también es posible visualizar estas agrupaciones en el biplot mediante elipses que agrupen un porcentaje predefinido de los datos.

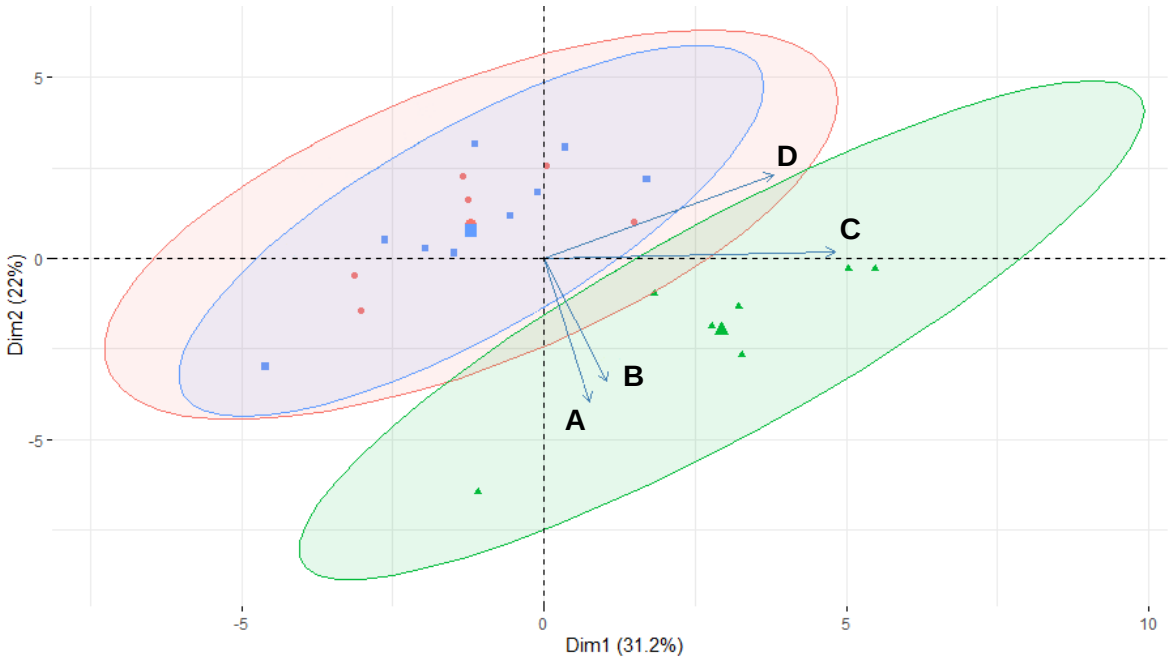


Figura 2. *Ejemplo biplot*

En la **Figura 2** se puede ver un ejemplo de un gráfico biplot. En este, la variable con mayor correlación positiva con respecto al primer componente es C. Esto quiere decir que a medida que este componente se incrementa, así mismo se incrementa C. Por otro lado, la variable con mayor correlación negativa con respecto al segundo componente es A, por lo que, a medida que este componente se incrementa, A disminuye. Las variables A y B presentan alta correlación, ya que el ángulo que forman es muy bajo, mientras la correlación entre A y D es casi cero, ya que el ángulo que forman sus vectores es de casi 90°. Siguiendo este criterio, de existir vectores que formen un ángulo de 180°, su correlación sería -1.

El mismo análisis aplica para los puntos graficados: un punto (una observación) ubicado a la derecha del gráfico tendrá gran correlación con el primer componente principal, así como con las variables correlacionadas con aquel componente.

2.6. Regresión Múltiple

La primera publicación sobre el uso de un método de regresión lineal corresponde a Legendré, en el año 1805, el cual correspondió al método de mínimos cuadrados. Sin embargo, quien introdujo el término *regresión* fue sir Francis Galton, médico y primo de Charles Darwin, quien, en 1886 en su artículo “Regression towards mediocrity in hereditary stature” (Regresión hacia la mediocridad en la estatura hereditaria), analizó la influencia de la estatura de los padres (variable predictora) sobre la estatura de sus descendientes (variable de respuesta).

Se utiliza una regresión cuando se está interesado en desarrollar un método de pronóstico de una variable respuesta, a partir de una o más variables predictoras que tienen alguna relación inherente entre sí. Cuando la relación entre las variables de respuesta (Y) y variables predictoras (X's) es exacta; es decir, una X siempre tendrá como resultado una misma Y, se la denomina una relación determinística. En muchos fenómenos de la vida real esta relación no es determinística, por lo que se los considera problemas de naturaleza probabilística. Teniendo esto en cuenta, el objetivo del análisis de regresión es encontrar la mejor relación entre X y Y, cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de Y, partiendo de los valores de X.

Cuando solo se tiene una variable de respuesta o regresor y la relación es lineal, el análisis se denomina regresión lineal simple, y se expresa mediante la siguiente ecuación:

$$\bar{Y} = \beta_0 + \beta_1 x + \mathcal{E}$$

Donde β_0 es la intersección de la recta con el eje Y y β_1 es su pendiente.

En la práctica, muchos problemas de investigación tienen un nivel de complejidad tal que, para poder realizar un modelo de regresión que pueda

predecir una respuesta precisa, requieren más de una variable predictora. Este tipo de análisis se denomina **regresión múltiple**. Cuando los valores de Y se generan a partir de una combinación lineal de los valores de las variables explicativas se la conoce como **regresión lineal múltiple** y se expresa como (Walpole, Myers, Myers, & Ye, 2012):

$$\bar{Y} = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_2 x + \mathcal{E}$$

Para definir los coeficientes de la ecuación, se busca que la suma de cuadrados de la diferencia entre los valores observados y los pronosticados por la ecuación sea lo más pequeña posible. En otras palabras, se va a minimizar la varianza residual. Para validar si un modelo es bueno, se utiliza un cociente denominado *coeficiente de determinación* (R^2), el cual nos indica cuánto de la variación total es explicada por el modelo, esto es:

$$R^2 = SEC/STC = 1 - SRC/STC$$

Donde *SEC* es la suma explicada de cuadrados, STC la suma total de cuadrados y SRC la suma residual de cuadrados. En caso de que se cuente con un modelo con ajuste perfecto, el valor de R^2 será de 1 y en caso de tener un modelo que no se ajusta en lo absoluto, su valor sería de 0.

Algo importante a considerar, es que R^2 incrementa a medida que se agregar variables predictoras al modelo, incluso cuando las variables agregadas tengan en realidad una contribución marginal. Por este motivo, muchas veces se ajusta el coeficiente de determinación utilizando la siguiente fórmula (Marco, s.f.):

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\right] * (1 - R^2)$$

Donde R_a^2 es el coeficiente ajustado, n es el número de observaciones de la muestra y k es el número de variables predictoras. Este coeficiente ajustado penaliza al R^2 a medida que se agregan variables al modelo.

Es posible encontrar varios modelos con R^2 similares, por lo que se requiere información adicional para definir cuál es mejor. Un criterio de selección es el

denominado criterio de información bayesiano (CIB), el cual se basa en la función de probabilidad. Como se indicó, al agregar parámetros al modelo se puede incrementar la probabilidad, pero resultar un sobreajuste. El CIB introduce un término de penalización para el número de parámetros utilizados. Está definido como (Wikipedia, 2020b):

$$BIC = k * \ln(n) - 2 * \ln(\hat{L})$$

Donde $\hat{L}$ es el valor maximizado de la función de verosimilitud del modelo, n es el tamaño de la muestra y *k* el número de parámetros estimados por el modelo.

Adicionalmente, para que un modelo sea adecuado se deben cumplir algunos supuestos estadísticos. Estos son (Santana, 2015):

- Existe *linealidad* entre la variable de respuesta y las variables predictoras. Esto quiere decir que un cambio unitario en *X* tiene el mismo efecto sobre *Y*.
- Los residuos son *independientes* y no están auto-correlacionados. Esto quiere decir que el residuo en la predicción de un valor no es afectado por la predicción del valor más cercano.
- La varianza de los residuos en las predicciones es constante, es decir, no varía a medida que se predicen valores mayores o menores. A esta constancia en los errores se la denomina *homocedasticidad*, mientras a la inconstancia se la denomina *heterocedasticidad*.
- Si se hiciera un histograma de frecuencias con los residuos de las predicciones se debe presentar una distribución *normal*. Esta normalidad se puede validar mediante la prueba de asimetría y kurtosis, prueba chi-cuadrado, prueba de residuos estandarizados, prueba de Kolmogorov-Smirnov-Liliefors, entre otras.

Hay ocasiones en las cuales un investigador se puede encontrar con un modelo irresoluble, por no poder determinar los coeficientes de la regresión. Uno se los motivos por lo que esto puede darse es debido a la colinealidad. La existencia de colinealidad se da cuando una variable predictora del modelo

es una combinación lineal de otra. En este mismo sentido, cuando se presenta una correlación fuerte entre dos o más variables predictoras se denomina multicolinealidad. La presencia de altos grados de correlación entre las variables predictoras hace que la varianza de los estimadores de pendiente (β) se incremente (Wooldridge, 2010).

Para determinar la severidad de la multicolinealidad, se puede utilizar un estadístico para coeficientes individuales denominado factor inflacionario de la varianza (FIV). El FIV para el coeficiente de pendiente j es $FIV_j = 1/(1 - R_j^2)$. En caso de encontrarse con un valor FIV mayor a 10, se puede decir que la multicolinealidad es alta (Kutner, Nachtsheim, & Neter, 2004).

Ahora que se cuenta con el marco teórico, se procederá a presentar la metodología con la que se realizó este trabajo.

CAPÍTULO 3

3. Metodología

Para el presente estudio se utilizaron datos obtenidos de diversas fuentes de información. Se obtuvieron datos generales sobre diversos indicadores de las provincias del país, entre los cuales están:

- Características de los hogares (características físicas inadecuadas, abastecimiento agua red pública, servicio eléctrico, alcantarillado, recolección de basura)
- Porcentaje de población por rangos de edad (0-14, 15-59, más de 60)
- Sobrepeso
- Nivel de educación
- Porcentaje de pobreza
- Servicios de salud (cantidad de respiradores, establecimientos médicos, médicos y enfermeras)
- Población, densidad poblacional

La información relacionada con la afectación de la COVID-19 fue obtenida de los Informes de Situación e Infografías emitidos por el Servicio Nacional de Gestión de Riesgos y Emergencias en su página web (Secretaría Nacional de Gestión de Riesgos y Emergencias, 2020):

- Casos confirmados
- Fallecidos por COVID-19

Con esta información, se calcularon los índices resaltados en el marco teórico:

- Tasa de mortalidad
- Tasa de letalidad
- Proporción de incidencia

El análisis univariante se realizó con las variables agrupadas de acuerdo con los criterios que representan:

- Variables de condiciones de vivienda: Incluye aquellas relacionadas con las condiciones de las viviendas. Estas variables son:
    - Hogares con características físicas inadecuadas
    - Viviendas con abastecimiento de agua de red pública
    - Viviendas con disponibilidad de servicio eléctrico
    - Viviendas con disponibilidad de alcantarillado
    - Viviendas disponen eliminación de basura mediante recolector
- Variables de condiciones socioeconómicas: Contiene variables que indican el rango de edad de la población, su nivel de educación:
    - Porcentaje población 0-14 años
    - Porcentaje población 15-59 años
    - Porcentaje población más de 60 años
    - Sobrepeso 19-59 años
    - Educación Nivel 1
    - Educación Nivel 2
    - Secundaria Terminada
    - Educación Superior
    - Porcentaje pobreza NBI
- Variables del sistema de salud: Contiene variables importantes para evaluar la capacidad de respuesta ante esta enfermedad:
    - Respiradores cada 10 mil habitantes
    - Establecimientos médicos cada 10 mil habitantes
    - Médicos y enfermeras cada 10 mil habitantes
- Variables de población: Variables relacionadas con la cantidad de habitantes y su ubicación geográfica:
    - Población
    - Agrupación población
    - Densidad poblacional
    - Región

- Variables de afectación: Corresponde a los indicadores epidemiológicos. Estos indicadores serán utilizados como variables dependientes en el estudio.
    - Tasa de mortalidad cada 10 mil habitantes
    - Tasa de letalidad
    - Proporción de incidencia

De cada variable mencionada anteriormente se obtuvo la media, mínimo, máximo, la desviación estándar y los cuartiles, además de que se obtuvo su histograma de frecuencias, para evaluar su distribución. En ciertos casos, se vio la necesidad de aplicarles logaritmo natural, para buscar normalizar su distribución. Esto se hizo siempre que la variable sea simétrica y unimodal, según la transformación de Box & Cox (1964):

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln x, \lambda = 0 \; y \; x > 0 \end{cases}$$

A continuación, se realizó un análisis bivariante, para evaluar la correlación de Pearson entre todas las variables independientes y dependientes. Se utilizó un gráfico tipo matriz en el cual se colocaron las variables en la diagonal principal. Sobre ésta se muestran las correlaciones entre las variables y debajo de la diagonal principal se muestran diagramas de dispersión de cada par de variables evaluadas.

Posteriormente, se procedió al análisis multivariante, realizando un análisis de componentes principales, para evaluar la influencia de cada variable independiente con respecto a las variables dependientes.

Finalmente, en este apartado se procuró obtener un modelo de regresión lineal múltiple no predictivo, que permita identificar las variables más influyentes para cada variable de respuesta.

Se utilizaron las metodologías Forward, Backward y Reemplazo Secuencial (combinación de las 2 anteriores según (Kassambara, 2018)) para determinar los predictores significativos de los modelos, para cada variable de respuesta. Una vez hecho esto, fue muy importante asegurarse de que los predictores sean independientes entre ellos, en otras palabras, que no presenten colinealidad.

