Instituto Tecnológico Autónomo de México

Estadística No Paramétrica y Estadística Aplicada 3

Proyecto Final

# Índice de vulnerabilidad para la CDMX

*Elaborado por:*
Pablo Barranco Soto-151528
Sofía De la Mora Tostado- 160062
Román Alberto Velez Jiménez - 165462
Santiago Muriel Vizcaino - 163195
Mariana Rivera Fares - 157957

*Profesor:*
Jorge Francisco De la Vega

28 de mayo del 2020

# Índice

# 1. Introducción

La crisis de salud que está viviendo la humanidad provocada por el Covid-19 llama a la implementación de una serie de políticas públicas por parte de las diferentes dependencias nacionales e internacionales. Sin duda, gracias a la masificación de datos y acceso a ellos, estas políticas ahora podrán estar basadas en datos reales sin aumentar el costo para obtener más y mejores resultados. Se reconoce que el problema que el mismo acceso masivo a la información trae consigo es la calidad de los datos, exigiendo una mejor calidad de su análisis.

Uno de los retos más grandes que presenta la pandemia es determinar en dónde y cómo se enfocarán los recursos, ya sea en el sector de salud, el sector de alimentación, el sector de economía u otros. Actualmente, contamos con mediciones como la tasa de infección, la tasa de contagio, la tasa de default, ente otras que miden una reyerta específica, empero creemos que es necesario instrumentar una medida holística que integre factores demográficos, socioeconómicos y de infraestructura del sector salud para auxiliar a la toma de decisiones aún más informada para enfrentar los retos que acompañan a la pandemia.

Se construyó un índice de vulnerabilidad multidimensional con ayuda de datos de la Ciudad de México para contribuir a que las políticas públicas y los esfuerzos comunales combatan deliberadamente las consecuencias de la pandemia y tengan el mayor impacto posible en la población. Este índice estará enfocado en cinco dimensiones que consideramos representativas de la vulnerabilidad desde distintas vertientes. Siendo la mayor preocupación la identificación de la población que tiene necesidades más inmediatas que, de no cubrirse, podrían a las personas en una situación paupérrima.

Principalmente, se encontraron a las delegaciones más indefensas ante una crisis sanitaria en términos económicos, médicos y laborales. Para eso se utilizaron una serie de variables de bases de datos públicas de la Ciudad de México para realizar un análisis de componentes principales y encontrar un índice que midiera, homogéneamente, la vulnerabilidad. Se encontró que, al menos en la Ciudad de México, se tienen diferentes necesidades según la región y será primordial el apoyo gubernamental para superar los estragos que sigue dejando la pandemia de Covid-19. Reconocemos que el índice no es solamente aplicable a la crisis del Covid-19, sino a cualquier crisis que se pueda presentar en el futuro próximo. Dado que en este momento se desea detener las consecuencias que puede dejar esta pandemia, nos referiremos a esta crisis en particular a lo largo del trabajo.

## 2. Datos

Se seleccionó a la Ciudad de México para realizar el análisis por ser la región del país con más datos disponibles, con mayor población y con gran diversidad social. Se reconoce que una cuestión que puede afectar la investigación al elegir a las delegaciones de la CDMX como unidades observacionales es que poco más de ocho millones habitan la ciudad, pero la población que ingresa a la ciudad diariamente para estudiar o trabajar provenientes del Estado de México, del Estado de Hidalgo y del Estado de Morelos asciende a un millón 720 mil 145 personas, de las cuales entre 39 % y 47 % tardan de una a dos horas en su traslado. Esto ocasiona que la población sea variable, incluso con la imposición de cuarentena.

A continuación se presentan las variables para realizar el análisis pertinente:

| **Nombre** | **Significado** | **Unidad** | **Fuente** |
|---|---|---|---|
| Alc | Delegación de la Ciudad de México | alfa | Datos CDMX [1] |
| Clave | Clave de la delegación | Delegación | Datos CDMX [1] |
| Pob | Población de la delegación | Personas | Censo, 2015 [3] |
| km2 | Tamaño de la delegación | $km^2$ | Datos CDMX [1] |
| In_inf | Población con ingreso inferior a la línea de bienestar | Personas | CONEVAL [2] |
| C_alim | Población con carencia por acceso a la alimentación | Personas | CONEVAL [2] |
| C_sal | Población con carencia por acceso a los servicios de salud | Personas | CONEVAL [2] |
| C_ser | Población con carencia por acceso a los servicios básicos | Personas | CONEVAL [2] |
| Sol_des | Solictudes seguro de desempleo (marzo-abril 2020) | Solicitudes | Datos CDMX [1] |
| Desem | Población total desempleada | Personas | CONEVAL [2] |
| P_medico | Número de personal trabajadoras de la salud | Personas | Salud CDMX [4] |
| C_Conf | Casos totales confirmados hasta el 18 de mayo de 2020 | Personas | Datos CDMX [1] |
| Remuneracion | Retribución economica de los hogares | Miles de pesos | Anuario estadístico [5] |
| P_af_ser | Población afiliada a servicios de salud | Porcentaje | Anuario estadístico [5] |
| esp_vida | Esperanza de vida | Años | Anuario estadístico [5] |
| Prom_escolaridad | Promedio de años de escolaridad | Años | Anuario estadístico [5] |

Cuadro 1: Datos para el cálculo del índice

Donde la delegación, la clave, la población y los kilómetros cuadrados se emplearon como variables de respuesta o auxiliares para estandarizar el resto de las variables que se tomaron como explicativas. Sobre esto se profundizará más adelante.

Inicialmente, estas fueron todas las variables que se utilizaron para comenzar con el análisis exploratorio de datos. Tras realizar un estudio eliminando las variables que causaban colinealidad, se encontró que no se cumplía el propósito del trabajo en la manera que se quería, pues al construir en índice se perdía variabilidad. Se procedió eliminando las variables cuya correlación con el resto era sumamente pequeña y se logró un mejor índice con dicha reducción.

Las variables explicativas que se emplearon para la construcción del índice en cuestión fueron las siguientes: Población con ingreso inferior a la línea de bienestar, Población con carencia por acceso a la alimentación,

Población con carencia por acceso a los servicios de salud, Número de personal trabajador de la salud, Remuneración total y Promedio de años de escolaridad. Para homogeneizar las medidas de cada variable se dividió entre la población de cada delegación. Así, cada valor está en términos de proporción con respecto al número de habitantes de cada delegación.

Otra observación importante es que la variable de *Remuneración* estaba medida en millones de pesos, pero se buscó transformarla a miles de pesos para que tuviera mayor impacto en el índice obtenido, dado que su relación con la vulnerabilidad de acceso a servicios de salud, y capacidad en general de enfrentar una crisis sanitaria, se ven altamente impactados por el ingreso familiar.

# 3. Metodología

## 3.1. Análisis Exploratorio de Datos

Para familiarizarse con los datos, se realizó un *Análisis exploratorio de datos (EDA*[1]*)* cuyo método carece de formalidad, sin embargo es de gran utilidad para responder a preguntas sencillas antes de manipular la información. Se seleccionaron las gráficas que, de manera visual, permitieran responder a preguntas sobre el comportamiento, relación y robustez de las variables. Para el análisis de datos es indispensable realizar este procedimiento, además, es apropiado aplicarlo tanto a datos cuantitativos cómo cualitativos.

El EDA supone que entre más se conozca sobre los datos con los que se realizará un estudio, mejores serán los resultados para probar una teoría o encontrar la respuesta a algunas preguntas. En este caso buscamos la vulnerabilidad de las personas habitantes de cada delegación al Covid-19, así que se analizó el comportamiento de las variables para cada una de las regiones. De esta manera, se reexpresaron los datos con otras unidades para obtener un índice que se entendiera de manera intuitiva.

## 3.2. Análisis de Componentes Principales

El *Análisis de Componentes Principales* es una técnica estadística o de aprendizaje no supervisado, que se utiliza para reducir la dimensionalidad de un conjunto grande de variables a conjunto más pequeño. Esto permite la visualización de los datos. Los componentes principales son los eigenvectores de la matriz de correlación o covarianza dependiendo de los datos. A partir de estos se pueden realizar transformaciones para representar los datos en un nuevo sistema de coordenadas, obtenido por rotación y traslación del sistema original a ejes en donde se maximiza la varianza, en cada dirección [7]. El resultado de las combinaciones lineales de los eigenvectores con los datos puede ser interpretable y, en muchos casos, incluso puede generar índices al ponderar variables.

## 3.3. Análisis de Correlación Canónica

El *Análisis de Correlación Canónica Simétrico* es una técnica estadística multivariante que tiene como propósito identificar y cuantificar las asociaciones *lineales* entre dos conjuntos de variables. El objetivo es encontrar las combinaciones lineales de dichos conjuntos que maximicen la correlación entre ellos. Una vez construido el índice, esta herramienta permite medir la relación de los conjuntos alternos y ver qué tanto aportan en términos de la varianza al índice. Como son distintas magnitudes, el análisis se hará vía la matriz de correlaciones.

