# Análisis de Componentes Principales

Ethan Enrique Verduzco Pérez

Instituto Tecnológico y de Estudios Superiores de Monterrey
Aplicación de Métodos Multivariados a la Ciencia de Datos

12 de septiembre de 2021

**Resumen**

*En este documento se presentará la metodología empleada y el proceso que se llevó a cabo para reducir la dimensionalidad de los datos de contaminantes y mediciones de tipo meteorológicos correspondientes a la problemática del reto integrador sobre la Contaminación del Aire en el Área Metropolitana de Monterrey a través del método multivariado del análisis de componentes principales.*

***Palabras claves:*** *Componentes principales, estadística multivariada, contaminantes ambientales.*

## I. Problematización

Unos de los principales retos a los que nos enfrentamos como seres humanos es la contaminación ambiental. La contaminación del aire tiene efectos adversos para el planeta, tanto para la flora, fauna y la vida humana. El desarrollo y crecimiento de nuestras ciudades, al mismo tiempo que presenta oportunidades de progreso económico y de movilidad social, trae consigo también el deterioro de diferentes condiciones de vida. Entre los aspectos negativos está la contaminación del aire ambiental.

Es por esto que, en conjunto con nuestro socio formador la Dirección de Gestión Integral del Aire, se nos ha planteado el reto integrador para el bloque de Aplicación de Métodos Multivariados en Ciencia de Datos acerca de la contaminación del aire en el Área Metropolitana de Monterrey, en el que se analizarán bases de datos abiertas relacionadas con los contaminantes y condiciones meteorológicas que se registran diariamente en la estación Centro de monitoreo pertenecientes al Sistema de Monitoreo del Aire (SIMA) durante enero de 2017 a julio de 2021 para aplicar el uso de la Estadística Multivariada para obtener conocimiento de la naturaleza de contaminantes y sus interrelaciones, en específico en este reporte, el uso del análisis de componentes principales.

## II. Preguntas de investigación

Una vez que ha comenzado el proceso de análisis de la base de datos y se ha generado un mayor entendimiento del comportamiento de los datos es importante generar preguntas de investigación que nos permitan guiar la investigación y marcar un camino, tales como:

- ¿Cuál es la cantidad óptima de componentes principales a obtener?
- ¿Es posible reducir la matriz de datos original a menos de la mitad de variables y seguir representando más del 75 % de la varianza total original?
- ¿Los resultados obtenidos varían según la herramienta computacional utilizada para aplicar el análisis?

## III. Enfoque

El enfoque y la metodología empleada para llevar a cabo el análisis de componentes principales a la base de datos fue la siguiente:

1. Carga y lectura de los datos
2. Importar las librerías necesarias
3. Análisis descriptivo de los datos
4. Limpieza de los datos

5. Escalado de las variables
6. Seleccionar número óptimo de componentes principales
7. Inicializar y entrenar el modelo
8. Cálculo eigenvectores y eigenvalores
9. Calcular porcentaje de varianza explicada por cada componente
10. Graficar porcentaje de varianza explicada acumulada
11. Reducir la dimensionalidad de la matriz de datos

## IV. Información

Las técnicas estadísticas multivariadas son aquellas que analizan múltiples características, medidas en un mismo individuo. Constituyen una generalización de las técnicas univariadas y bivariadas, donde todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que no tenga sentido interpretar de forma aislada sus diferentes efectos. Las técnicas multivariadas son herramientas que permiten al investigador extraer abundante información de los datos disponibles. Las mismas son complejas y requieren para su utilización de un conocimiento profundo de sus fundamentos. [**2**]. Una de las técnicas multivariadas más importantes es el análisis de componentes principales.

El análisis de componentes principales consiste en expresar un conjunto de variables en un conjunto de combinaciones lineales de factores no correlacionados entre sí, estos factores dando cuenta una fracción cada vez más débil de la variabilidad de los datos. Este método permite representar los datos originales (individuos y variables) en un espacio de dimensión inferior del espacio original, mientras limite al máximo la pérdida de información. [**3**] El método de PCA permite por lo tanto condensar la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering, entre otros. Aun así no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes.[**1**]

## V. Razonamiento

Durante el primer paso para ejecutar el análisis se cargó la base de datos en dos software computacionales, distintos e independientes entre sí, tal y como lo son Python y Minitab, comprobando los resultados finales obtenidos haciendo uso de ambas herramientas, por lo que se importortó en cada uno la bases de datos a partir de un formato .csv . Algunas de las librerías elementales utilizadas para el análisis fueron pandas (manipulación de las estructuras de la base de datos), numpy (manipulación y cálculos numéricos) y matplotlib (visualización). Por otra parte, algunas de las librerías más completas y específicas para el análisis que utilizaremos provienen de Scikit-learn, tales y como lo son PCA, *makepipeline*, StandardScaler y scale para el preprocesamiento y modelado de los datos.