Para determinar si existe colinealidad o multicolinealidad entre las variables predictoras de un modelo no se cuenta con un método estadístico en concreto. Lo que se recomienda es aplicar reglas prácticas que ayudan a determinar en qué medida se afecta el contraste y la estimación de un modelo. Los pasos que se pueden recomendar a seguir son:

- Si el coeficiente de determinación R^2 es elevado, pero ninguno de los predictores resulta significativo, puede haber indicios de colinealidad.
- Estudiar la relación lineal entre cada par de predictores mediante el cálculo de una matriz de correlación. Tener en cuenta que puede existir multicolinealidad, aunque no se obtenga ningún coeficiente de correlación elevado. Puede darse el caso de que las correlaciones simples entre algunos pares de variables no sean mayores que 0,5, pero exista una relación lineal casi perfecta entre tres o más de estas variables.
- Elaborar un modelo de regresión lineal simple entre cada uno de los predictores. Si en alguno de los modelos el R^2 conocido también como coeficiente de determinación es elevado, podría estar señalando una posible colinealidad.
- Factor de Inflación de la Varianza (FIV):
    - FIV = 1: Ausencia total de colinealidad
    - 1 < FIV < 5: La regresión puede ser afectada por cierta colinealidad.
    - 5 < FIV < 10: Existe una alta probabilidad de que exista colinealidad entre las variables.

Los datos fueron procesados mediante el software RStudio Versión 1.3.1073 y R versión 4.0.2.

CAPÍTULO 4

4. RESULTADOS

4.1. Análisis descriptivos

4.1.1 Evaluación condiciones de vivienda

Se consideraron dentro de este análisis 5 variables relacionadas con las condiciones de las viviendas en las diferentes provincias del país. Estas son:

$\boldsymbol{X_1}$**:** Hogares características físicas inadecuadas

$\boldsymbol{X_2}$**:** Viviendas abastecimiento agua red pública

$\boldsymbol{X_3}$**:** Viviendas disponibilidad servicio eléctrico

$\boldsymbol{X_4}$**:** Viviendas disponibilidad alcantarillado

$\boldsymbol{X_5}$**:** Viviendas disponen eliminación basura recolector

Tabla 1. *Medidas de centralidad, dispersión y posición para las variables que conforman la evaluación de las condiciones de vivienda*

var	media	mínimo	Máximo	desv.std.	q 0,25	q 0,5	q 0,75
$\boldsymbol{X_1}$	14,47	2,27	36,33	9,23	6,16	12,36	20,30
$\boldsymbol{X_2}$	59,55	38,82	82,60	13,79	49,37	57,30	70,50
$\boldsymbol{X_3}$	88,51	69,28	98,87	8,41	86,25	89,44	95,43
$\boldsymbol{X_4}$	37,55	13,92	69,48	13,98	31,38	35,63	45,37
$\boldsymbol{X_5}$	61,35	37,58	95,17	15,66	51,05	58,39	71,86

$\boldsymbol{X_1}$**:** Hogares características físicas inadecuadas

Indica el porcentaje de hogares que no cuentan con características físicas consideradas como adecuadas. Tal como se muestra en la **Tabla 1**, se encontró que, en promedio, casi un 15% del total de hogares por provincia tiene características inadecuadas. Las provincias con menor porcentaje de hogares con características físicas inadecuadas son Galápagos (2%), Napo (3%) y Pastaza (4%), mientras las que tienen el mayor porcentaje son

Manabí (27%), Chimborazo (28%) y Loja (36%). Su distribución se puede observar en la **Figura 3.a**.

X_2: Viviendas abastecimiento agua red pública

Como se puede ver en la **Tabla 1**, en promedio el 59,55% de viviendas de cada provincia cuenta con abastecimiento de agua pública. Las provincias con mayor porcentaje de abastecimiento de agua de red pública a sus viviendas son Santa Elena (79,5%), Carchi (81%) y Galápagos (82,6%). Las que tienen menor porcentaje son Orellana (38,8%), Manabí (41,4%) y Santo Domingo (41,4%). Su distribución se observa en la **Figura 3.b**.

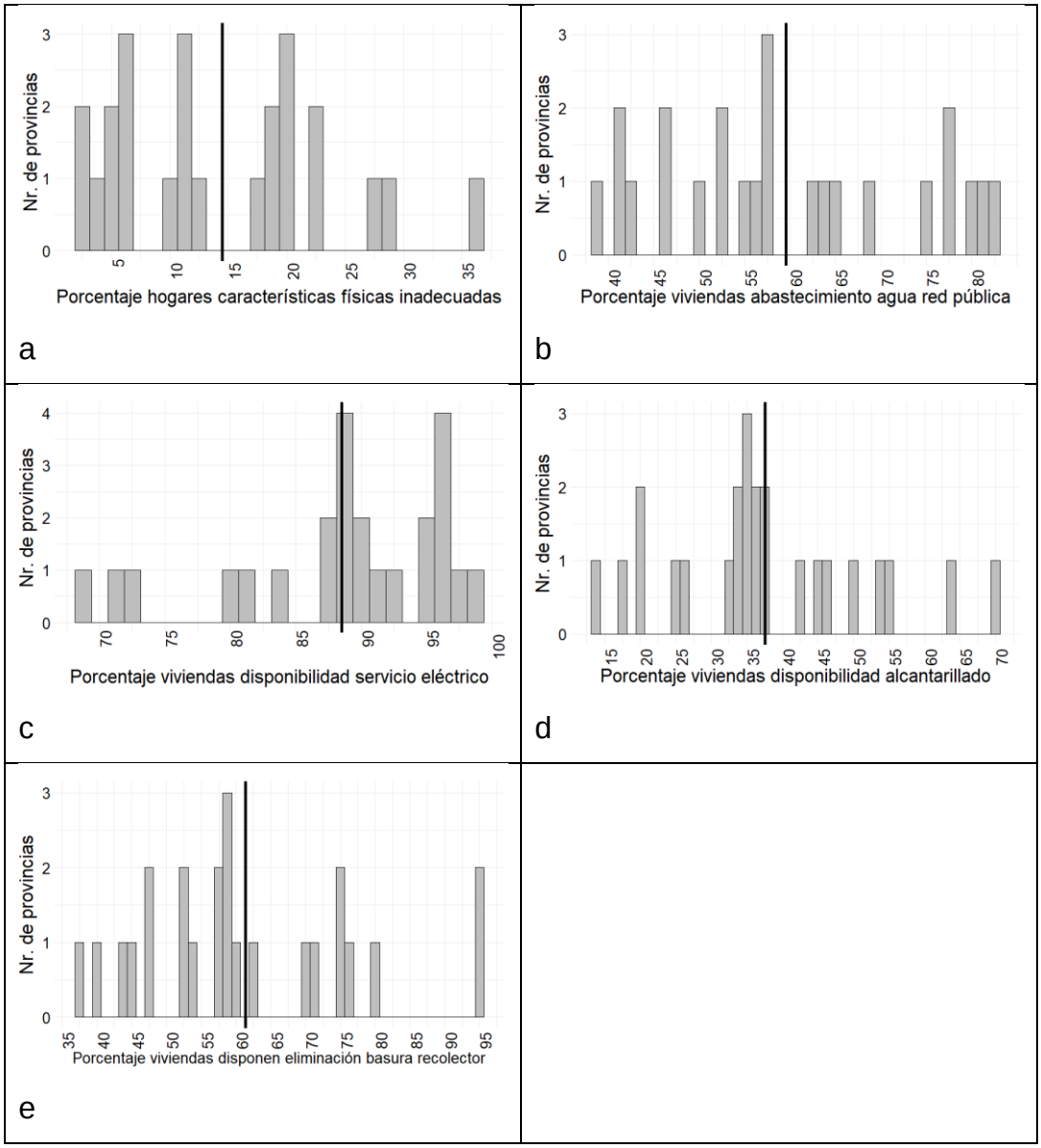


Figura 3. *Distribución de frecuencias de las variables que conforman la evaluación de las condiciones de vivienda*

Elaborado por: Daniel Salcedo

$\boldsymbol{X_3}$**:** Viviendas disponibilidad servicio eléctrico

En promedio, el 88,51% de viviendas tienen disponibilidad de servicio eléctrico (**Tabla 1**). Las provincias con porcentajes más altos son Imbabura (96,6%), Galápagos (97,6%) y Carchi (98,9%). Su distribución se encuentra en la **Figura 3.c**.

$\boldsymbol{X_4}$**:** Viviendas disponibilidad alcantarillado

La disponibilidad de alcantarillado es baja, con un promedio de apenas 37,55% y el 50% de los datos (cuartil 0,5) se encuentra incluso debajo del promedio (**Tabla 1**). Las provincias con porcentajes más bajos son Los Ríos (13,9%), Esmeraldas (17,6%) y Manabí (19,8%), mientras las que tienen mayor disponibilidad son El Oro (54,9%), Imbabura (63,5%) y Carchi (69,5%). Su distribución se encuentra en la Figura **3.d**.

$\boldsymbol{X_5}$**:** Viviendas disponen eliminación basura recolector

El 61,35% de viviendas disponen de servicio de recolección de basura, como se muestra en la **Tabla 1**. En cuanto a la distribución de los datos, llaman la atención 2 provincias con porcentajes considerablemente mayores a las demás: Santa Elena (94,8%) y Galápagos (95,2%). Las provincias con disponibilidad más baja son Loja (37,6%), Cotopaxi (39,4%), Chimborazo (44,3%). Su distribución se visualiza en la **Figura 3.e**.

4.1.2 Evaluación condiciones socioeconómicas

Este grupo de variables están relacionadas con la edad, educación y nivel de pobreza:

$\boldsymbol{X_6}$**:** Porcentaje población 0-14 años

$\boldsymbol{X_7}$**:** Porcentaje población 15-59 años

X_8: Porcentaje población más de 60 años

X_9: Sobrepeso 19-59 años

X_{10}: Educación Nivel 1

X_{11}: Educación Nivel 2

X_{12}: Secundaria Terminada

X_{13}: Educación Superior

X_{14}: Porcentaje pobreza NBI

Tabla 2. *Medidas de centralidad, dispersión y posición para las variables que conforman la evaluación de las condiciones socioeconómicas*

var	media	mínimo	máximo	desv.std.	q 0,25	q 0,5	q 0,75
X_6	0,31	0,26	0,38	0,04	0,28	0,31	0,34
X_7	0,59	0,54	0,63	0,02	0,57	0,59	0,61
X_8	0,10	0,06	0,13	0,02	0,09	0,11	0,12
X_9	0,63	0,52	0,70	0,05	0,59	0,64	0,67
X_{10}	0,23	0,09	0,33	0,06	0,20	0,23	0,28
X_{11}	0,40	0,32	0,49	0,04	0,37	0,40	0,42
X_{12}	0,17	0,13	0,26	0,04	0,14	0,17	0,20
X_{13}	0,19	0,11	0,34	0,06	0,15	0,17	0,22
X_{14}	0,30	0,09	0,48	0,10	0,23	0,31	0,38

X_6: Porcentaje población 0-14 años

El porcentaje de población entre 0 y 14 años es en promedio de 30,9%, como lo muestra la **Tabla 2**. Las provincias con porcentajes más bajos de población de 0 a 14 años son Pichincha y Tungurahua con 26% y aquellas con porcentajes más altos son Zamora Chinchipe y Napo (36%), Morona Santiago (37%) y Orellana (38%). Su distribución se puede ver en la **Figura 4.a**.

X_7: Porcentaje población 15-59 años

Como indica la **Tabla 2**, el promedio de población entre 15 y 59 años es del 59%. Tal como se puede ver en la **Figura 4.b**, la distribución de datos es asimétrica, con un valor separado del resto por debajo de la media (Bolívar, 54%) y mayores concentraciones sobre la media. Las provincias con

porcentajes más altos son El Oro, Galápagos y Guayas con 62%, y Pichincha con 63%.

X_8: Porcentaje población más de 60 años

El promedio de población mayor a 60 años es del 10% (ver **Tabla 2**).

La distribución (**Figura 4.c**) también es asimétrica, con una cola más larga debajo de la media y mayores frecuencias sobre la media.

Las provincias con porcentajes más altos son Chimborazo (12,6%), Tungurahua (12,7%) y Loja (13,3%). Las provincias con porcentajes más bajos son Orellana (6,2%), Morona Santiago (7%), Napo y Sucumbíos (7,4%).

X_9: Sobrepeso 19-59 años

El promedio de sobrepeso entre los 19 y 59 años se encuentra en 62,6% (ver **Tabla 2**).

Como se aprecia en la **Figura 4.d**, su distribución de frecuencias es asimétrica hacia la derecha, siendo las provincias con porcentajes más altos Galápagos (67,8%), Manabí (69%) y Carchi (70%), y con porcentajes más bajos Napo (52,4%), Cotopaxi (52,7%) y Orellana (54,9%).

X_{10}: Educación Nivel 1

Educación nivel 1 se refiere a las personas que no tienen educación o que han terminado la educación primaria. En promedio, el 23% de la población está en el nivel 1 de educación (**Tabla 2**).

La distribución de esta variable se encuentra en la **Figura 4.e**. Las provincias con menor porcentaje de personas con educación nivel 1 son Galápagos (9%), Pichincha (13%) y Guayas (17%), mientas las que tienen mayor porcentaje son Cotopaxi (31%), Bolívar (32%) y Cañar (33%).

$\boldsymbol{X_{11}}$**:** Educación Nivel 2

Educación nivel 2 corresponde a las personas que han terminado la educación primaria, pero no terminaron la educación secundaria. Como indica la **Tabla 2**, el promedio de personas con educación nivel 2 es del 40%. La distribución de esta variable se puede ver en la **Figura 4.f**. Las provincias con menor porcentaje son Pichincha (32%), Chimborazo (35%), Esmeraldas y Galápagos (36%). Las provincias con mayor porcentaje son Orellana (46%), Zamora Chinchipe (47%) y Santa Elena (49%).

$\boldsymbol{X_{12}}$**:** Secundaria Terminada

El promedio de personas con educación secundaria terminada es del 17% (**Tabla 2**). Su distribución se puede ver en la **Figura 4.g**.