## 3.4. Estimación de la densidad Kernel (KDE)

La estimación no paramétrica de densidades resulta útil para el propósito de este trabajo, dado que la densidad de una variable nos permite visualizar y analizar conceptos como dispersión, simetría, varianza, entre otros. La técnica consiste en estimar la función de densidad de una variable a lo largo de su soporte. Los conceptos clave detrás de este método son: estimador de Rosenblatt de la función de densidad a partir de la función de distribución empírica, la función kernel y el ancho de banda. La idea detrás del estimador de kernel es dispersar el peso $\frac{1}{n}$ de una observación alrededor de su vecindad. El kernel controla la forma y el ancho de banda controla la dispersión de este. [6]

[1]Por sus siglas en inglés: Exploratory Data Analysis.

En este caso, la aplicación será sobre los resultados de aplicar Bootstrap y obtener una muestra del índice para cada delegación, lo cual nos permitirá construir una distribución del índice por delegación. Exploraremos distintas funciones kernel como el Gaussiano, Triangular y Epanechnikov, además algunos métodos para obtener el ancho de banda como el "normal reference distribution"(nrd) propuesto por Silverman, el método propuesto por Sheater & Jones (SJ) y su variación utilizando "plug-in directo"(SJ-dpi). [6]

### 3.5. Bootstrap y Jackknife

Bootstrap y Jackknife son métodos de remuestreo con diversas aplicaciones, principalmente, inferir y describir las características poblacionales de un estimador. En este trabajo, se busca ilustrar la variabilidad del índice y su distribución [2], así como describir el error de esta variabilidad con remuestreo Jackknife, que de otra forma no podríamos conocer.

Una forma muy útil de ilustrar esta variabilidad, es con intervalos de confianza y estimación de la densidad Kernel. En cuanto a los intervalos de confianza Bootstrap, se explorarán diferentes métodos: Básico, Normal Estándar, Percentil, Studentizado, y con Corrección de Sesgo Acelerado (BCa) [3].

[2]Ya que al ser un score de la primer componente principal, es una variable aleatoria.
[3]Por sus siglas en inglés: Bias-Corrected, accelerated.

# 4. Resultados

## 4.1. EDA

Para seleccionar las variables que mejor se adecuaran al estudio, se buscaron valores atípicos con ayuda de visualizaciones de tipo gráfica de dispersión. Se graficó cada una de las variables de respuesta contra las variables explicativas y se encontraron visualizaciones como siguen 1:

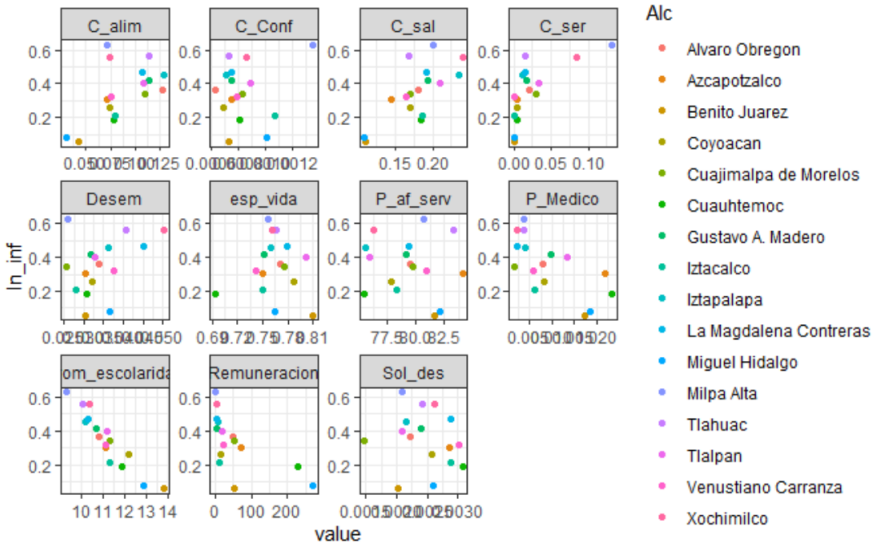
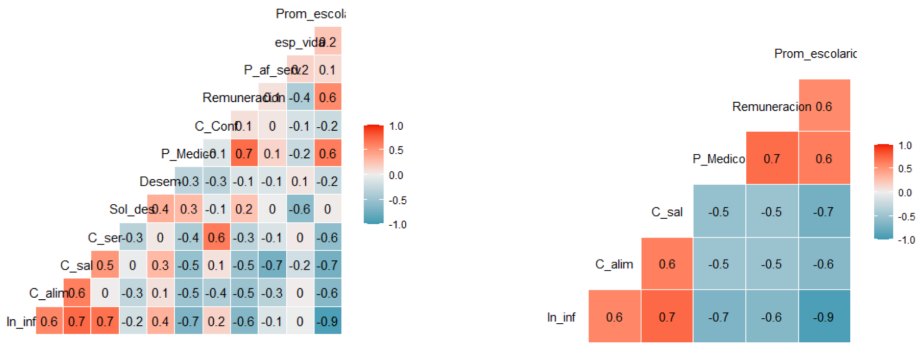



Figura 1: Relación Ingreso Inferior con el resto de variables

En todas se encontraron outliers u observaciones atípicas, sin embargo, todas están relacionados a las diferencias significativas entre las observaciones, pues no son una muestra aleatoria de unidades observacionales. Se comprobó lo anterior utilizando el respetable índice Gini que muestra que, en efecto, el grupo de outliers es sumamente distinto al resto. Las delegaciones que presentan anomalías lo hacen para todas las variables y, a su vez, pertenecen al mismo sector económico; estas son, en particular, las delegaciones Miguel Hidalgo y Cuauhtémoc, que tienen alta actividad económica; y, por otro lado, Milpa Alta, la cual tiene la menor población, tiene poca actividad económica.

Ahora bien, para seleccionar las variables finales enlistadas en la sección 2. Datos, se visualizaron las correlaciones entre las variables de respuesta [2]. Al notar que muchas de estas variables no aportarían variabilidad al índice que se construirá, se decidió categorizar las variables de desempleo y casos totales confirmados para obtener visualizaciones que sean de mayor ayuda.




(a) Variables iniciales (b) Variables utilizadas

Figura 2: Correlación entre variables

Se cuestionó, ¿qué nivel de desempleo tiene la remuneración más baja? Antes de dividir las variables de respuesta entre la población, se encontró una discrepancia entre el nivel de desempleo y la remuneración promedio (entre más desempleo, mayor remuneración), y fue cuando se decidió dividir entre la población total por delegación. Se encontró el comportamiento de las remuneración según el nivel de desempleo [3]:

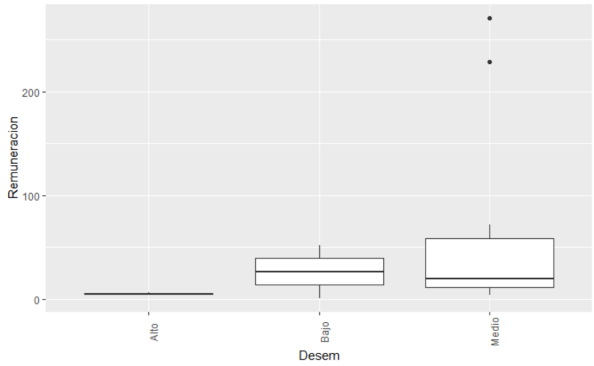



Figura 3: Remuneración para cada nivel de desempleo

Se observaron dos outliers para las observaciones de las delegaciones Cuauhtémoc y Miguel Hidalgo, se atribuyeron a que se clasificó con un desempleo medio porque está en la frontera de ser bajo. Los demás descubrimientos fueron consistentes: por ejemplo, la remuneración promedio más baja fue para las delegaciones que tienen desempleo alto. Por todo esto se concluyó que entre mayor sea el desempleo, menor será la remuneración.

Para encontrar si existe alguna relación entre el personal afiliado y la remuneración, se creó una gráfica de dispersión con burbujas cuyo tamaño indica el tamaño de la población de la delegación y el color la delegación correspondiente [4].

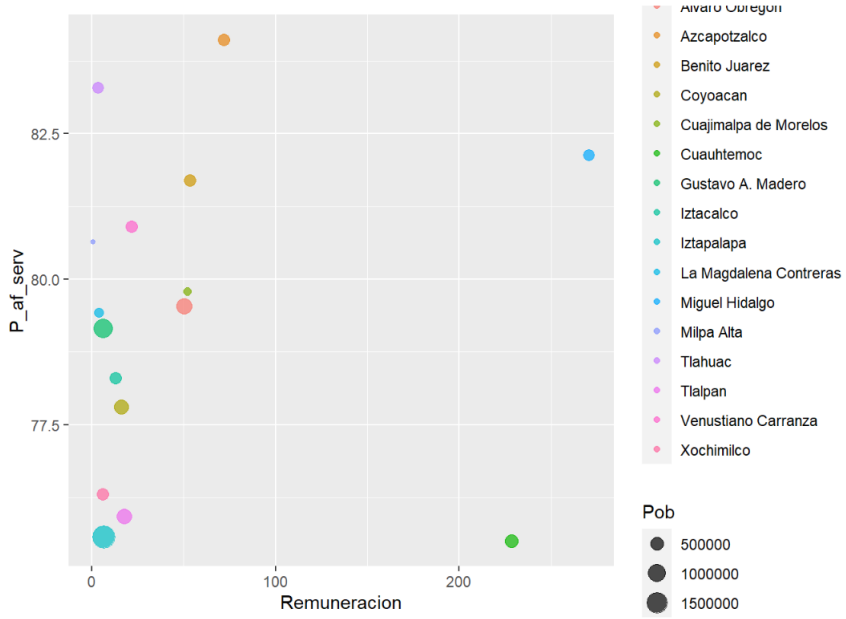



Figura 4: Personas afiliadas en servicios según la remuneración que reciben por delegación

No parece existir relación alguna y los valores extremos se dieron, una vez más, por las diferencias en niveles socioeconómicos en las delegaciones Cuauhtémoc y Miguel Hidalgo.