El siguiente paso fue realizar un análisis descriptivo a la base de datos, en la que se obtuvo un resumen estadístico, visualización, almacenamiento y organización computacional de las variables involucradas. Uno de los propósitos de dicho resumen, fue obtener y analizar medidas estadísticas de cada uno de los atributos y columnas de la base de datos, de todas aquellas que cumplieran con el requisito de tener datos de tipo numérico, tales como medidas de tendencia central (la media, la desviación estándar, varianza), medidas de dispersión (mínimo, máximo) y medidas de posición no-central (cuartiles), entre otros, tal y como se muestra en la Tabla 1 para tener un mejor entendimiento de los datos, en la que destacan la gran cantidad de valores perdidos por variable, cuya cantidad llega hasta cerca de diecinueve mil en algunas variables, y la gran diferencia en rango que existe entre los valores registrados por variable, teniendo teniendo valores máximo cercanos a 15 en algunas variables mientras

otras se encuentran superando los 700, tomando en cuenta que cada variable tiene su propia unidad de medición.

| Variable | N* | Mean | StDev | Variance | Minimum | Q1 | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|
| CO | 11085 | 2.0675 | 1.1098 | 1.2317 | 0.0800 | 1.2400 | 2.6500 | 14.6000 |
| NO | 10314 | 11.454 | 22.717 | 516.083 | 0.500 | 2.700 | 11.900 | 500.000 |
| NO2 | 9002 | 8.9927 | 6.9534 | 48.3499 | 0.0000 | 3.8000 | 12.5000 | 78.1000 |
| NOx | 8548 | 19.408 | 24.857 | 617.865 | 1.000 | 6.900 | 23.200 | 500.000 |
| O3 | 7437 | 28.361 | 20.078 | 403.143 | 1.000 | 12.000 | 40.000 | 148.000 |
| PM25 | 10505 | 23.947 | 15.705 | 246.642 | 2.170 | 12.600 | 31.255 | 225.370 |
| PM10 | 1672 | 52.205 | 34.522 | 1191.777 | 2.000 | 29.000 | 67.000 | 735.000 |
| PRS | 2071 | 711.48 | 3.53 | 12.47 | 689.40 | 709.10 | 713.50 | 726.00 |
| RAINF | 2907 | 0.000066 | 0.004588 | 0.000021 | 0.000000 | 0.000000 | 0.000000 | 0.700000 |
| RH | 1143 | 58.047 | 20.209 | 408.412 | 1.000 | 42.000 | 75.000 | 99.000 |
| SO2 | 18908 | 4.2945 | 2.5938 | 6.7277 | 0.5000 | 2.7000 | 5.2000 | 61.9000 |
| SR | 1046 | 0.18700 | 0.23234 | 0.05398 | 0.00000 | 0.00600 | 0.18400 | 1.09600 |
| TOUT | 1228 | 22.330 | 6.833 | 46.691 | -4.750 | 17.860 | 27.090 | 40.550 |
| WD | 1122 | 116.70 | 86.00 | 7396.19 | 1.00 | 59.00 | 141.00 | 360.00 |
| WS | 1203 | 6.339 | 3.852 | 14.837 | 0.100 | 3.600 | 8.600 | 122.400 |