Las provincias con menor porcentaje son Cañar, Chimborazo, Cotopaxi y Loja (13%), mientras la provincia con mayor porcentaje es Galápagos con 26%

$\boldsymbol{X_{13}}$**:** Educación Superior

El promedio de personas con educación superior es del 19% (ver **Tabla 2**). Su distribución se encuentra en la **Figura 4.h**.

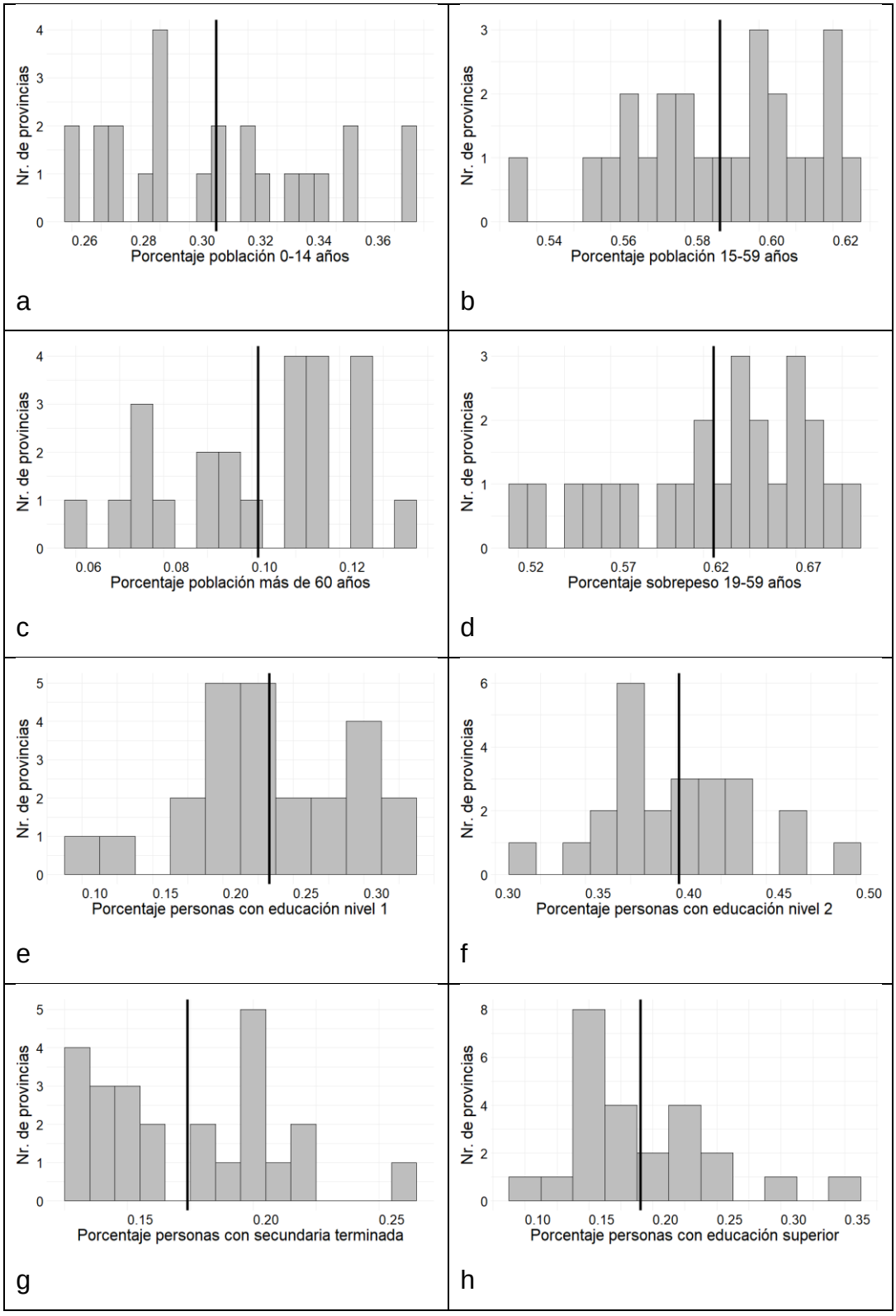
Las provincias con porcentajes más bajos son Sucumbíos (11%) y Orellana (12%), y las provincias con porcentajes más altos son Galápagos (29%) y Pichincha (34%).

$\boldsymbol{X_{14}}$**:** Porcentaje pobreza NBI

El promedio de porcentaje de pobreza por necesidades básicas insatisfechas es del 29,6% (ver **Tabla 2**).

La distribución es asimétrica, con mayores frecuencias en valores sobre la media (ver **Figura 4.i**). Las provincias con valores más bajos son Pichincha

(8,5%), Galápagos (12,2%) y Azuay (14,8%). Las provincias con valores más altos son Bolívar (40,6%), Orellana (44%) y Morona Santiago (47,9%).



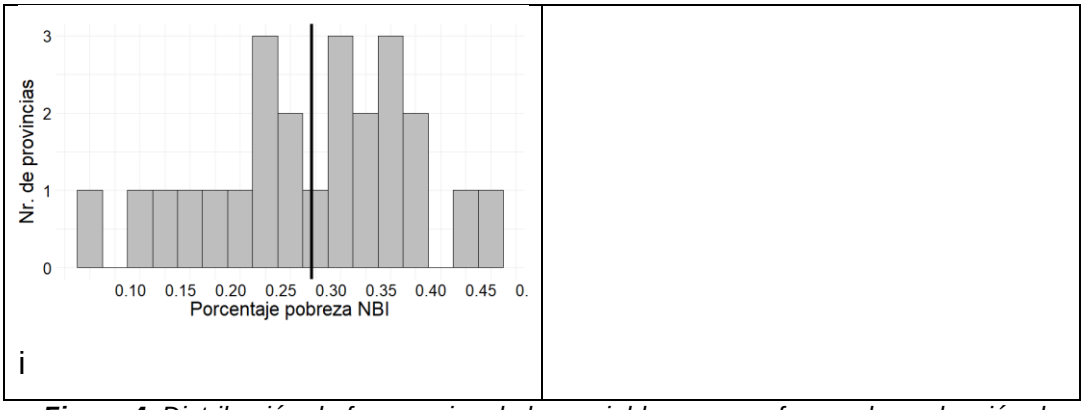


Figura 4. *Distribución de frecuencias de las variables que conforman la evaluación de las condiciones socioeconómicas*

Elaborado por: Daniel Salcedo

4.1.3 Evaluación sistema de salud

Este grupo de variables mide varios parámetros relacionados con el sistema de salud de cada provincia:

$\boldsymbol{X_{15}}$**:** Respiradores cada 10.000 habitantes

$\boldsymbol{X_{16}}$**:** Establecimientos médicos cada 10.000 habitantes

$\boldsymbol{X_{17}}$**:** Médicos y enfermeras cada 10.000 habitantes

Tabla 3. *Medidas de centralidad, dispersión y posición para las variables que conforman la evaluación del sistema de salud*

var	media	mínimo	máximo	desv.std.	q 0,25	q 0,5	q 0,75
$\boldsymbol{X_{15}}$	1,00	0,24	2,30	0,67	0,47	0,69	1,66
$\boldsymbol{X_{16}}$	3,44	1,54	6,11	1,35	2,51	3,04	4,27
$\boldsymbol{X_{17}}$	35,09	19,64	50,11	8,39	30,54	33,52	41,11

$\boldsymbol{X_{15}}$**:** Respiradores cada 10.000 habitantes

Viendo la **Tabla 3**, se podría llegar a la conclusión de que, en promedio, cada provincia cuenta con un respirador cada 10 mil habitantes. Sin embargo, la distribución de estos datos nos muestra una realidad diferente (ver **Figura 5.a**.

La distribución tiene una clara asimetría, con frecuencias más altas en valores sobre la media. Un cuarto de las provincias tiene entre 0,2 y 0,47 respiradores cada 10 mil habitantes y la mitad llega apenas a 0,7. Las provincias con menor cantidad de respiradores son Bolívar (0,2), Napo (0,3) y Cañar (0,3), mientras aquellas con mayor cantidad son Galápagos (2,1), Guayas (2,2) y Sucumbíos (2,3).

X_{16}: Establecimientos médicos cada 10.000 habitantes

En promedio, cada provincia cuenta con 3,4 establecimientos médicos cada 10 mil habitantes (ver **Tabla 3**). La distribución de estos datos se puede ver en la **Figura 5.b**.

Las provincias con menor cantidad de establecimientos médicos cada 10 mil habitantes son Guayas (1,5), Pichincha (1,94) y Santo Domingo (2,09). Las provincias con mayor cantidad son Pastaza (5,78), Zamora Chinchipe (5,90) y Morona Santiago (6,11).

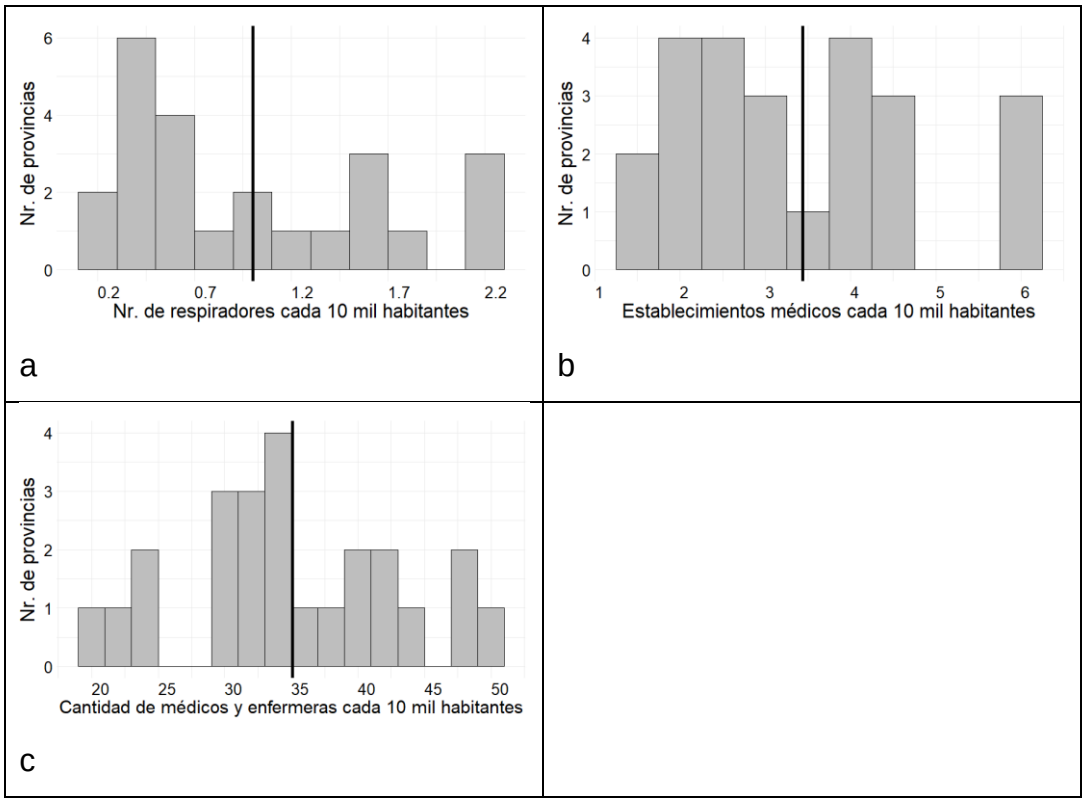


Figura 5. *Distribución de frecuencias de las variables que conforman la evaluación del sistema de salud*

Elaborado por: Daniel Salcedo

$\boldsymbol{X_{17}}$**:** Médicos más enfermeras cada 10.000 habitantes

Cada provincia cuenta en promedio con 35 médicos y enfermeras cada 10 mil habitantes (ver **Tabla 3**).

Las provincias con menor cantidad son Santa Elena (19,6), Los Ríos (21,2) y Cotopaxi (24,12). Las provincias con mayor cantidad son Pichincha (48,2), Pastaza (48,4) y Napo (50,1). La distribución de esta variable se encuentra en la **Figura 5.c**.

4.1.4 Evaluación de variables de población

Este grupo de variables mide la cantidad de habitantes en cada provincia, así como su densidad poblacional:

$\boldsymbol{X_{18}}$**:** Población

$\boldsymbol{X_{19}}$**:** Agrupación población

$\boldsymbol{X_{20}}$**:** Densidad poblacional

$\boldsymbol{X_{21}}$**:** Región

Tabla 4. *Medidas de centralidad, dispersión y posición para las variables que conforman la evaluación las variables relacionadas con la población*

var	media	mínimo	máximo	desv.std.	q 0,25	q 0,5	q 0,75
$\boldsymbol{X_{18}}$	727.864	33.042	4.387.434	1.021.291	194.119	482.487	622.250
$\boldsymbol{\ln(X_{18})}$	12,90	10,41	15,29	1.091,00	12,18	13,09	13,34
$\boldsymbol{X_{20}}$	85,24	3,93	333,10	83,37	12,14	78,50	108,80
$\boldsymbol{\ln(X_{20})}$	3,84	1,37	5,81	1,31	2,50	4,36	4,69

$\boldsymbol{X_{18}}$**:** Población

Como se puede ver en la **Tabla 4**, el promedio de población de cada provincia es de 727.864 habitantes. Sin embargo, se debe resaltar que el 75% de las provincias (18 de 24) tienen una población por debajo de 630 mil

personas. La distribución de los datos es marcadamente asimétrica, como se puede ver en la **Figura 6.a**.

Galápagos es la provincia con menor población (33 mil personas), mientras las provincias con mayor población son Pichincha (3,2 millones) y Guayas (4,4 millones).

Debido a la alta variabilidad de los datos de x18, se decide trabajar con $\mathbf{ln}(\boldsymbol{X_{18}})$. Su distribución se encuentra en la **Figura 6.b**.

