Análisis multivariante empleando correspondencias canónicas - COVID19 - Ecuador

by William F. Tandazo-Vargas

Abstract An abstract of less than 150 words.

Introducción

El 11 de marzo del 2020, La organización mundial de la salud (OMS) declara pandemia el brote del virus SARS-CoV-2, causante del COVID-19. Esta pandemia desato una serie de crisis dentro de diferentes aspectos sociales, económicos y políticos de las naciones del mundo. Los sistemas sanitarios tanto de potencias como de países con economías emergentes se mostraron frágiles e incapaces de satisfacer la demanda de pacientes infectados con esta enfermedad. Se ha mostrado que para el caso de algunos países que conforman el bloque G7 las acciones tomadas durante el brote causaron un impacto significativo en el desarrollo del virus SARS-CoV-2 (Zhang et al., 2020b).

En Ecuador, el caso cero fue reportado el 29 de febrero del 2020. Se determino que la paciente posiblemente lo contrajo desde España, país el cual la paciente llego dos semanas antes de confirmar su estado. Es desde este punto que se tiene constancia de la evolución del virus en Ecuador, situación que se complico a las pocas semanas después, lo que provocó el confinamiento total del país un 17 de marzo del 2020 (Santillan Haro, 2020). Las cifras de fallecidos después de esos días se dispararon llegando a sus picos más grandes los días del 22 de marzo al 9 de junio del 2020. Se pudo demostrar luego que los reportes generales descritos por el ministerio de salud pública del Ecuador (MSP) no contemplaban o ajustaban al verdadero impacto que genero esta pandemia desde los meses de febrero a octubre (Cevallos-Valdiviezo et al., 2021).

El análisis de datos se volvió fundamental en esta pandemia, se convirtió en un arma más para combatir el virus. Casos como el de Taiwán ponen en evidencia que el manejo e integración de los metadatos y su análisis previnieron la saturación de los hospitales, ofreciendo un servicio y manejo apropiado de los ciudadanos durante los días más fuertes del brote (Chen et al., 2020) . Con el análisis de los datos se logró determinar el impacto de las comorbilidades en pacientes que tienen el virus, lo que permitió darles una prioridad respecto a los demás que no poseen enfermedades como Diabetes o problemas cardiovasculares (Li et al., 2020) (Pal and Bhadada, 2020) (Zhang et al., 2020a), al igual demostrar que este virus afecta indiferentemente del sexo, pero se vuelve más crítico en pacientes de grupos etarios mayores a 65 años.
Para el caso de Ecuador, el subregistro de los factores o variables que tuvieron relación con el impacto de la pandemia no han permitido aclarar el desarrollo de esta. El estudio multivariante de las distintas variables de salud como camas disponibles, ventilación, admisiones dentro de los hospitales y las unidades de cuidado intensivo (UCI) tiene un poder predictivo importante al momento de medir la mortalidad del virus (de Fatima et al., 2020). Al igual que en materia económica, esto sirvió para evidenciar el "hoyo" que genero la pandemia en la economía global de los mercados (Sharma et al., 2020). Pese a que los datos de las distintas variables inherentes asociadas al COVID-19 para Ecuador no han sido debidamente liberadas es posible realizar un estudio multivariante con las existentes, como en el caso de Nepal (Devkota, 2021).

Gracias a las medidas de restricción de movilización humana dentro de Ecuador se ha mostrado que ciertos agentes contaminantes como el NO2 disminuyeron hasta en 5.6 veces menos con respecto a la media comparando el año de la pandemia con el 2018 y el 2019 (Zambrano-Monserrate and Ruano, 2020). El estudio de estos factores es importante puesto que estos agentes en altas concentraciones tienen una incidencia directa en la salud respiratoria de los ciudadanos. También, estudios realizados en Singapur han demostrado que este tipo de variables meteorológicas como la calidad del aire, temperatura y nivel de precipitación diaria fueron significativos en días donde hubieron varios casos confirmados de COVID-19 (Lorenzo et al., 2021).