**Tabla 1:** *Analísis descriptivo de los datos*

Una vez establecido cierto contexto sobre la base de datos y el objetivo de de reducir la dimensionalidad de los datos de contaminantes y mediciones de tipo meteorológicos para la estación de monitoreo Centro a partir del análisis de componentes principales, fue que se llevó a cabo el proceso de filtración y limpieza de los datos, en los que se identificaron valores atípicos, a través del uso de diagramas de caja, y se determinó si se mantenían o no en la base de datos esos valores especiales y aquellos fuera del rango habitual a partir de los criterios otorgados por el socio formador, en los que se nos otorgó un listado de posibles criterios que pudieron afectar las mediciones de los sensores, en los que clasificaron como datos válidos (a mantener) o dato inválido (a eliminar) según fuera el caso de cada registro como se muestra en la Tabla 3. Esto nos permitió tener una matriz de registros más limpia y manejable para nuestro análisis, por lo que favoreció mucho a la obtención de resultados finales, en donde manejamos 9646 datos para las variables CO, NO, NO, NOx, O3, PM25, PM10, PRS, RAINF, RH, SO2, SR, TOUT, WD y WS con sus respectivas unidades de medición mostradas en la Tabla 2 .

Es conocido que el proceso de componentes principales identifica aquellas direcciones en las que la varianza es mayor. Como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1 (distribución normal), aquellas

| | | |
|---|---|---|
| PM10 | Material Particulado menor a 10 micrométros | µg/m3 |
| PM2.5 | Material Particulado menor a 2.5 micrométros | µg/m3 |
| O3* | Ozono | ppb |
| SO2* | Dióxido de Azufre | ppb |
| NO2* | Dioxido de Nitrógeno | ppb |
| CO | Monóxido de Carbono | ppm |
| NO | Monóxido de Nitrógeno | ppb |
| NOx | Óxidos de Nitrógeno | ppb |
| TOUT | Temperatura | ºC |
| RH | Humedad Relativa | % |
| SR | Radiación Solar | kW/m2 |
| RAINF** | Precipitación | mm/Hr |
| PRS | Presión Atmosférica | mm Hg |
| WSR | Velocidad del Viento | Km/hr |
| WDR | Dirección del Viento | º |

**Tabla 2:** *Unidades de las variables de la base de datos*

| Flag | Description | Hora |
|---|---|---|
| P | Falla eléctrica | Valida |
| p | Falla eléctrica | Invalida |
| C | Calibración | Valida |
| c | Calibración | Invalida |
| D | Apagado | Valida |
| d | Apagado | Invalida |
| B | Malas condiciones | Valida |
| b | Malas condiciones | Invalida |
| m | Positivo sobre el rango | Invalida |
| l | Negativo sobre el rango | Invalida |
| z | Ceros y negativos | Invalida |
| o | PM10 mayor a 900 ug/m3 | Invalida |
| s | Valores repetidos | Invalida |
| r | comparativo PM10 vs PM2.5 | Invalida |
| e | Eliminar datos NO y Nox | Invalida |
| a | Eliminar PM menor a 5 ug/m3 y 0.05 ppm en CO | Invalida |
| s | Valores iguales consecutivos | Invalida |
| f | Valores 3 veces mayor que el valor anterior para PM10 | Invalida |
| h | Valores de temperatura con más de 10 grados o 10 mmHg de diferencia de una hora | Invalida |

**Tabla 3:** *Anotaciones sobre las banderas de los datos*

variables cuya escala sea mayor dominarán al resto. De ahí que se llevó a cabo el proceso de estandarización de los datos con la función de StandardScaler, la cual nos ayuda a poder convertir nuestros valores a unidades escalares de una unidad. La estandarización de los datos es importante para poder utilizar diferentes estimadores de aprendizaje automático.

Después, para poder elegir el número óptimo de componentes, dentro de Minitab realizamos una gráfica de Valores Propios (o también conocidos como valores característicos o raíces latentes) representando las varianzas de los componentes principales, tal y como se muestra en la Figura 1. Según el criterio de Kaiser, podemos utilizar los componentes con valores propios mayores a 1. En el caso de nuestros datos, observamos que podemos tomar en cuenta los 5 primeros componentes principales, ya que tienen valores propios mayores a 1. Sin embargo, tras analizar la varianza explicada acumulada de los primeros cinco, pudimos notar que la varianza era muy baja, dando un valor aproximado menor a 0.7, por lo que se decidió probar

con 7 componentes, donde el porcentaje de la varianza explicada subió a explicar el modelo en un 80 por ciento aproximadamente.