$\boldsymbol{X_{19}}$**:** Agrupación población

El 38% (9) de las provincias tiene menos de 250 mil habitantes, el 42% (10) tiene entre 250 mil y 750 mil, y el 21% (5) tiene más de 750 mil habitantes (ver **Figura 7.a**).

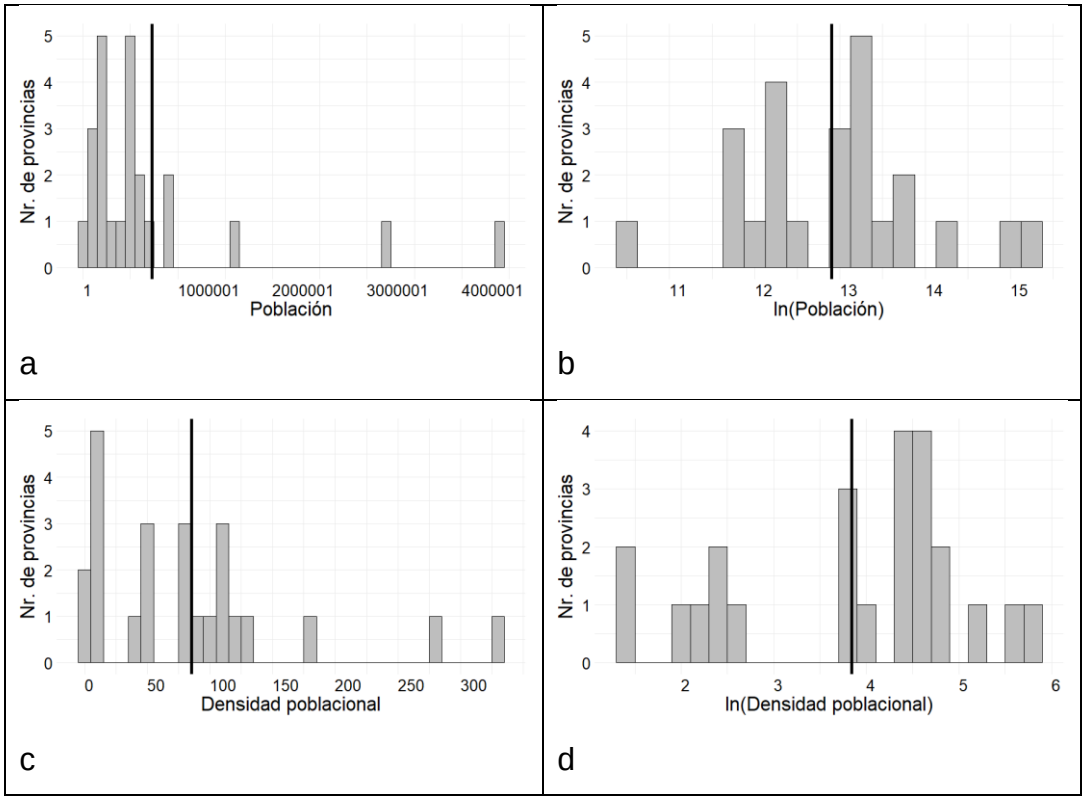


Figura 6. *Distribución de frecuencias de las variables que conforman la evaluación de las variables de población*

Elaborado por: Daniel Salcedo

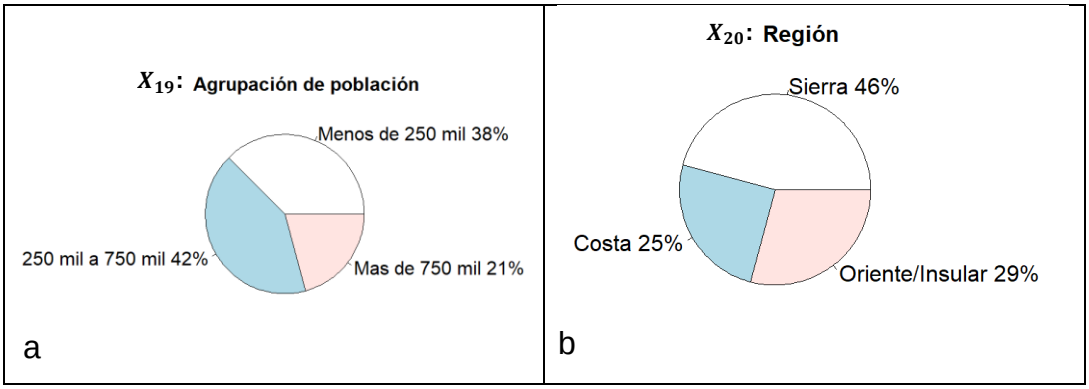


Figura 7. *Variables para agrupaciones*

X_{20}: Densidad poblacional

La densidad poblacional promedio es de 82,42 habitantes por kilómetro cuadrado, como se puede apreciar en la **Tabla 4**.

Su distribución muestra una clara asimetría negativa, con una larga cola hacia el lado derecho de la gráfica, lo cual se refleja en el último cuartil de los datos, los cuales abarcan desde 108,8 hasta 333 habitantes por kilómetro cuadrado (ver **Figura 6.c**).

Las provincias con menor densidad poblacional son Pastaza (3,9), Galápagos (4,1) y Morona Santiago (8,2), mientras que las provincias con mayor densidad son Tungurahua (183,3), Guayas (275,5) y Pichincha (333,1).

Debido a la alta varianza de la densidad poblacional, se decide también trabajar con $\mathbf{ln}(X_{20})$. Su distribución se puede visualizar en la **Figura 6.d**.

X_{21}: Región

El 45,8% de las provincias (11) corresponden a la región sierra, el 25% (6) a la región costa y el 29,2% (7) a las regiones oriente o insular (ver **Figura 7.b**).

4.1.5 Evaluación de indicadores de afectación

Esta agrupación corresponde a las variables de respuesta, las cuales indican el impacto que ha tenido la pandemia en cada provincia:

$\boldsymbol{Y_1}$**:** Tasa mortalidad 10000 hab 31.05

$\boldsymbol{Y_2}$**:** Tasa letalidad 31.05

$\boldsymbol{Y_3}$**:** Proporción incidencia 31.05

Tabla 5. *Medidas de centralidad, dispersión y posición para las variables que conforman la evaluación de los indicadores de afectación*

var	media	mínimo	máximo	desv.std.	q 0,25	q 0,5	q 0,75
$\boldsymbol{Y_1}$	1,35	0,15	7,00	1,49	0,40	1,09	1,61
$\boldsymbol{\ln(Y_1)}$	-0,15	-1,88	1,95	0,98	-0,93	0,08	0,46
$\boldsymbol{Y_2}$	0,10	0,01	0,36	0,08	0,04	0,08	0,13
$\boldsymbol{Y_3}$	0,13	0,04	0,32	0,07	0,08	0,14	0,16

$\boldsymbol{Y_1}$**:** Tasa de mortalidad 10 mil habitantes

La tasa de mortalidad indica la relación entre la cantidad de fallecidos y la cantidad de personas en una población y en un periodo de tiempo. En este caso, se utiliza el indicador cada 10 mil habitantes, debido a su baja magnitud.

La tasa de mortalidad promedio es de 1,35, con un mínimo de 0,15 y un máximo de 7 fallecidos cada 10 mil habitantes, como se puede ver en la **Tabla 5**. La distribución de estos datos se puede ver en la **Figura 8.a**.

Las provincias con tasa de mortalidad más baja cada 10 mil habitantes son Morona Santiago (0,15), Sucumbíos (0,22) y Orellana (0,25), mientras aquellas con tasa más alta son Manabí (2,78), Guayas (3,20) y Santa Elena (7,00).

Esta variable también demuestra una alta varianza, por lo que se decide trabajar con $\boldsymbol{\ln(Y_1)}$, cuya distribución se puede ver en la **Figura 8.b**.

$\boldsymbol{Y_2}$**:** Tasa de letalidad

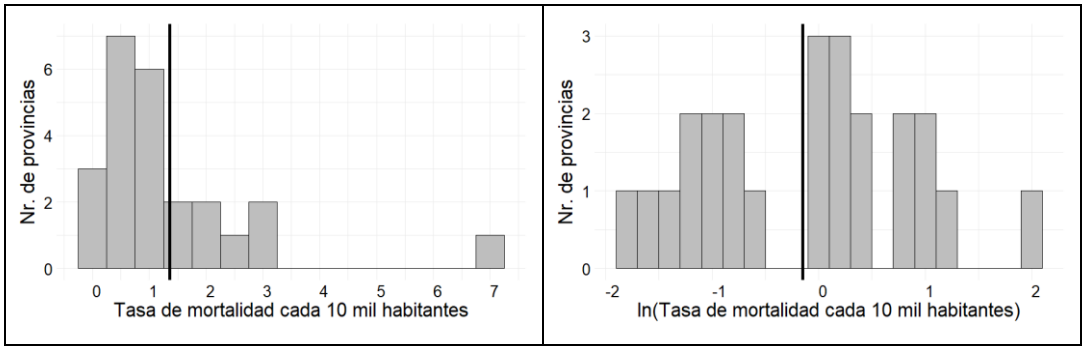
La tasa de letalidad indica la relación entre fallecidos y cantidad de personas que contrajeron la enfermedad en una población y en un período de tiempo. De acuerdo con la **Tabla 5**, en promedio la tasa de letalidad es de 0,1, es decir, una de cada 10 personas que contraen la enfermedad, falleció.

La distribución de los datos tiene una clara asimetría con tendencia a mayores frecuencias por debajo de la media (ver **Figura 8.c**). Las provincias con tasa de letalidad más baja son Galápagos (0,01) y Orellana (0,02), mientras aquellas con tasa más alta son Chimborazo (0,27) y Santa Elena (0,36).

$\boldsymbol{Y_3}$**:** Proporción de incidencia

La proporción de incidencia nos indica la relación entre los casos confirmados y la población. En otras palabras, nos indica qué tanto se ha esparcido la enfermedad. La proporción de incidencia promedio entre todas las provincias es del 0,13%, con un mínimo de 0,04% y un máximo de 0,32% (ver **Tabla 5**). La distribución de los datos muestra asimetría hacia la izquierda, como se puede ver en la **Figura 8.d**.

Las provincias con menor proporción de incidencia son Imbabura (0,04%), Morona Santiago y Sucumbíos (0,05%). Las provincias con mayor proporción de incidencia son Santo Domingo (0,20%), Galápagos (0,23%) y Guayas (0,32%).



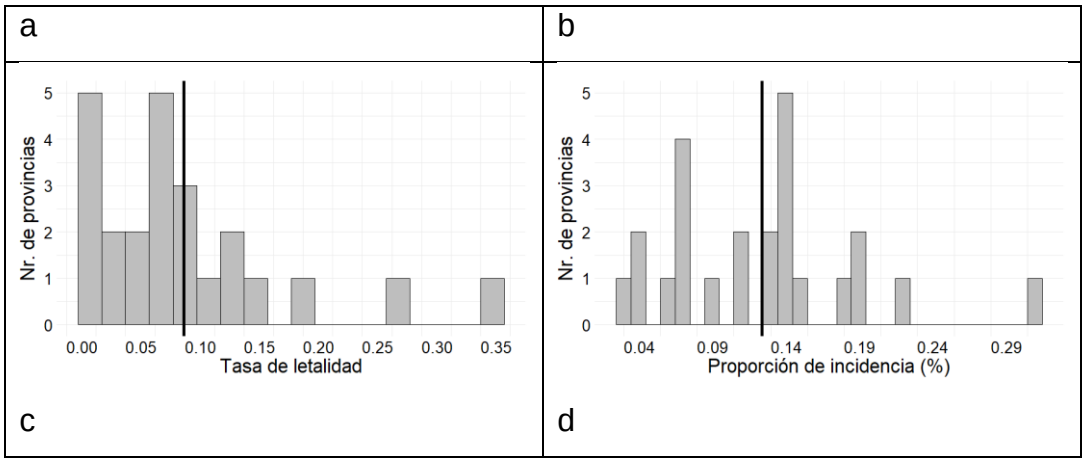


Figura 8. *Distribución de frecuencias de las variables que conforman la evaluación de los indicadores de afectación*

Elaborado por: Daniel Salcedo

4.2. Análisis bivariante

En este apartado se realizó el análisis bivariante de las variables, mediante el cual se determinó, entre otras cosas, si existe correlación entre alguna de las variables predictoras X con al menos una de las Y.