Se buscó si están relacionadas las carencias. Ya que se tenían datos de carencias de salud, de servicios y alimenticias, se construyó una visualización para encontrar su relación [5], empero se encontró que solo las alimenticias y las de salud están un tanto relacionadas, pero no las de servicios.

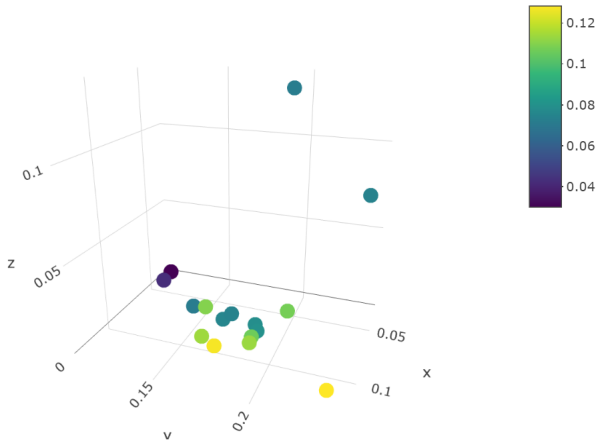



Figura 5: Comparación de las carencias

Se consideró que existía alguna relación entre el nivel de casos confirmados y la remuneración, pues la segunda trae otras cuestiones que ponen en riesgo a las personas. Por ejemplo, la remuneración determina la salud de las personas, los alimentos que ingieren, el trabajo que tienen, entre cosas, y esto, a su vez, determina sus probabilidades de enfermarse, de quedarse sin empleo o de conseguir alimentos. Eso se confirmó con la siguiente visualización en la figura [6].

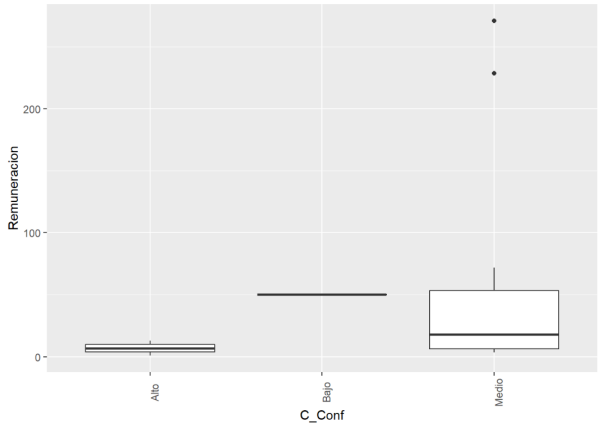



Figura 6: Remuneración para cada nivel de casos confirmados

Otra pregunta que se planteó sobre los datos fue ¿qué grupo de desempleo tiene más varianza en la esperanza de vida? Queríamos ver las observaciones de esperanza de vida para cada grupo de desempleados [7]. Encontramos

que la CDMX es una región donde la mayoría de las delegaciones tienen desempleo medio y a su vez la esperanza de vida más variada.

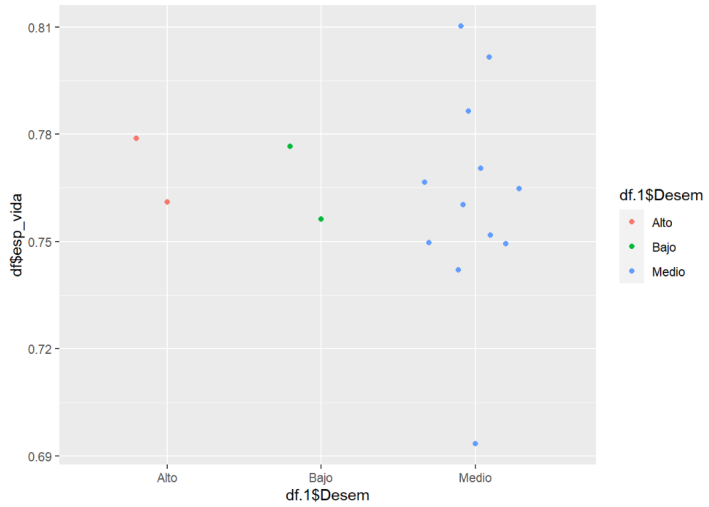



Figura 7: Esperanza de vida para cada grupo de desempleo

## 4.2. Análisis de Componentes Principales

### 4.2.1. Obtención de los Componentes Principales

Para simplificar la notación, se nombraron las variables de la siguiente forma:

$$
\begin{aligned}
X1 &:= \text{Ingresos Inferiores.} \\
X2 &:= \text{Carencia de Alimentos.} \\
X3 &:= \text{Carencia de Salud.} \\
X4 &:= \text{Personal médico.} \\
X5 &:= \text{Remuneración (miles de pesos).} \\
X6 &:= \text{Promedio años de escolaridad.}
\end{aligned}
$$

Para calcular los componentes principales, a pesar de que la mayoría de las variables estaban en términos poblacionales, se utilizó la matriz de correlación $R$ dado que las unidades de las variables eran distintas.

$$
R = \begin{pmatrix}
1.000 & 0.580 & 0.730 & -0.682 & -0.641 & -0.934 \\
0.580 & 1.000 & 0.623 & -0.504 & -0.517 & -0.632 \\
0.730 & 0.623 & 1.000 & -0.547 & -0.517 & -0.738 \\
-0.682 & -0.504 & -0.547 & 1.000 & 0.713 & 0.633 \\
-0.641 & -0.517 & -0.517 & 0.713 & 1.000 & 0.560 \\
-0.934 & -0.632 & -0.738 & 0.633 & 0.560 & 1.000
\end{pmatrix}
$$

Una vez construida la matriz, se procedió con el cálculo de los eigenvectores para obtener los componentes principales. Y se obtuvo lo siguiente:

$$
\begin{aligned}
Y_1 &= -0.449X_1 - 0.371X_2 - 0.405X_3 + 0.395X_4 + 0.380X_5 + 0.442X_6 \\
Y_2 &= -0.116X_1 - 0.307X_2 - 0.390X_3 - 0.531X_4 - 0.619X_5 + 0.275X_6 \\
Y_3 &= -0.415X_1 + 0.813X_2 - 0.034X_3 + 0.059X_4 - 0.200X_5 + 0.349X_6 \\
Y_4 &= -0.243X_1 - 0.309X_2 + 0.806X_3 + 0.111X_4 - 0.249X_5 + 0.348X_6 \\
Y_5 &= 0.211X_1 - 0.043X_2 - 0.181X_3 + 0.739X_4 - 0.598X_5 - 0.135X_6 \\
Y_6 &= 0.713X_1 + 0.097X_2 - 0.010X_3 + 0.016X_4 + 0.114X_5 + 0.684X_6
\end{aligned}
$$

### 4.2.2. Análisis e Interpretación

Una vez obtenidos los componentes principales $Y_i$ para $i = 1...6$ Se hizo la gráfica de codo correspondiente, para determinar cuáles son las componentes más significativas. Para realizar esto, se vio qué tanta varianza explicada aporta cada componente principal. Este resultado es obtenido con $\frac{\lambda_k}{p}$. Donde $\lambda_k$ es el k-ésimo eigenvalor asociado a la k-ésima componente principal y $p$ es la suma de la diagonal de la matriz de covarianza $R$.

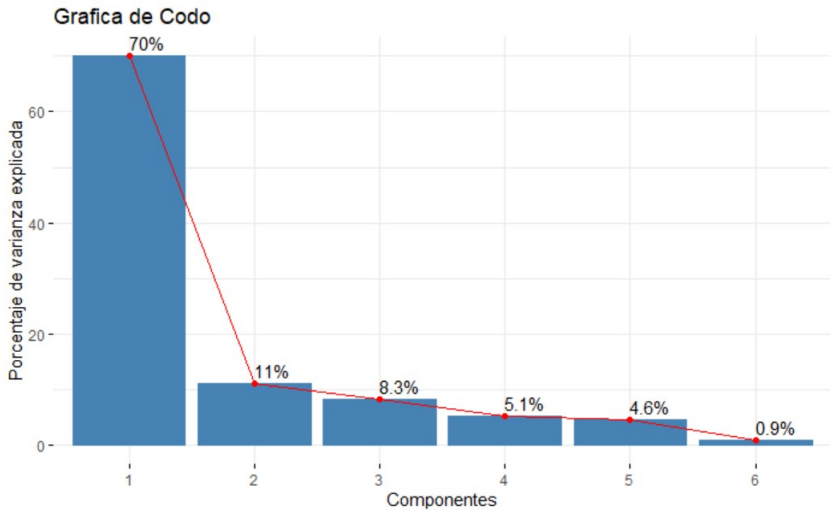



Figura 8: Gráfica de codo

Podemos ver que el primer componente principal $Y_1$ aporta un 70 % de variabilidad. Por lo cual, es factible decir que resume de buena manera los datos.