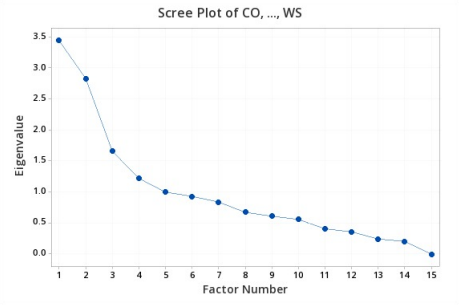


**Figura 1:** *Gráfica de sedimentación, valores propios*

Una vez definidos el número de componentes a utilizar, se inicializó el modelo haciendo uso de 'PCA' de Python y se entrenó el modelo a partir de la matriz de datos limpia obtenida anteriormente. A continuación se creó la matriz de vectores propios, a partir de los componentes generados por el modelo. El proceso de PCA genera siempre las mismas componentes principales independientemente del software utilizado, es decir, el valor de los loadings resultantes es el mismo. La única diferencia que puede darse es que el signo de todos los loadings esté invertido, tal y como fue el caso de lo obtenido según los resultados del modelo en Python y en Minitab, tal y como se muestra en las Tablas 4 y 5 respectivamente.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| CO | 0.288 | -0.015 | -0.383 | -0.133 | 0.007 | -0.183 | 0.053 |
| NO | 0.146 | 0.323 | 0.561 | -0.059 | -0.010 | -0.062 | -0.070 |
| NO2 | 0.344 | 0.263 | 0.018 | -0.016 | -0.014 | 0.035 | 0.159 |
| NOx | 0.251 | 0.369 | 0.478 | -0.049 | -0.015 | -0.029 | 0.008 |
| O3 | -0.372 | 0.234 | -0.156 | -0.146 | 0.003 | -0.151 | 0.079 |
| PM25 | 0.278 | 0.290 | -0.316 | 0.114 | -0.012 | 0.297 | 0.084 |
| PM10 | 0.210 | 0.366 | -0.345 | 0.002 | 0.000 | 0.118 | 0.114 |
| PRS | 0.173 | -0.220 | 0.037 | -0.681 | 0.041 | -0.267 | 0.060 |
| RAINF | -0.004 | 0.005 | 0.030 | 0.029 | 0.994 | 0.056 | 0.084 |
| RH | 0.247 | -0.334 | 0.164 | 0.082 | -0.044 | 0.455 | 0.124 |
| SO2 | 0.126 | 0.282 | -0.184 | -0.175 | 0.073 | -0.110 | -0.694 |
| SR | -0.171 | 0.286 | -0.030 | -0.229 | -0.050 | -0.035 | 0.630 |
| TOUT | -0.381 | 0.276 | 0.012 | 0.289 | -0.007 | 0.116 | -0.097 |
| WD | 0.163 | -0.035 | 0.001 | 0.503 | 0.018 | -0.726 | 0.158 |
| WS | -0.383 | 0.134 | 0.053 | -0.226 | 0.008 | 0.006 | -0.031 |

**Tabla 4:** *Vectores propios obtenidos en Python*

De igual manera, dentro de la Tabla 6 podemos encontrar representados los tamaños de los valores propios para poder determinar el

| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| CO | 0.288 | -0.015 | 0.383 | -0.133 | -0.007 | 0.183 | 0.053 |
| NO | 0.146 | 0.323 | -0.561 | -0.059 | 0.010 | 0.062 | -0.070 |
| NO2 | 0.344 | 0.263 | -0.018 | -0.016 | 0.014 | -0.035 | 0.159 |
| NOx | 0.251 | 0.369 | -0.478 | -0.049 | 0.015 | 0.029 | 0.008 |
| O3 | -0.372 | 0.234 | 0.156 | -0.146 | -0.003 | 0.151 | 0.079 |
| PM25 | 0.278 | 0.290 | 0.316 | 0.114 | 0.012 | -0.297 | 0.084 |
| PM10 | 0.210 | 0.366 | 0.345 | 0.002 | 0.000 | -0.118 | 0.114 |
| PRS | 0.173 | -0.220 | -0.037 | -0.681 | -0.041 | 0.267 | 0.060 |
| RAINF | -0.004 | 0.005 | -0.030 | 0.029 | -0.994 | -0.056 | 0.084 |
| RH | 0.247 | -0.334 | -0.164 | 0.082 | 0.044 | -0.455 | 0.124 |
| SO2 | 0.126 | 0.282 | 0.184 | -0.175 | -0.073 | 0.110 | -0.694 |
| SR | -0.171 | 0.286 | 0.030 | -0.229 | 0.050 | 0.035 | 0.630 |
| TOUT | -0.381 | 0.276 | -0.012 | 0.289 | 0.007 | -0.116 | -0.097 |
| WD | 0.163 | -0.035 | -0.001 | 0.503 | -0.018 | 0.726 | 0.158 |
| WS | -0.383 | 0.134 | -0.053 | -0.226 | -0.008 | -0.006 | -0.031 |