Las variables que muestran correlación con respecto a $\ln(Y_1)$ (Tasa de mortalidad cada 10 mil habitantes) son:

Tabla 6. *Correlación con respecto a tasa de mortalidad*

Variable	**Correlación (valor-p)**
X_8: Porcentaje población más de 60 años	0,35 (0.093)
X_{16}: Establecimientos médicos cada 10 mil habitantes	-0,62 (0.001)
X_{17}: Médicos + enfermeras cada 10 mil habitantes	-0,43 (0.036)
$\ln(X_{18})$: Población	0,53 (0.007)
$\ln(X_{20})$: Densidad poblacional	0,64 (0.001)

Las variables que muestran correlación con respecto a Y_2 (Tasa de letalidad) son:

Tabla 7. *Correlación con respecto a tasa de letalidad*

Variable	**Correlación (valor-p)**
X_1: Hogares características físicas inadecuadas	0,35 (0.012)
X_{16}: Establecimientos médicos cada 10 mil habitantes	-0,49 (0.015)
X_{17}: Médicos + enfermeras cada 10 mil habitantes	-0,58 (0.003)
$\ln(X_{18})$: Población	0,36 (0.080)
$\ln(X_{20})$: Densidad poblacional	0,51 (0.010)

Por otro lado, las variables que muestran correlación con respecto a Y_3 (Proporción de incidencia) son:

Tabla 8. *Correlación con respecto a proporción de incidencia*

Variable	**Correlación (valor-p)**
X_5: Hogares características físicas inadecuadas	0,43 (0.038)
X_{10}: Establecimientos médicos cada 10 mil habitantes	-0,47 (0.021)
X_{12}: Médicos + enfermeras cada 10 mil habitantes	0,54 (0.007)

En las **Figuras 9 y 10** a continuación, se puede ver en la diagonal principal las diferentes variables predictoras y de respuesta. Sobre la diagonal se muestran las correlaciones entre cada par de variables y debajo de la diagonal se muestran sus gráficos de dispersión. Debido a la cantidad de variables, se dividieron las variables predictoras en 2 grupos, danto como resultado las 2 figuras mencionadas:

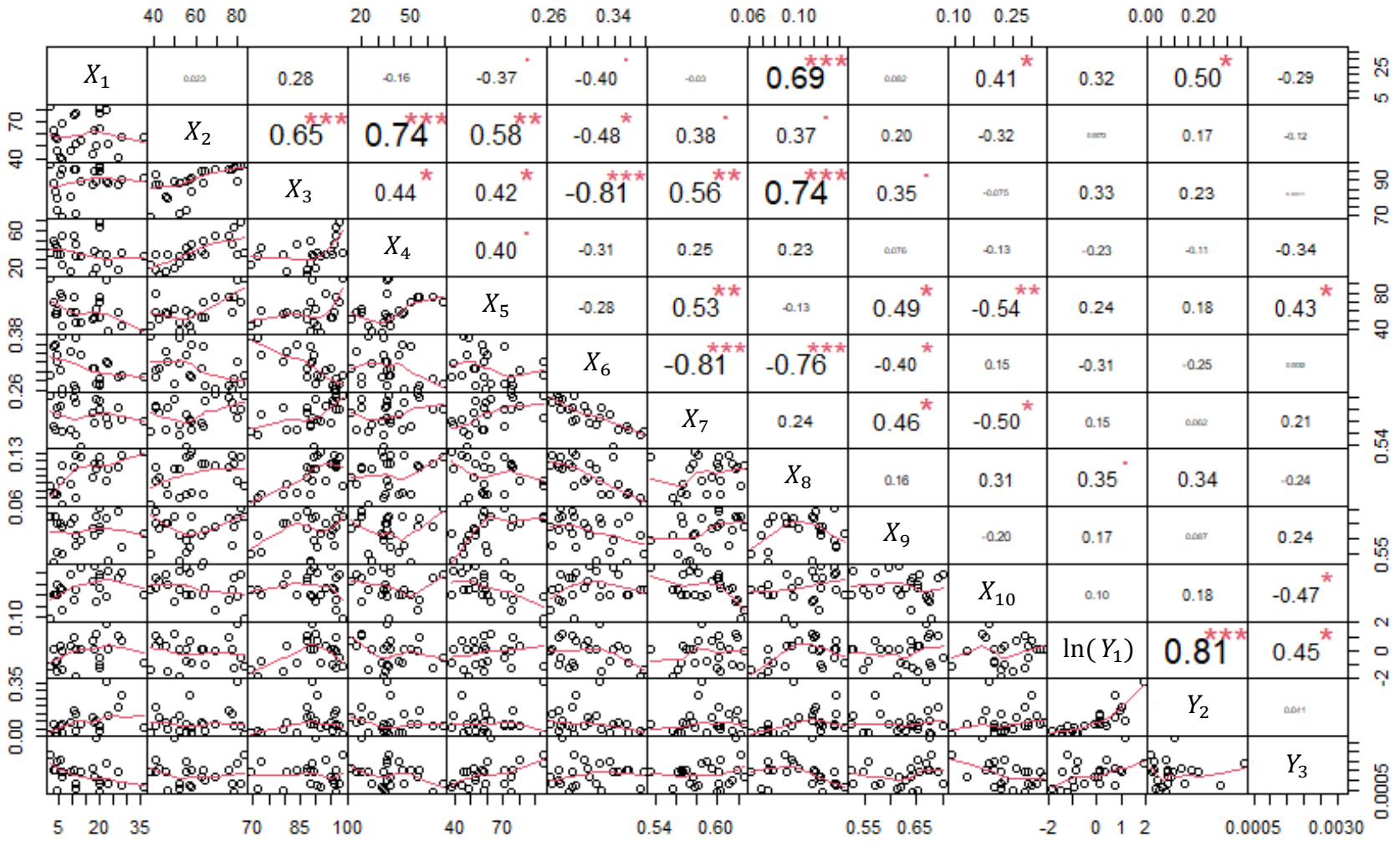


Figura 9. *Gráfico de correlación entre variables*

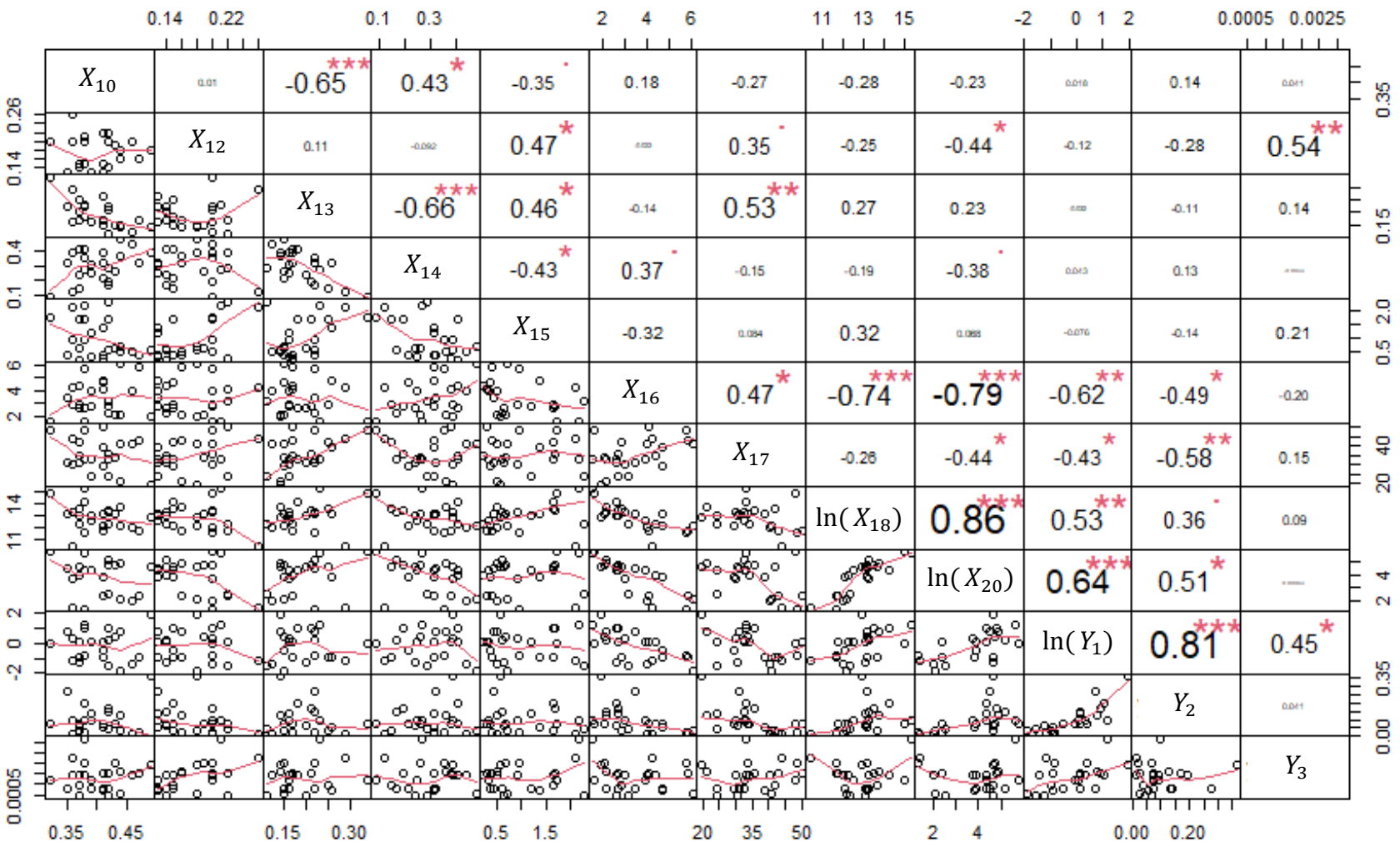


Figura 10. Gráfico de correlación entre variables

En las **Figuras 9 y 10** se puede observar que también se presenta una alta correlación entre las variables Y_2 (Tasa de letalidad) y Y_3 (Proporción de incidencia) con $\ln Y_1$ (Tasa de mortalidad):

Tabla 9. *Correlación entre variables de respuesta*

Variable	**Correlación (valor-p)**
Y_2: Tasa de letalidad cada 10 mil habitantes	0,81 ($1.8e^{-6}$)
Y_3: Proporción de incidencia	0,45 (0.027)

Se muestra a continuación un nuevo gráfico (**Figura 11**), esta vez incluyendo únicamente las variables predictoras que tienen correlación con las variables de respuesta:

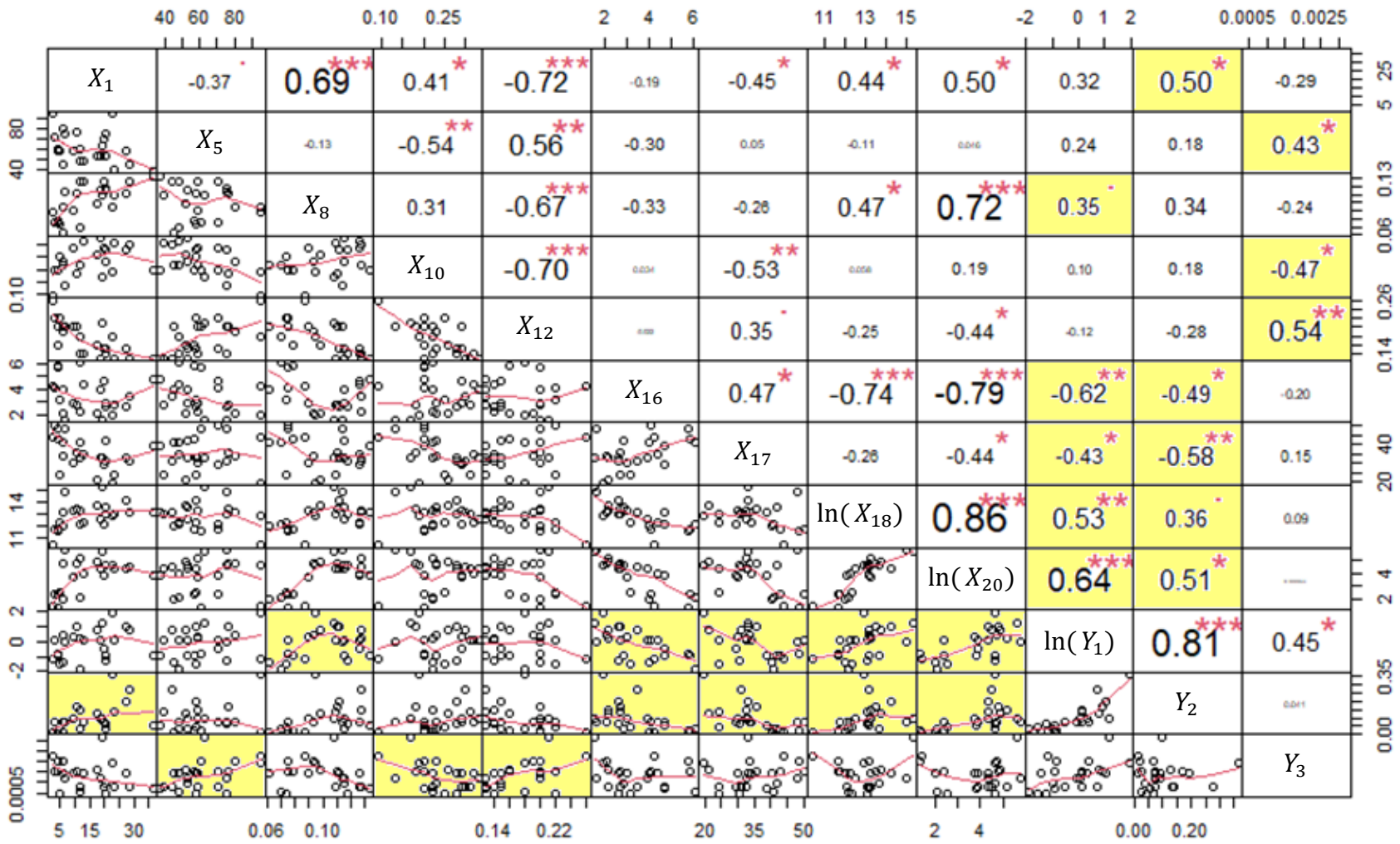


Figura 11. *Gráfico de correlación entre variables*

4.3. Análisis multivariante

Se realizó un análisis de componentes principales con todos los datos, excepto X_{19}: Agrupación de población y X_{21}: Región, dado que son factores. El resultado fue el siguiente:

Tabla 10. *Componentes principales - Importancia de los componentes*

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Desv. Std.	2,620	2,201	1,748	1,483	1,124	0,949	0,863	0,808
Proporción de var.	0,312	0,220	0,139	0,100	0,057	0,041	0,034	0,030
Proporción acum.	0,312	0,532	0,671	0,771	0,828	0,869	0,903	0,933
	PC9	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**	**PC16**
Desv. Std.	0,688	0,504	0,425	0,414	0,387	0,287	0,243	0,225
Proporción de var.	0,022	0,012	0,008	0,008	0,007	0,004	0,003	0,002
Proporción acum.	0,954	0,966	0,974	0,982	0,989	0,992	0,995	0,998
	PC17	**PC18**	**PC19**	**PC20**	**PC21**	**PC22**		
Desv. Std.	0,168	0,122	0,106	0,035	0,022	0,000		
Proporción de var.	0,001	0,001	0,001	0,000	0,000	0,000		
Proporción acum.	0,999	0,999	1,000	1,000	1,000	1,000		

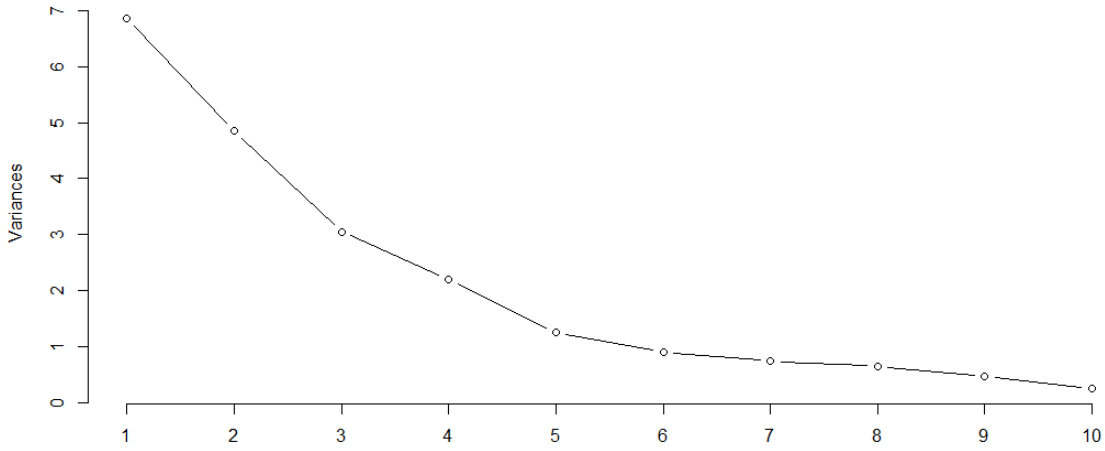


Figura 12. *Análisis de componentes principales*

Los dos primeros componentes explican alrededor del 53% de la variabilidad de los datos. Para obtener al menos un 80% de la explicación, se requerirían 5 componentes.