Al buscar la interpretación de la primera componente principal se destacó lo siguiente: los coeficientes de las variables $X_1$, $X_2$,$X_3$ asignadas a *Ingresos Inferiores*, *Carencia de Alimentos*, *Carencia de Salud* tienen signo negativo. Contrastándolo con los coeficientes correspondientes a las variables $X_4$, $X_5$,$X_6$ asignadas a *Personal médico* , *Remuneración*, *Promedio años de escolaridad* tienen signo positivo. De donde se infiere que se pondera

negativamente las variables que representan a una persona más vulnerable. Y, al contrario, se pondera positivamente las variables que representan a una persona menos vulnerable. Es por esto que usaremos esta variable para para medir la vulnerabilidad de la población.

La segunda componente principal no forma parte del índice, pero es de conveniencia hacer una interpretación sobre esta para el análisis. $Y_2$ puede verse como una variable que separa en categorías de salud y economía en un bloque con signo negativo, y pondera positivamente la variable $X_6$ *Promedio de años de escolaridad*.

Una vez hecho esto, podemos visualizar las componentes principales como proyecciones en el espacio de las dos primeras componentes principales.

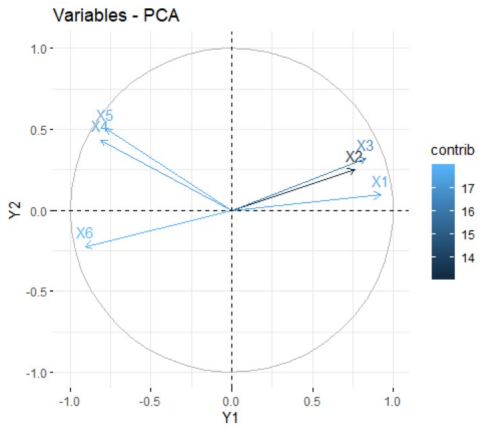



Figura 9: Variables

Se observa que el conjunto de variables $X_2$, $X_3$ apuntan prácticamente a la misma dirección mientras que $X_1$ apunta en el mismo sentido pero con unos cuantos grados de diferencia. Por otro lado, $X_4$, $X_5$,$X_6$ apuntan completamente en sentido opuesto y se ve que, en particular, $X_6$ es totalmente opuesta a $X_2$ y $X_3$. También, se ve que la variable $X_2$ es la que tiene una menor contribución.

Para ver que tanta aportación proporcionan cada una de las variables de las componentes principales se proporciona la siguiente gráfica.

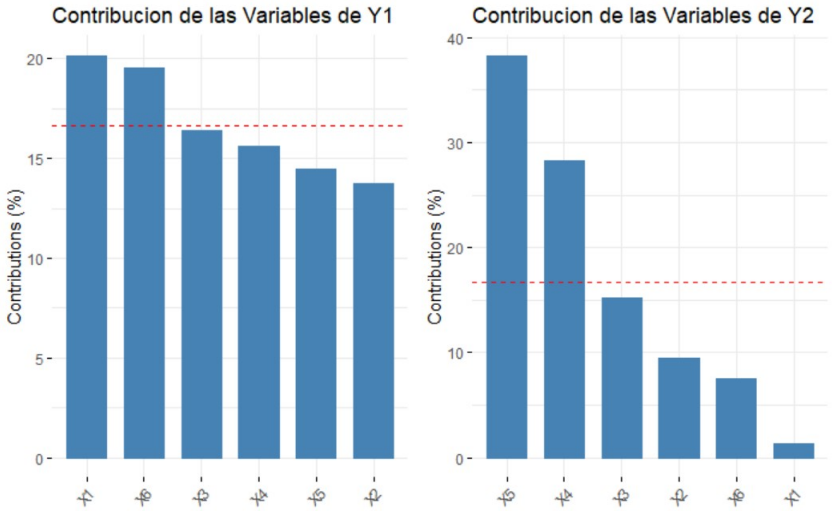



Figura 10: Contribuciones

Es factible observar que para la primer componente principal, es decir para nuestro índice, las dos variables que más aportan son $X_1$ y $X_6$ correspondientes a *Ingresos Inferiores* y *Promedio años de escolaridad* .Mientras que, para la segunda dimensión son las variables $X_5$ y $X_4$, correspondientes a *Remuneración* y *Personal médico*.

Una vez analizadas las componentes, podemos plasmar los datos en dos dimensiones. De esta manera se puede visualizar el comportamiento de las delegaciones con respecto a las primeras dos componentes principales $Y_1$ y $Y_2$.

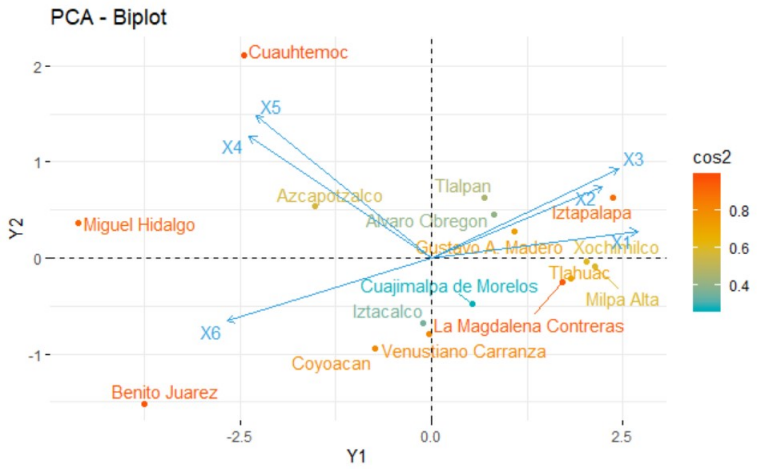



Figura 11: Biplot Delegaciones

Podemos ver que a medida en que estén hacia la derecha de la gráfica, las delegaciones son más vulnerables, mientras que, si se colocan más hacia la izquierda son menos vulnerables. Podemos ver que las delegaciones *Miguel Hidalgo*, *Benito Juárez* y *Cuauhtemoc* son las menos afectadas. Por otro lado, las más afectadas serian *Milpa alta*, *Iztapalapa* y *Magdalena Contreras*.

### 4.2.3. Formulación del Índice

Una vez obtenido el índice de vulnerabilidad (**IV**) se le aplicó una transformación para facilitar la interpretación del mismo.

$$\mathbf{IV} = \frac{1}{Y_1} * 100$$

Dónde $Y_1$ es el primer componente principal. Con la última transformación, entre mayor sea la vulnerabilidad mayor sera el índice.

A continuación, se presenta el mapeo de la Ciudad de México y la tabla (2), donde se ilustran los scores correspondientes a las delegaciones. Entre más oscuro, quiere decir que esa delegación es mal vulnerable, análogamente si es más claro quiere decir que es menos vulnerable.

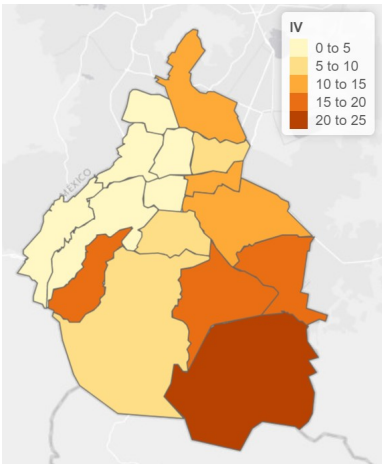



Figura 12: Mapa Ciudad México

| **Delegacion** | **Score** |
|---|---|
| Álvaro Obregón | 4.237 |
| Azcapotzalco | 3.120 |
| Benito Juárez | 3.792 |
| Coyoacán | 8.809 |
| Cuajimalpa de Morelos | 4.082 |
| Cuauhtémoc | 1.087 |
| Gustavo A. Madero | 14.688 |
| Iztacalco | 10.308 |
| Iztapalapa | 14.886 |
| La Magdalena Contreras | 17.383 |
| Miguel Hidalgo | 0.921 |
| Milpa Alta | 24.680 |
| Tláhuac | 18.251 |
| Tlalpan | 8.728 |
| Venustiano Carranza | 7.704 |
| Xochimilco | 15.310 |

Cuadro 2: Scores

Se puede ver, cómo lo mencionamos en la sección de análisis de componentes principales, que las delegaciones más vulnerables son *Milpa alta*, *Iztapalapa*, *Magdalena Contreras* y las menos vulnerables *Miguel Hidalgo* y *Cuauhtémoc*.