Este estudio tiene como objetivo realizar un estudio multivariante tomando variables de registros publicos y de subregistros; variables como de mortalidad, condiciones de vida, alertas ciudadanas, casos confirmados de COVID-19 y variables de carácter meteorológicas dentro de las fechas del [por confirmar]. Para esto se usará un modelo de análisis de correspondencia canónica para encontrar la naturaleza de las relaciones existentes entre estas variables obtenidas en los comienzos de la pandemia de COVID-19.

Métodos y Materiales.

Redución de dimensiones.

Con el inicio de la pandemia surgen distintas formas de analizar el impacto en el que esta afecta a un determinado sector. Estas variables son numerosas para distintos ámbitos en los que se quiera enfocar, es decir, si nuestro estudio posee p-variables explicativas entonces estas serán de dimensión p. Reducir las dimensiones será fundamental para nuestro análisis, llevar a una expresión condensada de estas que explique su interrelación nos ayudará a concluir los objetivos de este estudio. Tomar que variables son las que más aportan dentro de un estudio ha sido tema de investigación durante mucho tiempo dentro del campo estadístico. Existen dos distintos enfoques para esto, uno que queda en manos del investigador el cual de todas las covariables disponibles toma un grupo de estas según su criterio, estas siendo las que el cree son las menos redundantes y las más significativas al momento de explicar el fenómeno que estudia. El segundo enfoque, el cual es el que desarrollaremos más de fondo en este estudio, será el de reducción de dimensiones suficiente. Este enfoque producirá una combinación lineal del grupo de covariables tratadas. De lo anterior, se pueden producir varias combinaciones lineares de un grupo de covariables, la meta es decir que combinación o combinaciones son las que mejor explican el efecto que se estudia. Para el análisis de correspondencias, este paso es crucial y permite hacerse con un costo de perdida de información mínima, nuestro objetivo será el restringir esta mínima perdida de información para que este explique el máximo de información.

Ordenamiento en espacio reducido.

Al reducir la dimensión de los datos es necesario que estos estén ordenados alrededor de una serie de ejes ortogonales reducida, estos de manera descendiente siendo el primero el que más información provee de esta estructura. Esto se hace por medio de la extracción de valores propios asociado a la matriz de datos de nuestras variables. Si esta matriz de datos es de nxp, n observaciones en p variables estudiadas, estas n observaciones pueden ser representadas por grupos en un espacio p dimensional. Este grupo de datos usualmente se acumula más en ciertas direcciones y es plano en otras dentro de nuestro espacio p dimensional y no necesariamente este tomara el curso de alguna de las variables. Donde más esta alargado esta nube de puntos es donde reside la mayor varianza de este grupo, será el mayor gradiente que representan nuestros datos. Este será nuestro primer eje que se extraerá y es el que más información dota nuestro análisis, el siguiente eje será el segundo más impórtate solo si este es ortogonal al primero. El numeró de ejes que interpretan la estructura de los datos se determina por el método a tomar que en nuestro caso ser el de análisis de correspondencia.

Análisis de Correspondencia.

El análisis de correspondencia (AC) es una técnica estadística la cual es útil cuando en la matriz de datos se tienen filas y columnas de naturaleza categórica. En particular, se usa cuando se tiene tablas de contingencia la cual contiene frecuencias numéricas de los perfiles, produciendo un análisis gráfico más simple y formal, haciendo que su interpretación sea eficiente. Esta técnica buscara la mejor representación de las variables de la matriz en un plano de baja dimensión, los cuales llamaremos ejes que de manera ordinal se organizaran siendo el primero el que mejor explique la asociación entre los perfiles filas y columnas. Los otros factores trataran explicar la mayor parte del residuo que no explico su inmediato anterior, organizándose de manera descendiente. Este método buscara encontrar estas dimensiones tal que la variabilidad geométrica de la nube de puntos (también llamada inercia) sea máxima.
Esencialmente, el análisis de correspondencia tiene 3 fases: La transformación inicial de la matriz de datos, la descomposición en valores singulares de la transformación y el cambio de escala de los vectores proprios resultantes. Del primer paso, la transformación de la matriz se lleva acabo de la siguiente forma:

$$H = S^{-\frac{1}{2}} X C^{-\frac{1}{2}} \qquad (1)$$

La matriz H tiene los valores transformados de la matriz X. Las matrices $S^{-\frac{1}{2}}$ y $C^{-\frac{1}{2}}$ son matrices diagonales que contienen la raíz cuadrada de las marginales totales reciprocas tanto de las dimensiones de la fila y columna. Esta transformación remueve los efectos de las magnitudes que se formas de las diferencias entre marginales totales. En el caso de tablas de contingencia lo anterior seria equivalente a remover los valores esperados del estadístico Chi-cuadrado del en el modelo de independencias.

La segunda fase constaría en obtener los vectores característicos de nuestra matriz, mediante descomposición de valores singulares.

$$X_{n,m} = U_{n,m} d_{m,m} V^T_{n,m} \quad (2)$$

Donde:

$$U_{n,m} : \textit{ es la matriz de la dimensin } fila.$$

$$V^T_{n,m} : \textit{ es la matriz de la dimensin columna.}$$

$$d_{m,m} : \textit{ es la matriz diagonal de valores propios.}$$

Las matrices U, V y d se pueden encontrar directamente con una descomposición de valores singulares en la matriz de datos X o indirectamente mediante la extracción de los valores propios de las matrices de productos cruzados de X. Como última fase, las columnas tanto de las matrices U y V deben ser reescaladas para obtener los valores óptimos, también llamados valores canónicos. El resultado de esto da un ordenamiento donde se preservará la distancia chi-cuadrado (χ^2) entre los perfiles. Gráficamente el resultado de un análisis de correspondencias puede ser descrito por medio de un "Biplot" el cual grafica puntos del perfil fila x_i y puntos del perfil columna y_j tal que los productos escalares entre los vectores de fila y columna se aproximan a los elementos correspondientes de los datos de la matriz lo más cerca posible en un espacio reducido.

Análisis de Correspondencia Múltiple.

El análisis de Correspondencia (ACM) visto en el punto anterior es el caso más sencillo cuando se tiene dos variables categóricas. El análisis de correspondencia múltiple involucra 3 o más variables dentro del conjunto de variables a estudiar. El análisis de correspondencia múltiple equivale a realizar un análisis de correspondencia simple, pero a la matriz indicadora de nuestros datos. En estas matrices indicadoras cada fila representa un caso y las columnas representan todas las categorías de las variables. Para el caso de variables de naturaleza continua estas pueden ser redefinidas por categorías, con una variable por categoría. Empleando un análisis por medio del producto cruz o matriz de Burt de las variables indicadoras se puede llegar a la solución óptima.

Análisis de Correspondencia Canónica.

Como hemos visto, el análisis de correspondencia nos ayuda a visualizar los datos de una matriz dentro de un espacio de dimensión reducida, donde la inercia de esta nube de puntos será la más optima. Un análisis de correspondencia canónico (ACC) se puede emplear cuando encontramos estas dimensiones a partir del AC, pero con la condición de que estas son una combinación lineal de variables explicativas adicionales. Podemos ver esto como un problema de regresión lineal múltiple, solo que en vez de hacer regresión sobre las variables explicativas lo haremos a las dimensiones de estas variables explicativas. El ACC restringe el espacio donde este buscara los ejes principales óptimos de un espacio total, el complemento de este será un espacio no restringido el cual podría también tener la solución óptima. El porcentaje total de inercia estará repartido entre estos dos espacios, siendo el restringido el lugar donde se encontraría la inercia máxima. En consecuencia, a lo anterior el ACC explicara un poco menos del total de inercia existente. Para interpretar el modelo de manera gráfica se sigue la de un AC clásico y se sigue la misma explicación que la del "Biplot".