A continuación, se utilizaron los dos primeros componentes para realizar biplots, agrupando los datos por X_{19}: Agrupación de población y luego por X_{21}: Región.

Primero se analizarán las variables, sin considerar los elipsoides. En las **Figuras 13 y 14** se puede ver que hay variables que no están muy representadas por los 2 componentes principales, como X_4 y Y_3. Esto se evidencia en la poca longitud de los vectores. Analizando las variables predictoras, se puede ver que X_6, X_{14} y X_{16} tienen la mayor correlación positiva con respecto al primer componente principal, mientras X_3, X_7, X_8, X_{13}, $\ln(X_{18})$ y $\ln(X_{20})$ tienen la mayor correlación negativa. Con respecto al segundo componente, las variables con mayor correlación positiva son X_1 y X_{10}, y aquellas con mayor correlación negativa son X_{12} y X_{17}.

En cuando a las variables de respuesta, Y_1 y Y_2 presentan correlación negativa con respecto al primer componente y correlación positiva con respecto al segundo. Se confirma además que estas dos variables presentan alta correlación entre sí. Como se indicó en el párrafo anterior, Y_3 no está muy representada por ninguno de los dos componentes principales, sin embargo, se podría decir que su relación con Y_1 y Y_2 es baja, ya que forman un ángulo de casi 90°.

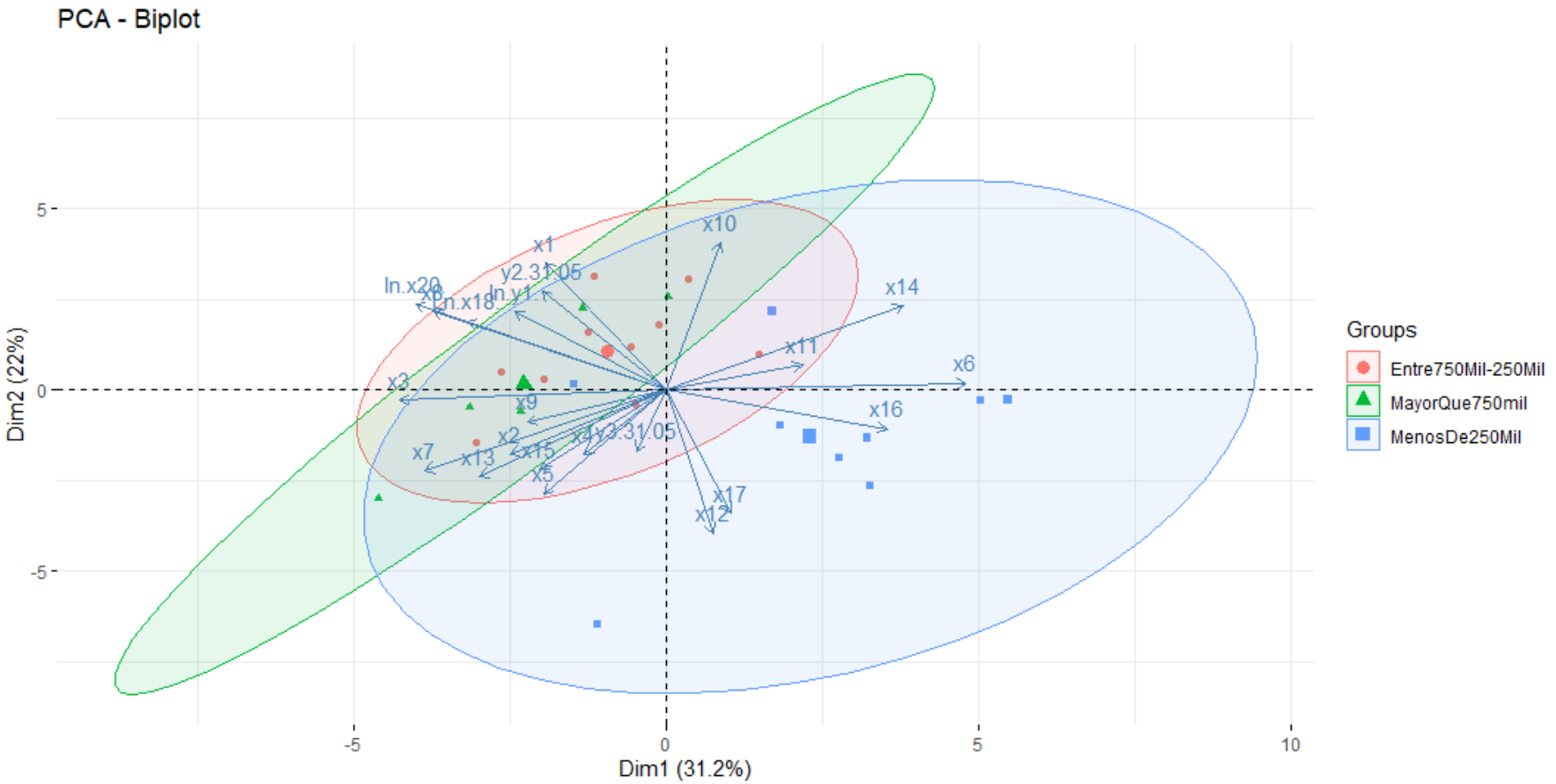


Figura 13. *Biplot agrupación población*

Pasaremos ahora al análisis de los datos agrupados, empezando por la agrupación por $\boldsymbol{X_{19}}$**: *Agrupación población*** (**Figura 13**). Las provincias con menos de 250 mil habitantes forman un grupo muy grande y que ocupa los 4 cuadrantes del gráfico. Esto no facilita relacionar la correlación de las observaciones contenidas en este grupo con respecto a las variables estudiadas ni a los componentes principales. Sin embargo, se puede decir que se evidencia cierta tendencia de esta agrupación hacia el cuadrante inferior derecho del Biplot (correlación positiva con el primer componente, negativa con el segundo), mientras que la dirección de las variables Y_1 y Y_2 es hacia el cuadrante superior izquierdo (correlación negativa con el primer componente, positiva con el segundo). Esto indica la influencia inversa de esas variables sobre el grupo. Teniendo esto en cuenta, podríamos afirmar que, aunque esta agrupación presenta observaciones de todo tipo, en ella están mayormente las provincias menos afectadas, ya que en su mayoría no son representadas por Y_1 y Y_2.

Las provincias con población entre 250 mil y 750 mil habitantes, muestran más homogeneidad, ya que forman una figura más pequeña y con cierta tendencia (elipsoide rojo en la **Figura 13**). Por su ubicación principalmente en el cuadrante superior izquierdo (correlación negativa con el primer componente principal y positiva con el segundo), se puede inferir que este grupo está fuertemente caracterizado por las variables de respuesta Y_1, Y_2. En cuanto a las variables predictoras, el grupo está caracterizado por X_1, X_8, $\ln X_{20}$ y $\ln X_{18}$ (son las que más han influido para que las provincias con esta población sean afectadas), mientras que las variables X_{12}, X_{17} y X_{16}, son las que desfavorecieron el incremento en las variables Y_1 y Y_2 (actuaron para contener la epidemia). En la **Tabla 11** se detallan las relaciones entre las principales variables predictoras, en relación con las variables de respuesta.

Tabla 11. *Relación entre variables predictoras y variables de respuesta Y_1 y Y_2, para agrupación de población entre 250 mil y 750 mil*

Variable	Descripción	Tipo de relación con Y_1 y Y_2
X_1	Hogares características físicas inadecuadas	Directamente proporcional
X_3	Viviendas disponibilidad servicio eléctrico	Directamente proporcional
X_8	Porcentaje población más de 60 años	Directamente proporcional
X_{10}	Educación Nivel 1	Directamente proporcional
X_{12}	Secundaria Terminada	Inversamente proporcional
X_{16}	Establecimientos médicos cada 10 mil hab.	Inversamente proporcional
X_{17}	Médicos+enfermeras cada 10 mil hab.	Inversamente proporcional
$\ln X_{18}$	Población	Directamente proporcional
$\ln X_{20}$	Densidad poblacional	Directamente proporcional

Algo similar sucede con las provincias con población mayor que 750 mil habitantes (elipsoide verde en la **Figura 13**). Este grupo también está caracterizado tanto por Y_1 y Y_2. No es posible caracterizar al grupo con respecto a Y_3, ya que hay tanto observaciones caracterizadas (ubicadas en los cuadrantes superior derecho e inferior izquierdo) como no caracterizadas (cuadrante superior izquierdo) por ésta. Al igual que con la agrupación entre 250 y 750 mil habitantes, se mantiene la fuerte relación positiva de con las variables Y_1 y Y_2 con X_1, X_8, $\ln X_{20}$ y $\ln X_{18}$ (han influido para que las provincias de este grupo sean más afectadas), y se tiene relación inversa con X_6, X_{14} y X_{16} (han contribuido para una menor afectación).

Tabla 12. *Relación entre variables predictoras y variables de respuesta Y_1 y Y_2, para agrupación de población mayor que 750 mil*

Variable	Descripción	Tipo de relación con Y_1 y Y_2
X_1	Hogares características físicas inadecuadas	Directamente proporcional
X_3	Viviendas disponibilidad servicio eléctrico	Directamente proporcional
X_6	Porcentaje población 0-14 años	Inversamente proporcional
X_7	Porcentaje población 15-59 años	Directamente proporcional
X_8	Porcentaje población más de 60 años	Directamente proporcional
X_9	Sobrepeso 19-59 años	Directamente proporcional
X_{11}	Educación Nivel 2	Inversamente proporcional
X_{14}	Porcentaje pobreza NBI	Inversamente proporcional
X_{16}	Establecimientos médicos cada 10 mil hab.	Inversamente proporcional
X_{18}	Población	Directamente proporcional
X_{20}	Densidad poblacional	Directamente proporcional