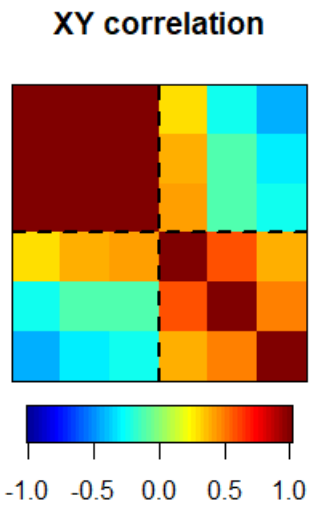
Para finalizar con este análisis, podemos concluir que, ante los scores de nuestro índice, las delegaciones más

afectadas son aquellas en donde hay mayor carencia de servicios públicos y menor actividad económica. Sin embargo, es importante destacar que, al haber convertido nuestras variables a términos poblacionales, como era de esperarse, las delegaciones con menor población reflejaban mayor carencia de servicios y menor actividad económica. Por lo mismo estas fueron ponderadas como delegaciones muy vulnerables. Es el caso en particular de *Milpa Alta* que esta categorizada como la delegación en peores condiciones. Se propone como trabajo futuro, hacer un análisis análogo a este sin tomar en cuenta esta delegación ya que podría estar afectando los resultados del estudio al ser un posible *outlier*.

## 4.3. Análisis de Correlación Canónica

### 4.3.1. Correlaciones *a priori*

En primer lugar, se quiso observar la correlación de entre todas las variables, ya sea para su propio conjunto o para comparar la correlación cruzada marginal de una variable con otra de otro grupo. En la siguiente imagen se puede analizar cómo la relación lineal entre el primer conjunto es altísima y todas en el mismo sentido, pues las correlaciones entre variables son mayores al 97 % [4], mientras que, en el otro grupo las relaciones, a pesar de ser todas positivas, sí varían en magnitud. Específicamente para la correlación cruzada entre grupos, se observa que solo la variable *personal médico* se relaciona positivamente con el grupo de carencias, mientras que las demás variables tiene relaciones negativas. Esto sugirió el cambio de signo dentro de la primera componente principal. Aunado a esto, se empieza a tener una mejor interpretación del índice, pues al ser cien veces el inverso de los scores del primer componente principal, comunica que entre mayor sea el puntaje para el grupo de carencias, más vulnerable se es y viceversa.




[4]Debido a que estas variables fueron tomadas del censo del CONEVAL y se usaron las mismas muestras para las distintas variables, es muy probable que existe colinealidad en este grupo. Sin embargo, a falta de más instituciones que realicen este tipo de censos, no se pudo prescindir de otras variables.

### 4.3.2. Variables canónicas

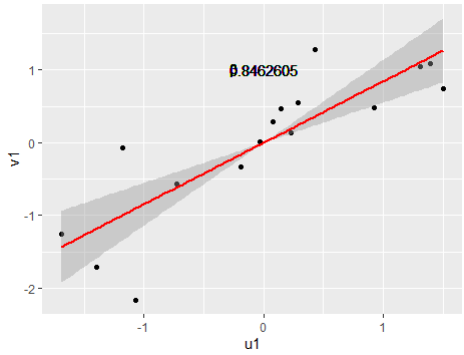
Como los datos muestran distintas escalas y tienen distintas medidas, se optó por hacer el análisis con la matriz de correlaciones, como en el caso de componentes principales, ya que esta escala cada uno de los datos y por ende se quitan las magnitudes, pero también hace que las aportaciones a las variables canónicas sean unitarias.

Una vez realizadas las combinaciones lineales intragrupos $U = \sum_{k=1}^{3} a_k X_k$, $V = \sum_{k=4}^{6} a_k X_k$ que maximizan la correlación $\rho = \frac{a' R_{XY} b}{\sqrt{a' R_X a}\sqrt{b' R_Y b}}$, sujetas a que $a' R_X a = 1$ y $b' R_Y b = 1$ , se llegaron a las siguientes variables canónicas primarias:

$$U = (1.9948X_1 - 0.5256X_2 - 3.1031X_3) \times 10^{-5},$$
$$V = 8.6266 \times 10^{-4} X_4 - 1.3924 \times 10^{-7} X_5 - 0.7255X_6.$$

Se observó que en la primer variable canónica $U$, la *carencia de salud* $X_3$ es la que mayor representación tiene y por ende la que más aporta a la relación en el grupo de carencia. En la otra variable canónica $V$, se observó sin lugar a dudas que el *promedio de años de escolaridad* es la que mayor aporta al grupo de opulencias.

Ahora bien, la correlación canónica $\rho_{U,V}$ fue del 0.8462, la cual es bastante alta. A continuación, se muestra esta relación y el impacto que tiene un cambio en $U$ en $V$. Por lo que, un modelo de regresión lineal simple sin intercepto, captura la misma correlación. La siguiente regresión muestra dicha relación:




Entonces podemos interpretar que un cambio de magnitud unitaria en el grupo de carencias generará un cambio del 84 % en el grupo de opulencias y para el caso contrario se tendrá que un aumento unitario en el grupo de opulencias generará un cambio del $\beta^{-1} = 118\,\%$ en el grupo de carencias con los datos actuales.

### 4.3.3. Correlación entre variables originales y canónicas

A continuación, se muestran las relaciones lineales marginales intragrupos y extragrupos con respecta a las variables canónicas.

**Grupo de carencias**

A continuación, se muestran las correlaciones de Pearson entre *ingresos inferiores* $X_1$, *carencia de alimentos* $X_2$, *carencia de salud* $X_3$ y las tres variables canónicas del presente grupo $U_1, U_2, U_3$ , como también con las tres variables canónicas del otro grupo, respectivamente.


Figura 13: $\rho_{X_i,U_j} \quad \forall i,j \in \{1,2,3\}$


Figura 14: $\rho_{X_i,V_j} \quad \forall i,j \in \{1,2,3\}$

Figura 15: Correlaciones del grupo carencias con las variables canónicas

Como se mencionó anteriormente, al estar íntimamente relacionadas este grupo entre sí, las correlaciones marginales con las variables canónicas son muy parecidas para cada variable original. Pero es de observarse que, a pesar que la *carencia de salud* es la que numéricamente aporta más a la primer variable canónica, la que mayor correlación tiene son los *ingresos inferiores* y por lo dicho anteriormente, también la que más se relaciona linealmente con la primer variable canónica del grupo de opulencias.

**Grupo de opulencias** Para el grupo de opulencias se realizó el mismo análisis que para el grupo de carencias.

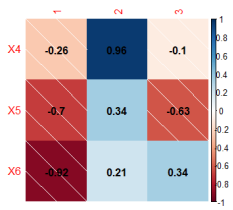



Figura 16: Correlación entre opulencias y variables canónicas del mismo grupo

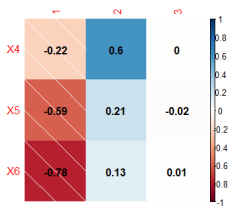



Figura 17: Correlación entre opulencias y variables canónicas de carencias

Se concluyó que la relación de este grupo con cada una de las variables opulentas es inversa, lo cual tiene mucho sentido, ya que al tener una correlación canónica positiva pero marginalmente negativa, se dice que el aumento en $X_{4,5,6}$ será inversamente proporcional. Esto se traduce en el índice que un valor alto en cualquiera de estas tres variables hará que tu índice (vulnerabilidad) sea menor. Además, el *promedio de años de escolaridad* son los datos que mayor relación tiene con este grupo, por lo que específicamente esta variable impacta con mayor fuerza en la vulnerabilidad de las delegaciones.

## 4.4. Remuestreo para los scores de las delegaciones

### 4.4.1. Procedimiento

Después de la obtención de los índices vía PCA se buscó observar cuánto variaban los mismos scores para distintas muestras de la CDMX. Por lo que se remuestrearon los índices para cada delegación vía *boostrap* para poder estimar la densidad de cada índice para cada aclaldía. El procedimiento fue el siguiente:

**Data:** Matriz original de las delegaciones **X**, semilla **S**
**Result:** $Y_* \in \mathbb{R}^{B\times n}$ matriz de scores para cada delgación (columnas).
seed = S;
**for** $i \to B$ **do**
1. Remuestrear con reemplazo los renglones de X, obteniendo $X_i^*$;
2. Obtener la primera componente principal de $X_i^*$;
3. Guardar los i-ésimos *scores* en $Y_i^*$;

**end**

**Algorithm 1:** Bootstrap de los índices

La semilla que se usó fue la 42 y se usó un tamaño 200 para la muestra bootstrap, como sugiere Efron y De la Góngora, debido a que el aporte a partir de la iteración 201 en adelante es raquítica y visualmente generaba ruido.

### 4.4.2. Ajuste de kernel a los scores

Como se adelantó en la explicación de la metodología, se comenzó tomando una delegación experimental y probamos tres funciones kernel para la estimación de la densidad. Después, se eligió un kernel y se probaron los tres métodos para la selección del ancho de banda. Por último, con el kernel y método elegidos se realizó la estimación de la densidad del índice para las dieciséis delegaciones. Solo para efectos de mostrar la experimentación, se eligió la delegación Álvaro Obregón.