Datos

Bibliography

H. Cevallos-Valdiviezo, A. Vergara-Montesdeoca, and G. Zambrano-Zambrano. Measuring the impact of the covid-19 outbreak in ecuador using preliminary estimates of excess mortality, march 17october 22, 2020. *International Journal of Infectious Diseases*, 104:297–299, 2021. ISSN 1201-9712. doi: https://doi.org/10.1016/j.ijid.2020.12.045. URL https://www.sciencedirect.com/science/article/pii/S1201971220325674. [p1]

F.-M. Chen, M.-C. Feng, T.-C. Chen, M.-H. Hsieh, S. Kuo, H.-L. Chang, C.-J. Yang, and Y.-H. Chen. Big data integration and analytics to prevent a potential hospital outbreak of covid-19 in taiwan. *Journal of Microbiology, Immunology and Infection*, 54, 04 2020. doi: 10.1016/j.jmii.2020.04.010. [p1]

A. de Fatima, B. Boger, R. Vilhena, M. Fachi, J. Marlei, M. Fernandes, and Santos. A multivariate analysis of risk factors associated with death by covid-19 in the usa, italy, spain, and germany (2). *Journal of public health*, 1, 10 2020. doi: 10.1007/s10389-020-01397-7. [p1]

J. Devkota. Multivariate analysis of covid-19 for countries with limited and scarce data: Examples from nepal. *Journal of Environmental and Public Health*, 2021, 01 2021. doi: 10.1155/2021/8813505. [p1]

M. Li, Y. Dong, H. Wang, W. Guo, H. Zhou, Z. Zhang, C. Tian, K. Du, R. Zhu, L. Wang, L. Zhao, H. Fan, S. Luo, and D. Hu. Cardiovascular disease potentially contributes to the progression and poor prognosis of covid-19. *Nutrition, Metabolism and Cardiovascular Diseases*, 30, 04 2020. doi: 10.1016/j.numecd.2020.04.013. [p1]

J. S. L. Lorenzo, W. W. S. Tam, and W. J. Seow. Association between air quality, meteorological factors and covid-19 infection case numbers. *Environmental Research*, 197:111024, 2021. ISSN 0013-9351. doi: https://doi.org/10.1016/j.envres.2021.111024. URL https://www.sciencedirect.com/science/article/pii/S0013935121003182. [p1]

R. Pal and S. Bhadada. Covid-19 and diabetes mellitus: An unholy interaction of two pandemics. *Diabetes and Metabolic Syndrome Clinical Research and Reviews*, 05 2020. doi: 10.1016/j.dsx.2020.04.049. [p1]

A. Santillan Haro. Caracterización epidemiológica de covid-19 en ecuador. *InterAmerican Journal of Medicine and Health*, 3:1 – 7, Apr. 2020. doi: 10.31005/iajmh.v3i0.99. URL https://www.iajmh.com/iajmh/article/view/99. [p1]

N. Sharma, S. Yadav, M. Mangla, A. Mohanty, and S. Mohanty. Multivariate analysis of covid-19 on stock, commodity and purchase manager indices: A global perspective. 09 2020. doi: 10.21203/rs.3.rs-68388/v1. [p1]

M. Zambrano-Monserrate and M. Ruano. Has air quality improved in ecuador during the covid-19 pandemic? a parametric analysis. *Air Quality, Atmosphere and Health*, 13, 08 2020. doi: 10.1007/s11869-020-00866-y. [p1]

J. Zhang, X. Wang, X. Jia, J. Li, K. Hu, G. Chen, J. Wei, Z. Gong, C. Zhou, H. Yu, M. Yu, H. Lei, F. Cheng, B. Zhang, Y. Xu, G. Wang, and W. Dong. Risk factors for disease severity, unimprovement, and mortality of covid-19 patients in wuhan, china. *Clinical Microbiology and Infection*, 26, 04 2020a. doi: 10.1016/j.cmi.2020.04.012. [p1]

X. Zhang, R. Ma, and L. Wang. Predicting turning point, duration and attack rate of covid-19 outbreaks in major western countries. *Chaos, Solitons Fractals*, 135:109829, 2020b. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2020.109829. URL https://www.sciencedirect.com/science/article/pii/S0960077920302290. [p1]

William F. Tandazo-Vargas
ESPOL
line 1
line 2

wtandazo@espol.edu.ec