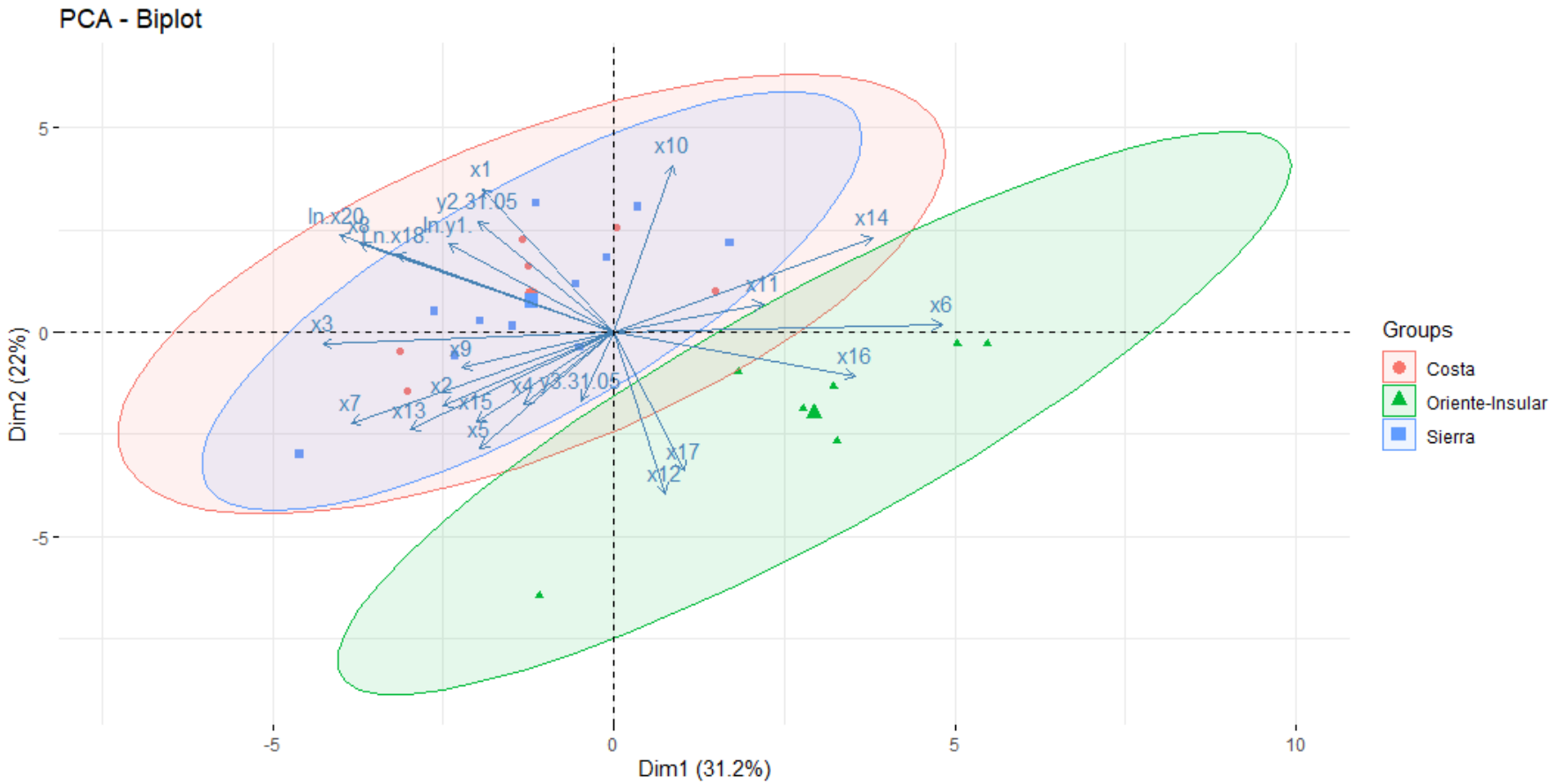


Figura 14. *Biplot agrupación Región*

En cuanto al gráfico **agrupado por X_{21}: *Región (Figura 14)***, las regiones costa y sierra no difieren mucho en su comportamiento. Ambas regiones están fuertemente caracterizadas por las variables Y_1 y Y_2, por lo que se puede decir que estas regiones han sido las más afectadas por la pandemia. En este sentido, X_1, X_8, $\ln X_{20}$ y $\ln X_{18}$ han influido para que las provincias de este grupo sean más afectadas, mientras X_6, X_{12}, X_{16} y X_{17} han contribuido para una menor afectación. A continuación, se presenta la relación de las variables con respecto a Y_1 y Y_2.

Tabla 13. *Relación entre variables predictoras y variables de respuesta Y_1 y Y_2, para agrupaciones Costa y Sierra*

Variable	Descripción	Tipo de relación con Y_1 y Y_2
X_1	Hogares características físicas inadecuadas	Directamente proporcional
X_2	Viviendas abastecimiento agua red pública	Directamente proporcional
X_3	Viviendas disponibilidad servicio eléctrico	Directamente proporcional
X_6	Porcentaje población 0-14 años	Inversamente proporcional
X_7	Porcentaje población 15-59 años	Directamente proporcional
X_8	Porcentaje población más de 60 años	Directamente proporcional
X_9	Sobrepeso 19-59 años	Directamente proporcional
X_{11}	Educación Nivel 2	Inversamente proporcional
X_{12}	Secundaria Terminada	Inversamente proporcional
X_{16}	Establecimientos médicos cada 10.000 hab	Inversamente proporcional
X_{17}	Médicos+enfermeras cada 10.000 hab	Inversamente proporcional
X_{18}	Población	Directamente proporcional
X_{20}	Densidad poblacional	Directamente proporcional
Y_1	Tasa mortalidad cada 10 mil habitantes	Directamente proporcional
Y_2	Tasa de letalidad	Directamente proporcional

Por otro lado, el área del grupo correspondiente a las regiones Oriente e Insular claramente tiende a la región inferior derecha (correlación positiva con el primer componente principal y negativa con el segundo). Se encuentra en dirección opuesta a la de las variables Y_1 y Y_2, por lo que este grupo no está caracterizado por ambas variables. Esto quiere decir que estas provincias han sido las menos afectadas por la pandemia, a la fecha de realización de este trabajo. Las variables predictoras que más representan a estas provincias son

X_6, X_{12}, X_{16} y X_{17}, mientras las que están menos representadas son X_1, X_3, X_8, $\ln X_{18}$ y $\ln X_{20}$.

Tabla 14. *Relación entre variables predictoras y variables de respuesta Y_1 y Y_2, para agrupación Oriente e Insular*

Variable	Descripción	Tipo de relación con Y_1 y Y_2
X_1	Hogares características físicas inadecuadas	Directamente proporcional
X_3	Viviendas disponibilidad servicio eléctrico	Directamente proporcional
X_6	Porcentaje población 0-14 años	Inversamente proporcional
X_8	Porcentaje población más de 60 años	Directamente proporcional
X_{10}	Educación Nivel 1	Directamente proporcional
X_{12}	Secundaria Terminada	Inversamente proporcional
X_{16}	Establecimientos médicos cada 10.000 hab	Inversamente proporcional
X_{17}	Médicos+enfermeras cada 10000 hab	Inversamente proporcional
X_{18}	Población	Directamente proporcional
X_{20}	Densidad poblacional	Directamente proporcional

Identificación de descriptores a través de Regresión Lineal Múltiple

Se utilizaron las metodologías Forward, Backward y Reemplazo Secuencial para identificar las variables predictoras que se relacionan de manera significativa con las variables de respuesta Y_1 y Y_3. No se considera relevante elaborar un modelo para Y_2, debido a su alta correlación con Y_1 (0,81, 1,8e^{-6} valor-p). Se definieron los mejores modelos para cada metodología utilizando los indicadores R^2 ajustado y CIB, y según esto se identificó las variables que más se repitieron, obteniendo los siguientes resultados:

Modelos con Y_1 como variable de respuesta:

Tabla 15. *Modelos con Y_1 como variable de respuesta*

	X_3	X_4	X_5	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{16}	X_{17}	$\ln(X_{18})$	$\ln(X_{20})$
Forward 4 variables		o	o							o				o
Forward 5 variables	o	o	o							o				o
Forward 6 variables	o	o	o		o					o				o
Backward 4 variables	o							o		o			o	

Backward 8 variables	o		o			o	o	o	o	o			o		
Backward 10 variables	o	o	o		o	o	o	o	o	o			o		
Backward 12 variables	o	o	o		o	o	o	o	o	o	o	o	o		
Sequential 4 variables	o							o		o					o
Sequential 5 variables	o	o	o							o					o
Sequential 6 variables		o	o	o	o					o					o
	8	**7**	**8**	**1**	**4**	**3**	**3**	**5**	**3**	**10**	**1**	**1**	**4**		**6**

Las variables que más se repiten entre todos los modelos son:

X_3 Viviendas disponibilidad servicio eléctrico

X_4 Viviendas disponibilidad alcantarillado

X_5 Viviendas disponen eliminación basura recolector

X_{14} Porcentaje pobreza NBI

$\ln(X_{20})$ Densidad poblacional

Estas variables corresponden a las que se obtuvieron con el forward con 5 descriptores:

Coeficientes	Estimado	Error Std.	Valor t	Valor p
Intercepto	-7,944	2,638	-3,012	0,007
X_3	0,037	0,029	1,293	0,212
X_4	-0,018	0,011	-1,671	0,112
X_5	0,027	0,009	2,858	0,010
X_{14}	5,524	1,932	2,859	0,010
$\ln(X_{20})$	0,493	0,130	3,784	0,001

Error estándar residual: 0,594 con 18 grados de libertad
R^2: 0,7139 R^2 ajustado: 0,634 Valor p: 0,0002

Si se eliminan del modelo las variables X_3 y X_4, ya que aparentemente no son significativas según su valor p, se obtiene el siguiente modelo:

Coeficientes	Estimado	Error Std.	Valor t	Valor p
Intercepto	-5,662	1,054	-5,373	2,94E-05

X_5	0,027	0,009	2,919	0,008
X_{14}	5,023	1,478	3,398	0,003
$\ln(X_{20})$	0,621	0,108	5,758	1,24E-05

Error estándar residual: 0,619 con 20 grados de libertad
R^2: 0,655 R^2 ajustado: 0,603 Valor p: 7,312E-05

Este modelo es más parsimonioso que el anterior, sin disminuir considerablemente el R^2 ajustado y mejorando el valor p del modelo.

Modelos con Y_3 como variable de respuesta:

Tabla 16. *Modelos con Y_3 como variable de respuesta*

	X_4	X_5	X_9	X_{12}	X_{16}	X_{17}	$\ln(X_{20})$
Forward 4 variables	o	o		o		o	
Forward 5 variables	o	o		o		o	o
Forward 6 variables	o	o		o	o	o	o
Forward 7 variables	o	o	o	o	o	o	o
Backward 3 variables	o	o			o		
Backward 4 variables	o	o			o		o
Backward 5 variables	o	o	o		o		o
Sequential 3 variables	o	o				o	
Sequential 4 variables	o	o				o	o
Sequential 5 variables	o	o			o	o	o
Sequential 6 variables	o	o	o		o	o	o
	11	**11**	**3**	**4**	**7**	**8**	**8**

Las variables que más se repiten entre todos los modelos son:

X_4 Viviendas disponibilidad alcantarillado

X_5 Viviendas disponen eliminación basura recolector

X_{17} Médicos + enfermeras cada 10.000 habitantes

$\ln(X_{20})$ Densidad poblacional

Las variables que más se repiten corresponden justamente al modelo de reemplazo secuencial con 4 variables. Los datos de este modelo son:

Coeficientes	Estimado	Error Std.	Valor t	Valor p
Intercepto	-5,04E-01	7,28E-01	-0,692	0,497
X_4	-3,43E-02	7,88E-03	-4,346	3,48E-04
X_5	2,91E-02	6,68E-03	4,357	3,40E-04
X_{17}	3,03E-02	1,34E-02	2,260	0,036
$\ln(X_{20})$	7,36E-02	8,24E-02	0,893	0,383

Error estándar residual: 4,58E-04 con 19 grados de libertad
R^2: 0,599 R^2 ajustado: 0,514 Valor p: 0,001

A pesar de que $\ln(X_{20})$ tiene un coeficiente elevado, se lo decide eliminar meramente por motivos estadísticos, ya que al hacerlo se mejoró el valor p del modelo, se lo hizo más parsimonioso al disminuirle una variable y prácticamente se mantuvo el valor de R^2 ajustado:

Coeficientes	Estimado	Error Std.	Valor t	Valor p
Intercepto	-7,00E-02	5,39E-01	-0,130	0,898
X_4	-3,33E-02	7,76E-03	-4,285	3,62E-04
X_5	2,92E-02	6,65E-03	4,390	2,83E-04
X_{17}	2,48E-02	1,19E-02	2,093	0,049

Error estándar residual: 4,55E-04 con 20 grados de libertad
R^2: 0,582 R^2 ajustado: 0,519 Valor p: 4,79E-04

Colinealidad

Se utilizaron los criterios indicados en la metodología para determinar la existencia de colinealidad:

- *Si el coeficiente de determinación R^2 es elevado, pero ninguno de los predictores resulta significativo, puede haber indicios de colinealidad.*

No aplica, ya que ambos modelos tienen R^2 elevados y sus predictores son significativos.

- *Estudiar la relación lineal entre cada par de predictores mediante el cálculo de una matriz de correlación. Tener en cuenta que puede existir multicolinealidad, aunque no se obtenga ningún coeficiente de correlación elevado. Puede darse el caso de que las correlaciones simples entre algunos pares de variables no sean mayores que 0,5, pero exista una relación lineal casi perfecta entre tres o más de estas variables.*

  Del análisis de correlación elaborado anteriormente se obtiene que, de las variables en cuestión, la única con una correlación elevada con respecto a Y_1 es $\ln(X_{20})$:

Variable	**Correlación (valor-p)**
$\ln(X_{20})$: Densidad poblacional	0,64 (0.001)

  Con respecto a Y_3, la variable con correlación más elevada de las que se presentaron en el modelo es X_5:

Variable	**Correlación (valor-p)**
X_5: Hogares características físicas inadecuadas	0,43 (0.038)

- *Elaborar un modelo de regresión lineal simple entre cada uno de los predictores. Si en alguno de los modelos el R^2 conocido también como coeficiente de determinación es elevado, podría estar señalando una posible colinealidad.*

- *Factor de Inflación de la Varianza (FIV):*
  - *FIV = 1: Ausencia total de colinealidad*

- *1 < FIV < 5: La regresión puede ser afectada por cierta colinealidad.*
- *5 < FIV < 10: Existe una alta probabilidad de que exista colinealidad entre las variables.*

Los coeficientes obtenidos fueron los siguientes:

Modelo para Y_1:

X_5: 1,229

X_4: 1,436

$\ln(X_{20})$: 1,192

Modelo para Y_3:

X_4: 1,306

X_5: 1,202

X_5: 1,096

Teniendo en cuenta todas estas evaluaciones, se puede concluir que no hay colinealidad en los modelos.

Interpretación de los modelos

Los modelos obtenidos finalmente serían los siguientes:

$$\boldsymbol{Y_1} = 5{,}662 + 0.027\,\boldsymbol{X_5} + 5.023\,\boldsymbol{X_{14}} + 0.621\ln(\boldsymbol{X_{20}})$$

Los coeficientes de este modelo nos indican lo siguiente:

- El hecho de que las viviendas dispongan de eliminación de basura mediante recolector (X_5) afecta directamente a la tasa de mortalidad (Y_1). Esto parecería no tener sentido, sin embargo, la explicación radicaría en que las provincias más grandes fueron las más afectadas al inicio de la pandemia y son justamente en las que más se cuenta con este servicio.