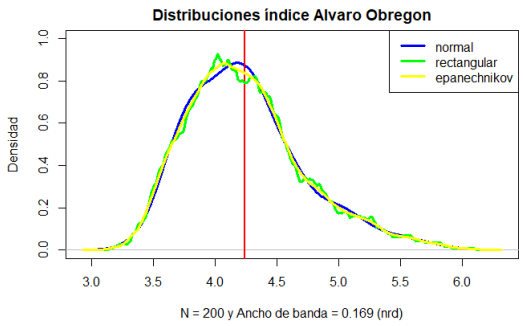



Figura 18: Distribuciones primera iteración

En esta gráfica se puede observar el ajuste para los distintos kerneles, la línea vertical roja representa el valor

del índice para dicha delegación. Se utilizó el método *nrd* para encontrar el ancho de banda. Es importante recordar que se utilizó un tamaño de muestras Bootsrap $B = 200$, por esto se observa que las tres distribuciones son muy similares, la única diferencia se presenta en los picos. Se eligió el kernel Gaussiano pues el pico de la distribución y se presentó casi en el mismo nivel que el valor del índice. Fijando dicho kernel, se probaron los tres métodos para calcular el ancho de banda.

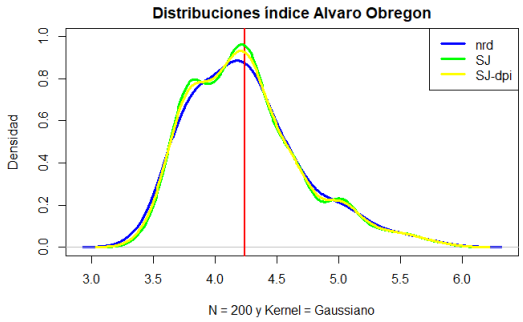



Figura 19: Distribuciones segunda iteración

Es factible observar que ahora el pico se pronuncia más con el método $SJ$ y su variación. Además, se pierde suavidad en la estimación. Es importante mencionar que la elección del kernel no es tan relevante, pero la elección del ancho de banda sí lo es. Esta última busca suavizar lo suficiente para eliminar saltos insignificantes y no suavizar demasiado para ocultar picos reales. [6]

Por los resultados anteriores, se eligió el kernel Gaussiano y con el método $SJ$, pues logra el balance entre suavizar y no ocultar picos reales. A continuación, se presenta la estimación de la densidad para el índice de cada una de las delegaciones.

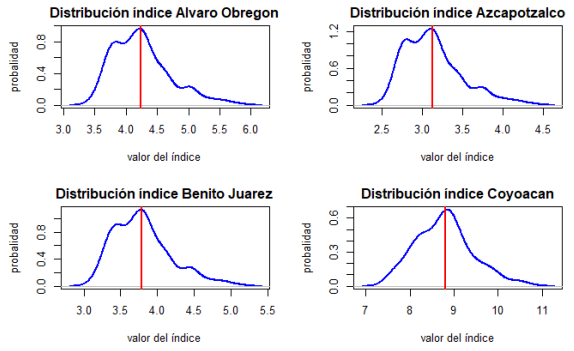



Figura 20: Alcaldías 1 a 4

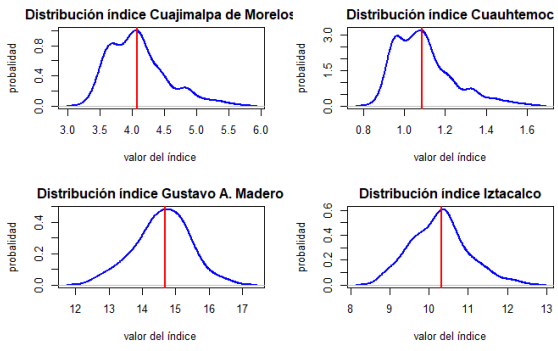



Figura 21: Alcaldías 5 a 8

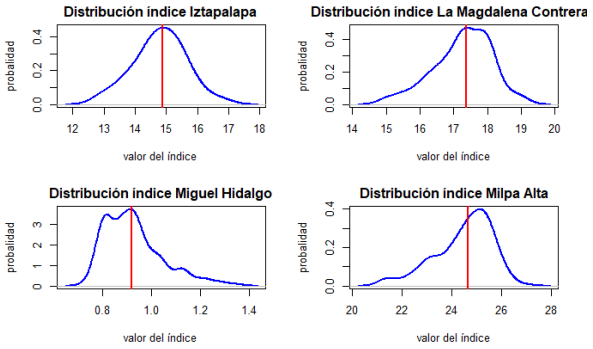



Figura 22: Alcaldías 9 a 12

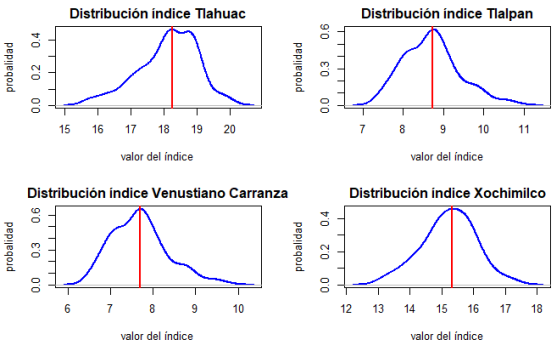



Figura 23: Alcaldías 13 a 16

Lo más interesante de estos resultados es la gran similitud entre el pico más alto de las distribuciones estimadas y los valores del índice, con excepción de Milpa Alta. Esto ayudó a validar la robustez de los resultados, en cuanto al valor del índice obtenido. Por otro lado, la varianza de algunas distribuciones es más amplia. Para el caso de Gustavo A. Madero, Iztapalapa y Xochimilco la distribución parece una campana gaussiana, con simetría casi perfecta y el valor índice parece representar la media de la distribución. Estas delegaciones presentan niveles del índice medios-altos. Para el caso de Coyoacán, Iztacalco, Tlalpan y Venustiano Carranza, se preserva la simetría y representan los niveles del intermedio del índice. Los otros casos pierden simetría y se presenta asimetría. Para la Álvaro Obregón, Azcapotzalco, Benito Juárez, Cuajimalpa, Cuauhtémoc y Miguel Hidalgo, se presenta asimetría positiva y los niveles más bajos del índice. Para la Magdalena Contrera, Milpa Alta y Tláhuac se presenta asimetría negativa y los niveles más altos del índice. Esto nos da indicios de una posible clasificación por rangos para las delegaciones.

### 4.4.3. Intervalos Bootstrap y Jackknife

Otra medida para comprender la variabilidad del índice fue computar Intervalos de Confianza Bootstrap (IC) al, con $\alpha = 0.10$, a partir del remuestreo descrito en la sección 4.4.1. Se consideraron diferentes métodos para observar cómo cambiaban los intervalos. Recalcamos que todos los intervalos obtenidos en remuestreo contienen a nuestro estimador de la muestra original.

**Intervalos de Confianza Bootstrap 90%**

| | A. Obregón | Azcapotzalco | B. Juárez | Coyoacán |
|---|---|---|---|---|
| **INDICE** | **4.24** | **3.12** | **3.79** | **8.81** |
| **Básico** | ( 3.34 , 4.83 ) | ( 2.40 , 3.57 ) | ( 3.05 , 4.32 ) | ( 7.63 , 9.86 ) |
| **Normal** | ( 3.47 , 4.99 ) | ( 2.52 , 3.70 ) | ( 3.14 , 4.43 ) | ( 7.78 , 9.90 ) |
| **Percentil** | ( 3.64 , 5.13 ) | ( 2.67 , 3.84 ) | ( 3.27 , 4.53 ) | ( 7.76 , 9.99 ) |
| **Bca** | ( 3.74 , 5.59 ) | ( 2.74 , 4.20 ) | ( 3.36 , 4.91 ) | ( 8.01 , 10.45 ) |
| **Student** | ( 3.74 , 6.40 ) | ( 2.79 , 4.05 ) | ( 3.38 , 5.61 ) | ( 7.94 , 13.02 ) |
| | **Cuajimalpa** | **Cuauhtémoc** | **G. A. Madero** | **Iztacalco** |
| **INDICE** | **4.08** | **1.09** | **14.69** | **10.31** |
| **Básico** | ( 3.22 , 4.65 ) | ( 0.79 , 1.25 ) | ( 13.55 , 16.33 ) | ( 9.06 , 11.54 ) |
| **Normal** | ( 3.34 , 4.80 ) | ( 0.85 , 1.31 ) | ( 13.45 , 16.11 ) | ( 9.18 , 11.52 ) |
| **Percentil** | ( 3.51 , 4.94 ) | ( 0.92 , 1.38 ) | ( 13.05 , 15.83 ) | ( 9.07 , 11.55 ) |
| **Bca** | ( 3.60 , 5.38 ) | ( 0.95 , 1.52 ) | ( 13.42 , 16.25 ) | ( 9.35 , 12.03 ) |
| **Student** | ( 3.50 , 5.64 ) | ( 1.04 , 1.20 ) | ( 13.38 , 21.49 ) | ( 9.19 , 15.25 ) |
| | **Iztapalapa** | **M. Contreras** | **M. Hidalgo** | **Milpa Alta** |
| **INDICE** | **14.89** | **17.38** | **0.92** | **24.68** |
| **Básico** | ( 13.63 , 16.61 ) | ( 16.20 , 19.19 ) | ( 0.67 , 1.06 ) | ( 23.38 , 27.07 ) |
| **Normal** | ( 13.56 , 16.40 ) | ( 16.08 , 18.93 ) | ( 0.72 , 1.11 ) | ( 22.96 , 26.77 ) |
| **Percentil** | ( 13.17 , 16.14 ) | ( 15.57 , 18.57 ) | ( 0.78 , 1.17 ) | ( 22.29 , 25.98 ) |
| **Bca** | ( 13.57 , 16.64 ) | ( 15.93 , 18.80 ) | ( 0.80 , 1.29 ) | ( 22.33 , 26.00 ) |
| **Student** | ( 12.79 , 18.73 ) | ( 15.33 , 21.54 ) | ( 0.89 , 1.00 ) | ( 17.07 , 30.21 ) |
| | **Tláhuac** | **Tlalpan** | **V. Carranza** | **Xochimilco** |
| **INDICE** | **18.25** | **8.73** | **7.70** | **15.31** |
| **Básico** | ( 17.03 , 20.13 ) | ( 7.46 , 9.81 ) | ( 6.52 , 8.68 ) | ( 14.12 , 17.01 ) |
| **Normal** | ( 16.91 , 19.85 ) | ( 7.61 , 9.90 ) | ( 6.64 , 8.80 ) | ( 14.01 , 16.80 ) |
| **Percentil** | ( 16.37 , 19.47 ) | ( 7.64 , 9.99 ) | ( 6.72 , 8.89 ) | ( 13.61 , 16.50 ) |
| **Bca** | ( 16.69 , 19.69 ) | ( 7.90 , 10.61 ) | ( 6.94 , 9.52 ) | ( 13.98 , 16.95 ) |
| **Student** | ( 15.65 , 22.97 ) | ( 7.62 , 12.19 ) | ( 6.63 , 12.65 ) | ( 13.37 , 19.77 ) |