- Porcentaje pobreza por necesidades básicas insatisfechas (X_{14}) afecta directamente a la tasa de mortalidad, como era de esperarse, ya que estas personas deben salir diariamente a la calle a buscar sustento para llevar a sus hogares y no pueden resguardarse del virus.
- Por último, la densidad poblacional (X_{20}) también afecta directamente a la tasa de mortalidad. Esto tiene sentido, ya que el virus tiene mayor facilidad para transmitirse a mayor concentración de población.

$$\boldsymbol{Y_3} = -7 * 10^{-5} - 3 * 10^{-5}\boldsymbol{X_4} + 3 * 10^{-5}\,\boldsymbol{X_5} + 2 * 10^{-5}\,\boldsymbol{X_{17}}$$

Los coeficientes de este modelo nos indican lo siguiente:

- El hecho de que las viviendas tengan disponibilidad de alcantarillado (X_4) afecta inversamente a la proporción de incidencia (Y_3). La proporción de incidencia indica la relación entre los casos confirmados y la población.
- Igual que con el modelo anterior, el servicio de recolección de basura (X_5) parecería afectar directamente a la proporción de incidencia. Como se indicó anteriormente, la explicación radicaría en que las provincias más grandes fueron las más afectadas al inicio de la pandemia y son las que más cuentan con este servicio.
- La cantidad de médicos + enfermeras cada 10 mil habitantes (X_{17}) parecería afectar directamente a la proporción de incidencia. La explicación sería la misma que la indicada para el servicio de recolección de basura.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Base de datos depurada y completa

- Al finalizar el trabajo, se esperaba contar con una base de datos que contenga únicamente las variables que realmente tienen relación con las variables de respuesta.
- Se eliminaron las variables que no tienen esta relación y aquellas que podrían aparecer como duplicadas, debido a su alta relación con otras.
- Gracias a los resultados del análisis, se cumplió el objetivo de contar con una base de datos únicamente con las variables significativas.

Estudiar de manera univariante el comportamiento de cada variable

- Las provincias más afectadas por la pandemia son: Chimborazo, El Oro, Galápagos, Guayas, Los Ríos, Manabí, Pastaza, Santa Elena, Santo Domingo.
- Entre el 30% y 40% de estas provincias se encuentran sobre el 3er cuartil de las variables: hogares características físicas inadecuadas (X_1), sobrepeso 19-59 años (X_9), población (X_{18}) y densidad poblacional (X_{20}).
- Entre el 30% y 40% de estas provincias se encuentra debajo del 1er cuartil de las variables: abastecimiento agua red pública (X_2), disponibilidad de alcantarillado (X_4) y establecimientos médicos cada 10 mil habitantes (X_{16}).

Analizar las relaciones existentes entre las variables objeto de estudio

- Al analizar la correlación entre variables, se detectó que muchas predictoras tienen correlaciones fuertes, por lo que muchas de las variables resultan redundantes. En algunos casos esta correlación es lógica, ya que se dio entre variables que forman parte de una misma agrupación. Por ejemplo, *Viviendas con abastecimiento agua red pública* (X_2) presenta alta correlación con *Viviendas con disponibilidad servicio eléctrico* (X_3) y *Viviendas con disponibilidad alcantarillado* (X_4). Esto tiene sentido, ya que una vivienda sin abastecimiento de agua es muy probable que no cuente con los otros dos servicios.

  Debido a esta alta correlación entre algunas predictoras, se vio la necesidad de utilizar métodos de selección de variables. Así mismo, se encontró una elevada correlación entre la Tasa de mortalidad (Y_1) y la Tasa de letalidad (Y_2), por lo que en análisis posteriores se utilizó solo una de estas variables.

Identificar posibles factores de riesgo

- Luego de utilizar las metodologías Forward, Backward y Reemplazo Secuencial, para identificar los factores que se relacionan de manera significativa con las variables de respuesta Tasa de mortalidad (Y_1) y Proporción incidencia (Y_3), se determinó que:
  - El hecho de que las viviendas dispongan de eliminación de basura mediante recolector (X_5) afecta directamente a la tasa de mortalidad (Y_1). Esto parecería no tener sentido, sin embargo, la explicación radicaría en que las provincias más grandes fueron las más afectadas al inicio de la pandemia y son justamente en las que más se cuenta con este servicio.
  - Porcentaje pobreza por necesidades básicas insatisfechas (X_{14}) afecta directamente a la tasa de mortalidad, como era de esperarse, ya que estas personas deben salir diariamente a la calle a buscar sustento para llevar a sus hogares y no pueden resguardarse del virus.

- Por último, la densidad poblacional (X_{20}) también afecta directamente a la tasa de mortalidad. Esto tiene sentido, ya que el virus tiene mayor facilidad para transmitirse a mayor concentración de población.
- El hecho de que las viviendas tengan disponibilidad de alcantarillado (X_4) afecta inversamente a la proporción de incidencia (Y_3). La proporción de incidencia indica la relación entre los casos confirmados y la población.
- Igual que con el modelo anterior, el servicio de recolección de basura (X_5) parecería afectar directamente a la proporción de incidencia. Como se indicó anteriormente, la explicación radicaría en que las provincias más grandes fueron las más afectadas al inicio de la pandemia y son las que más cuentan con este servicio.
- La cantidad de médicos + enfermeras cada 10 mil habitantes (X_{17}) parecería afectar directamente a la proporción de incidencia. La explicación sería la misma que la indicada para el servicio de recolección de basura.

Otras conclusiones

- Al analizar los datos agrupados por nivel de población en mediante el biplot (**Figura 13**), se determinó que las variables con menor cantidad de habitantes (menor que 250 mil) han sido las menos afectadas por la pandemia, hasta la fecha del estudio, ya que la tendencia de su elipsoide es opuesta a la dirección de las variables Y_1 y Y_2. Esto tiene lógica, ya que la pandemia empezó y afectó más fuertemente a las provincias con mayor población, como son Guayas, Manabí y Pichincha.
- En el mismo biplot (**Figura 13**), se encontró que las provincias con población entre 250 mil y 750 mil, y mayor a 750 han sido las más afectadas por la pandemia. Además, se evidenció que las variables correspondientes a Hogares características físicas inadecuadas (X_1),

Porcentaje población más de 60 años (X_8), Población ($\ln X_{18}$) y Densidad poblacional ($\ln X_{20}$) influyen positivamente sobre la Tasa de mortalidad (Y_1) y la Tasa de letalidad (Y_2). Esto es lógico, ya que las personas mayores son las más afectadas por la enfermedad y la cantidad y aglomeración de personas favorecen la multiplicación de ésta.

Las variables que desfavorecieron el impacto de la pandemia fueron Secundaria Terminada (X_{12}), Médicos y enfermeras cada 10 mil habitantes (X_{17}) y Establecimientos médicos cada 10 mil habitantes (X_{16}) en las provincias con población entre 250 mil y 750 mil habitantes, y Porcentaje población 0-14 años (X_6), Porcentaje pobreza NBI (X_{14}) y Establecimientos médicos cada 10 mil habitantes (X_{16}) en aquellas con más de 750 mil habitantes.

- Al analizar los datos agrupados por región (**Figura 14**), se detectó que las regiones Costa y Sierra tienen comportamientos similares y que, al igual que las provincias con población mayor a 250 mil personas, están fuertemente caracterizadas por la Tasa de mortalidad (Y_1) y la Tasa de letalidad (Y_2), es decir, han sido las más afectadas por la pandemia. Las variables que favorecieron la afectación de la pandemia son: Hogares con características físicas inadecuadas (X_1), Porcentaje de población con más de 60 años (X_8), Población ($\ln X_{18}$) y Densidad poblacional ($\ln X_{20}$).
- Como se evidencia en la Figura 14, a la fecha de corte de los datos, las provincias de las regiones Oriente e Insular han sido las menos afectadas por la pandemia. Las variables que ayudaron a disminuir el impacto de la pandemia fueron: Porcentaje población 0-14 años (X_6), Secundaria Terminada (X_{12}), Establecimientos médicos cada 10 mil habitantes (X_{16}) y Médicos y enfermeras cada 10 mil habitantes (X_{17}).

6.2. Recomendaciones

- Se recomienda realizar un nuevo análisis dentro de unos meses con datos actualizados, para determinar si han cambiado los factores de riesgo o los comportamientos de las agrupaciones. Por ejemplo, al haber avanzado el contagio por COVID-19 en todas las provincias, las variables disponibilidad de alcantarillado (X_4) y eliminación de basura mediante recolector (X_5), relacionadas con las condiciones de vivienda, podrían cambiar su influencia, la cual en este momento es directa.

Bibliografía

Arbós, D., & Belles, M. (2020). *14 maneras de destruir a la humanidad.* Angle Editorial.

BBC News. (11 de febrero de 2020). *BBC News*. Obtenido de Coronavirus disease named Covid-19: https://www.bbc.com/news/world-asia-china-51466362

Chen, F.-M., Feng, M.-C., Chen, T.-C., Hsieh, M.-H., Kuo, S.-H., Chang, H.-L., . . . Chen, Y.-H. (2020). Big data integration and analytics to prevent a potential hospital outbreak of COVID-19 in Taiwan. *Journal of Microbiology, Immunology and Infection*.

Cuadras, C. M. (2004). *Análisis Multivariante.*

Departamento de Seguridad Nacional de España. (11 de Marzo de 2020). *Departamento de Seguridad Nacional de España*. Obtenido de Coronavirus (COVID-19) - 11 de marzo 2020: https://www.dsn.gob.es/es/actualidad/sala-prensa/coronavirus-covid-19-11-marzo-2020

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis.*

Firth, S. (5 de marzo de 2020). *MedPage Today*. Obtenido de Singapore: The Model for COVID-19 Response?: https://www.medpagetoday.com/infectiousdisease/covid19/85254

IBM Knowledge Center. (s.f.). *IBM Knowledge Center*. Obtenido de IBM: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/chaidnode_general.html

Infosalus. (22 de marzo de 2020). *Infosalus*. Obtenido de Breve historia de las pandemias globales: cómo hemos luchado contra los mayores asesinos: https://www.infosalus.com/salud-investigacion/noticia-breve-historia-pandemias-globales-hemos-luchado-contra-mayores-asesinos-20200322075937.html

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 119-127.

Kassambara, A. (11 de 3 de 2018). *Statistical tools for high-throughput data analysis*. Obtenido de Stepwise Regression Essentials in R:

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied Linear Regression Models.* McGraw-Hill Education.

Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., . . . Shao, Y. (2020a). Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling*, 282-292.

Li, M., Dong, Y., Wang, H., Guo, W., Zhou, H., Zhang, Z., . . . Hu, D. (2020c). Cardiovascular disease potentially contributes to the progression and poor prognosis of COVID-19. *Nutrition, Metabolism and Cardiovascular Diseases*.

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., . . . Zhao, J. (2020b). Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *Journal of Allergy and Clinical Immunology*.

Marco, F. (s.f.). *Economipedia*. Obtenido de R cuadrado ajustado (Coeficiente de determinación ajustado): https://economipedia.com/definiciones/r-cuadrado-ajustado-coeficiente-de-determinacion-ajustado.html

Ministerio de Salud Pública. (6 de 2020). *Gestión de Riesgos EC*. Obtenido de Infografía Nacional Covid-19: https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/06/INFOGRAFIA-NACIONALCOVI-19-COE-NACIONAL-30062020-08h00.pdf

Moreno-Altamirano, A., López-Moreno, S., & Corcho-Berdugo, A. (2000). Principales medidas en epidemiología. *Salud Pública de México*, 337-348.

OMS. (2020a). *OMS*. Obtenido de Coronavirus: https://www.who.int/es/health-topics/coronavirus

OMS. (2020b). *OMS*. Obtenido de Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19): https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses

OMS. (11 de febrero de 2020c). *OMS*. Obtenido de Intervención del Director General de la OMS en la conferencia de prensa sobre el 2019-nCoV del 11 de febrero de 2020: https://www.who.int/es/dg/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020

OMS. (11 de marzo de 2020d). *OMS*. Obtenido de Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19 celebrada el 11 de marzo de 2020: https://www.who.int/es/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Pal, R., & Bhadada, S. K. (2020). COVID-19 and diabetes mellitus: An unholy interaction of two pandemics. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 513-517.

Palacios-Cruz, L., Pérez, M., Rivas-Ruiz, R., & Talavera, J. O. (213). Investigación clínica XVIII. *Revista Médica del Instituto Mexicano del Seguro Social*, 656-661.

Peña, D. (2002). *Análisis de datos multivariantes.*

Reuters. (5 de febrero de 2020). *Reuters*. Obtenido de WHO: 'no known effective' treatments for new coronavirus: https://www.reuters.com/article/us-china-health-treatments-who-idUSKBN1ZZ1M6

Santana, E. (19 de Abril de 2015). *Apuntes R*. Obtenido de Machine Learning con R: http://apuntes-r.blogspot.com/2015/04/supuestos-en-regresion-lineal.html

Secretaría Nacional de Gestión de Riesgos y Emergencias. (2020). *Servicio Nacional de Gestión de Riesgos y Emergencias*. Obtenido de Informes de Situación e Infografías – COVID 19: https://www.gestionderiesgos.gob.ec/informes-de-situacion-covid-19-desde-el-13-de-marzo-del-2020/

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias.*

Wikipedia. (2020a). *Wikipedia*. Obtenido de Análisis de componentes principales: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_componentes_principales

Wikipedia. (14 de Diciembre de 2020b). *Wikipedia*. Obtenido de Bayesian information criterion: https://en.wikipedia.org/wiki/Bayesian_information_criterion

Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART.

Wooldridge, J. (2010). *Introducción a la econometría.* Cengage Learning Editores, S.A. de C.V.

Yang, C.-J., Chen, T.-C., & Chen, Y.-H. (2020). The preventive strategies of community hospital in the battle of fighting pandemic COVID-19 in Taiwan. *Journal of Microbiology, Immunology and Infection*.

Yen, M.-Y., Schwartz, J., Chen, S.-Y., & King, C.-C. (2020). Interrupting COVID-19 transmission by implementing enhanced traffic control bundling: Implications for global prevention and control efforts. *Journal of Microbiology, Immunology and Infection*.

Zhang, J., Wang, X., Jia, X., Li, J., Hu, K., Chen, G., . . . Dong, W. (2020b). Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China. *Clinical Microbiology and Infection*.

Zhang, X., Ma, R., & Wang, L. (2020a). Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. *Chaos, Solitons & Fractals*.