Figura 24: Intervalos de Confianza Bootstrap

Ahora bien, nótese que los intervalos *Studentized* en general son más amplios que el resto, esto de debe a que este método consiste en Remuestreo Bootstrap anidado, y así, considera el error estándar del error estándar. Es decir, este intervalo contempla aún más variabilidad en para cada índice. Haciendo un estudio Jackknife After Bootstrap este error detrás de los intervalos *Studentized*:

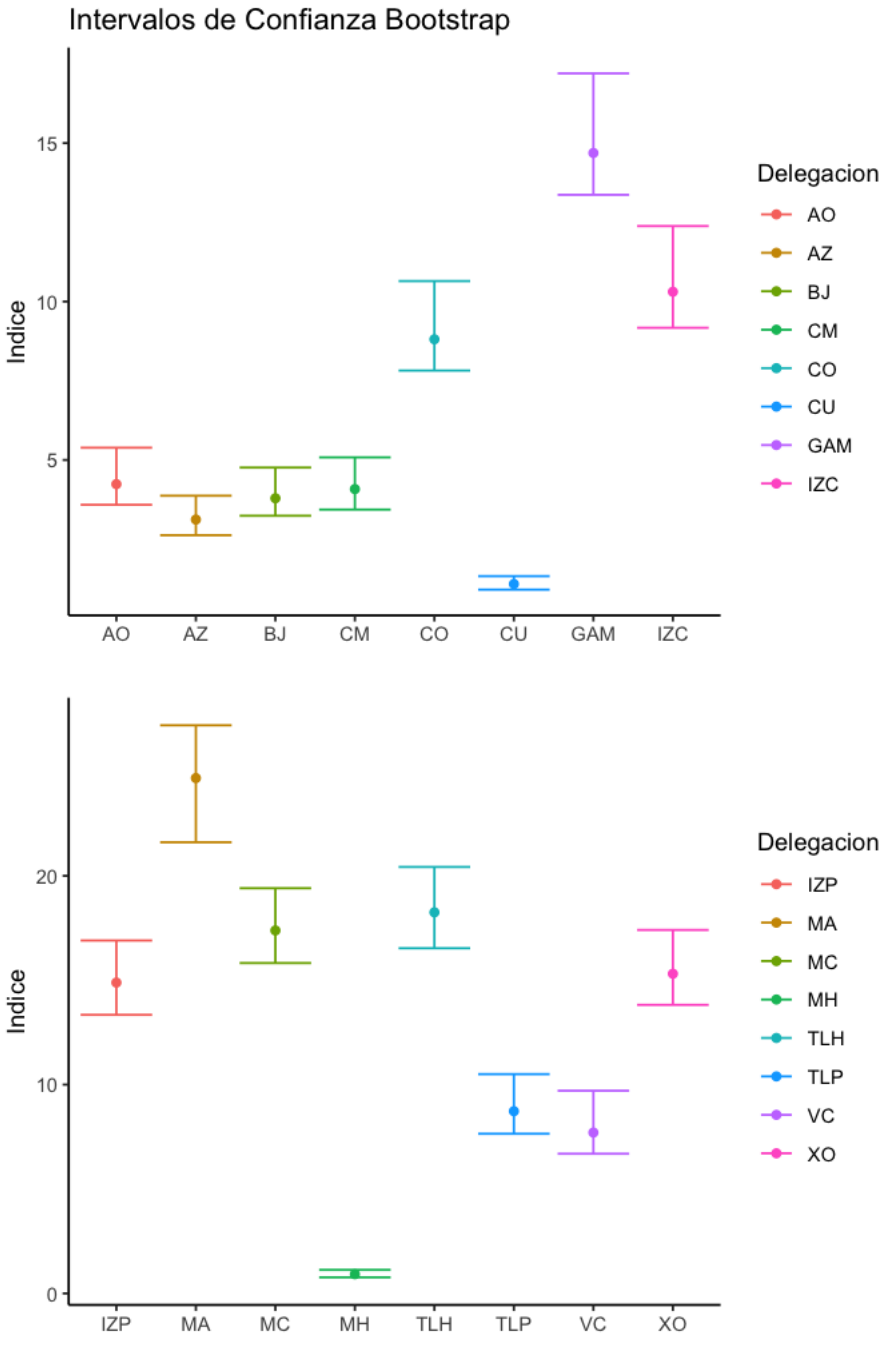



Figura 25: Intervalos de Confianza Bootstrap Promedio (alpha = .1)

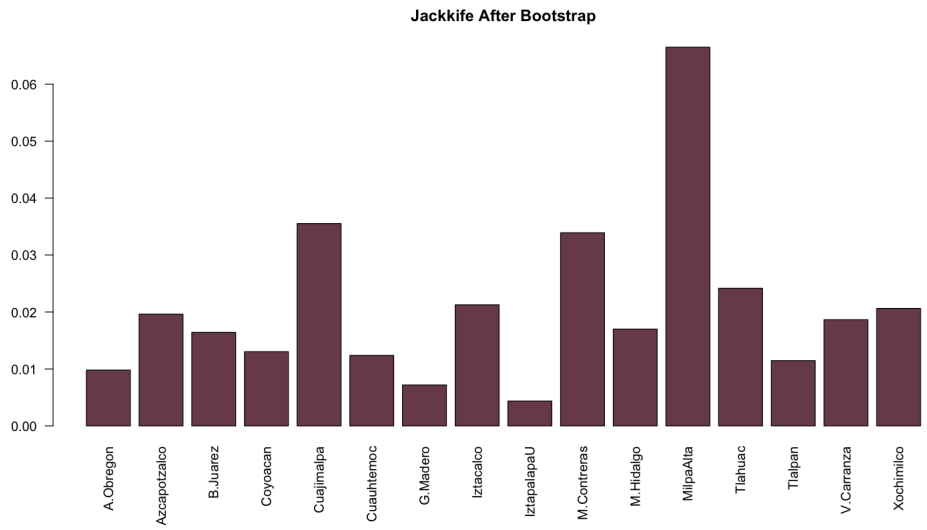



Figura 26: Jackknife After Bootstrap. Error estándar del error estándar para Remuestreo Bootstrap

Observamos en especial Milpa Alta, cuyo intervalo sufre el cambio más dramático con el último método, y comparando con la Figura 25, vemos que el análisis Jackknife es congruente.

## 4.5. Comparaciones

Dado que la crisis sanitaria sigue presente, la situación nos permite hacer comparaciones entre el índice y otros indicadores que muestran la vulnerabilidad de la población. Primero, tomamos el número de casos confirmados de Covid-19 por delegación, con corte del día 19 de mayo. Cómo las variables para la construcción del índice fueron tomadas en términos de porcentaje poblacional, se realizó lo mismo para el número de casos y obtenemos el porcentaje de población contagiada. Después, se realizó la comparación entre el índice por delegación y el porcentaje de población contagiada.

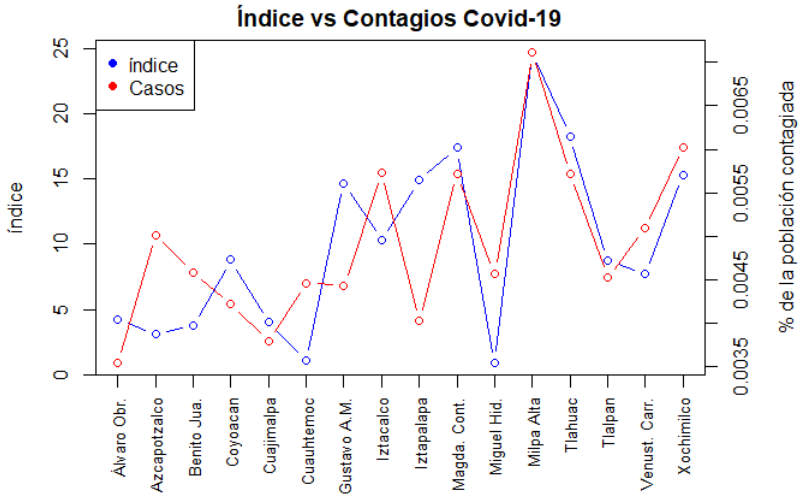



Figura 27: Índices vs Contagios Covid-19

Los puntos azules representan el valor del índice obtenido para cada delegación y los rojos el porcentaje de personas contagiadas. Se puede observar que la dinámica de ambas variables coincide en algunas delegaciones en términos relativos, pues las magnitudes son distintas. Lo que sí es claro es que Milpa Alta presenta el valor máximo del índice, es decir, es la delegación más vulnerable, así como el mayor porcentaje de personas contagiadas. Por otro lado, si se divide la gráfica en tres secciones equiespaciadas se nota que casi todas las delegaciones coinciden en la magnitud del índice y la magnitud de contagios. Esto confirma la validez del índice y es una primera muestra de que se logró condensar la situación real de cada comunidad.

Para la segunda comparación, se utilizaron las solicitudes del seguro de desempleo que se han registrado desde que inició la contingencia hasta el mismo corte del 19 de mayo. Para ponerla en términos porcentuales haremos la división usando el número de personas económicamente activas, pensando que esas son el total de personas que tenían un empleo.

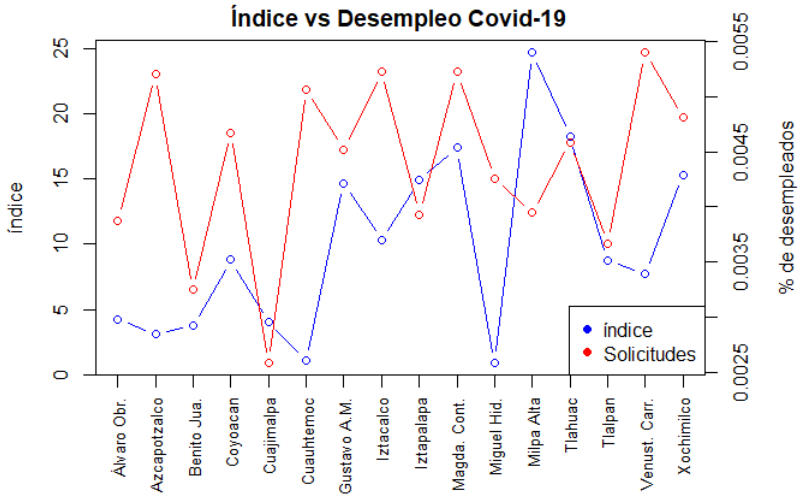



Figura 28: Índices vs Desempleo Covid-19

Se observa que en términos porcentuales casi todas las delegaciones presentan un porcentaje de desempleo en un mismo rango. Esto se puede deber al origen de los datos o, quizá, no tiene por qué coincidir el número de solicitudes de desempleo con el número de personas que en realidad han perdido su trabajo. El análisis se deja para trabajos futuros.

# 5. Discusión

Las limitaciones del trabajo radican en la calidad y la cantidad de los datos, pues hay datos que no están actualizados día a día, otros que no parecían coherentes con la realidad y algunos otros que estaban incompletos. Para disminuir dicho problema elegimos a las delegaciones de la Ciudad de México como unidades observacionales, sin embargo, como ya se admitió, el movimiento de entrada y salida de personas de la ciudad es muy grande y los cálculos solo se realizaron con información de los residentes de las delegaciones.

Una vez realizado este estudio, si se estuviesen al alcance datos de calidad que proporcionen información sobre la vulnerabilidad del país, sería posible realizar el estudio a nivel nacional. Esto podría ser de gran beneficio, no solo para saber a que sectores dirigir políticas públicas sino que también como marco de referencia para ver si las políticas públicas han funcionado con el paso del tiempo.

Como continuación del presente trabajo, se propondrán los siguientes ejes de análisis que pudieras resultar interesantes. A partir de los estimadores del índice Bootstrap, sugerimos buscar sus respectivos UMVUE's, y, con base en ellos, ver qué delegación es la que tiene mayor vulnerabilidad ante los peligros que se incluyeron en este trabajo. Añadir, con información de otras fuentes, una réplica del mismo estudio y ver los resultados para compararlos. Por otro lado, se propone replicar el estudio para los 2,458 municipios y dieciséis alcaldías de la República. Al tener más observaciones, se pueden incluir más variables y por ende hacer un índice mucho más robusto que el que se presentó en este trabajo.

# 6. Conclusiones

La pandemia por Covid-19 es una de las crisis más grandes que ha tenido la humanidad en la historia moderna. Esto sigue y seguirá siendo por años una amenaza latente en la economía, salud y estabilidad general de los mexicanos. El presente trabajo, se presentó un índice de vulnerabilidad ante una contingencia sanitaria. Encontramos que los recursos públicos deberán de ser distribuidos dependiendo de la vulnerabilidad de la población, pues encontramos, tal como lo esperábamos, qué la desigualdad del país empuja a que la ayuda no sea necesitada por todos.

Con ayuda del análisis de componentes principales se construyó un índice multidimensional para la toma de decisiones informadas. Donde pudimos ver con respecto a nuestras variables, cuáles eran las delegaciones más vulnerables. Por otra parte se utilizaron los métodos Bootstrap y Jackkine para evaluar la calidad [5] de este índice. Se encontró que es un índice aplicable a cualquier situación en la que la población se vea vulnerada, en especial una crisis sanitaria, pero también podría aplicarse en casos de inestabilidad política. Quizá lo más complejo del trabajo fue encontrar datos de fuentes confiables que tuvieran los menos errores posibles.

Para el caso del *análisis de correlación canónica simétrico* se pudo identificar que el *promedio de años de escolaridad* es la variable que tiene mayor influencia en la variabilidad y peso del índice. Se puede observar incluso en las densidades de la delegación Miguel Hidalgo que es la que menor variabilidad presenta en el índice, pues es la que más cercana está al promedio de años de escolaridad. Se puede concluir lo mismo a partir del *análisis por componentes principales*. En la misma mano también enfatizamos que tanto el *grupo de opulencias* como de *carencias* están íntimamente relacionados [6], por lo que un cambio en cualquiera de los grupos afectará con casi la misma magnitud al siguiente. Al tener que *ingresos inferiores* y *años de escolaridad* son las que mayor aportación a la correlación canónica tienen, se sugiere concentrar esfuerzos en estas mismas: disminuyendo el número de personas con ingresos inferiores como aumentando la escolaridad en las delegaciones para bajar la vulnerabilidad.

Por otro lado, en el *KDE*, vemos que se pueden formar 4 clusters de densidades:

1. Delegaciones cuyo índice es menor a 4.5.
2. Delegaciones cuyo índice está entre 7 y 10.5.
3. Delegaciones cuyo valor del índice está alrededor del 15.
4. Delegaciones con índice entre 17 y 18.
5. Milpa Alta.

De lo anterior inferimos que entre menor sea el índice, menor variabilidad tiene la delegación y viceversa. El caso de Milpa Alta se debe a que es la delegación con menor población de la CDMX, esto ocasiona que al tomar el porcentaje poblacional sea la delegación donde mayor peso tiene cada persona. Si nos interesa saber la vulnerabilidad en términos porcentuales, podemos afirmar estadísticamente que esta última delegación debe

[5] El sesgo y la variabilidad de cada índice. Deseamos insesgamiento y la menor varianza posible. El sesgo se pudo corregir gracias al estimador Jacknife del Boostrap, sin embargo, la varianza mínima no se tomó en consideración.

[6] Se mencionó una correlación entre ambos grupos del 0.8462.

ser prioridad en la agenda política. Siguiendo el mismo hilo, el cluster uno, tiene una asimetría positiva y los clusters cuatro y cinco una asimetría negativa. Lo anterior nos llevaría a concluir que ante una crisis sanitaria los primeros clusters podrían mostrar menor vulnerabilidad que el valor calculado del índice y los últimos mayor vulnerabilidad que el índice.

# Referencias


[1] Datos CDMX(2020). *Portal de datos de la Ciudad de México*

[2] Consejo Nacional de Evaluación de la Política de Desarrollo Social(2015). *Datos Abiertos del Coneval.*

[3] Instituto Nacional de Estadística y Geografía(2015). *Encuesta intercensal 2015.*

[4] Secretaría de Salud de la Ciudad de México (2018). *Datos de personal médico*

[5] Instituto Nacional de Estadística y Geografía (2017). *Anuario estadístico y geográfico de la Ciudad de México 2017*

[6] J. de la Vega (2020). *Estimación de densidades.* Presentación parte de la clase de Estadísitca No Paramétrica Primavera 2020.

[7] J. de la Vega (2020). *Analisis de Componentes Principales.* Presentación parte de la clase de Estadísitca Aplicada III 2020.

[8] Bradley Efron y Gail Gong (1983). *A leisurely look at the bootrstap, the jacknife and cross-validation.* American Statistical Association.

# A. Apéndice A

Liga al repositorio que contiene los datos y códigos: https://github.com/pbarrancs/proyIndVul