**Tabla 5:** *Vectores propios obtenidos en Minitab*

número de componentes principales que sería ideal elegir. Por norma, se debe conservar los componentes principales con los valores propios más grandes. De igual manera, la influencia de las variables en cada componente analizarse visualmente con un gráfico de tipo heatmap, tal y como se muestra en la Figura 2, en donde podemos observar qué variables son las que tienen un vector propio más alto según el componente que se quiera trabajar.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Eigenvalue | 3.4398 | 2.8312 | 1.66 | 1.2212 | 1.0008 | 0.9323 | 0.8408 | 0.6751 | 0.6127 | 0.5613 |
| Proportion | 0.229 | 0.189 | 0.111 | 0.081 | 0.067 | 0.062 | 0.056 | 0.045 | 0.041 | 0.037 |
| Cumulative | 0.229 | 0.418 | 0.529 | 0.61 | 0.677 | 0.739 | 0.795 | 0.84 | 0.881 | 0.918 |
| Eigenvalue | 0.41 | 0.3607 | 0.2453 | 0.2061 | 0.0026 | | | | | |
| Proportion | 0.027 | 0.024 | 0.016 | 0.014 | 0 | | | | | |
| Cumulative | 0.946 | 0.97 | 0.986 | 1 | 1 | | | | | |

**Tabla 6:** *Análisis de valores propios obtenidos en Minitab*

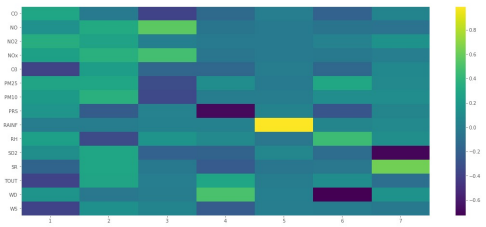

**Figura 2:** *Heatmap de los componentes*

El siguiente paso consistió en calcular y visualizar el porcentaje de varianza explicada y el porcentaje de varianza explicada acumulada por cada uno de los siete componentes principales que utilizamos, tal y como se muestra en las Figuras 3 y 4 respectivamente. Lo obtenido en dichas visualizaciones nos muestra que el primer componente principal es el que la primera componente explica el 23 % de la

varianza observada en los datos, la segunda el 19 %, y la tercera el 11 %. Las tres últimas componentes se aproximan al 7 % de varianza explicada cada uno, a partir de ahí ya existe cierta tendencia lineal a mantenerse en el mismo rango hasta que se explique por completo la varianza, por lo que se decidió, como se mencionó anteriormente, utilizar sólo los siete componentes que representan cerca del 80 % de la varianza total.

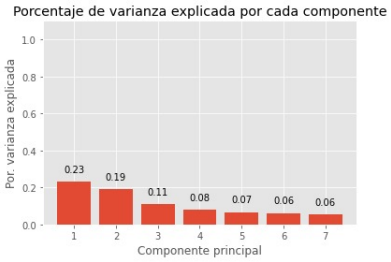


**Figura 3:** *Porcentaje de varianza explicada por cada componente*

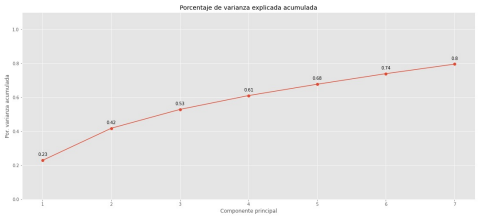


**Figura 4:** *Porcentaje de varianza explicada acumulada*

Finalmente, se reduce la dimensionalidad de la matriz de datos proyectándolas en el espacio definido por las componentes, a través de una transformación, la cual es el resultado de multiplicar los vectores que definen cada componente con el valor de las variables y da como resultado una matriz de datos nueva, tal y como se muestra en la Tabla 7. Pasando a analizar las nuevas variables creadas, las cuales, recordemos que son combinaciones lineales de las variables originales, es posible observar que son independientes entre sí, lo cual es justamente el objetivo. Esto se puede observar en la Figura 5 en donde se muestra el diagrama de dispersión entre el primer contra el segundo componente y se observa que no existe una relación entre las variables, pues se visualiza una nube de puntos con forma de elipse.

| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|
| 2.026 | -0.014 | 0.761 | 1.811 | -0.015 | 0.638 | 0.004 |
| 1.794 | -0.493 | 0.714 | 1.747 | -0.006 | 0.459 | -0.021 |
| 2.590 | 0.279 | 1.425 | 1.894 | 0.019 | -0.329 | 0.152 |
| -0.525 | 1.605 | 1.333 | -1.170 | -0.017 | -0.370 | 0.389 |
| 1.643 | 5.574 | 5.398 | -0.837 | 0.127 | -3.144 | 2.343 |
| -0.391 | 5.089 | 4.398 | -2.051 | 0.062 | -1.160 | 2.620 |
| 0.593 | 6.021 | 5.492 | -1.153 | 0.180 | -3.103 | 3.024 |
| -2.231 | 2.360 | 1.795 | -1.791 | 0.088 | -0.032 | 2.025 |
| -2.816 | 1.581 | 0.951 | -1.697 | 0.047 | 0.679 | 1.552 |
| -1.492 | 0.539 | 0.877 | -0.889 | -0.103 | 0.148 | -0.554 |
| -1.631 | -0.018 | 0.422 | -3.121 | -0.058 | 1.711 | 1.513 |

**Tabla 7:** *Primeras 10 proyecciones obtenidas*

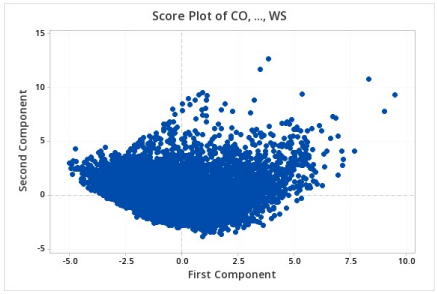


**Figura 5:** *Gráfica de puntuaciones del modelo PCA*

Finalmente, al realizar el análisis de componentes principales en Minitab, fue posible graficar las cargas de los primeros dos componentes para observar qué variables son las que predominan más en cada componente, y si lo hacen de manera positiva o negativa. Esto también ayuda a conocer qué variables se sobreponen a otras. Por ejemplo, en la Figura 6, se observa la gráfica de las cargas del primer componente contra el segundo. En ella podemos observar que el único contaminante criterio que pone un contrapeso a los demás, es el ozono, pues este va en el sentido negativo, mientras que todos los demás en dirección positiva. También, podemos observar que la mayoría de contaminantes tienen un 'brazo' de tamaño similar, aunque son el PM10 y NOx los que resaltan.

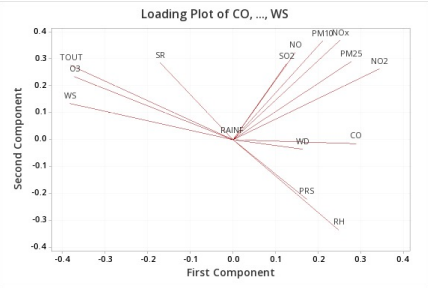


**Figura 6:** *Gráfica de cargas del modelo PCA*

## VI. Resultados

El proceso del análisis de Componentes Principales nos permitió reducir la dimensionalidad de los datos de contaminantes y mediciones de tipo meteorológicos para la estación de monitoreo Centro, de 14 variables originales (contaminantes) a 7 componentes, en los que se encuentran representado cerca del 80 % de la varianza total de la matriz de datos originales. Dichos componentes principales son representados de la siguiente manera, según la ponderación obtenida por los vectores propios de cada variable.

Dentro de estos valores, podemos observar que para cada componente las variables más influyentes son las siguientes:

- Componente Principal 1
  - TOUT: Temperatura, WS: Velocidad del viento, O3: Ozono, NO2: Dióxido de nitrógeno
- Componente Principal 2
  - NO: Monóxido de nitrógeno, NOx: Óxidos de Nitrógeno, PM10: Material particulado menor a 10 micrométros, RH: Humedad relativa
- Componente Principal 3
  - NO: Monóxido de nitrógeno, NOx: Óxidos de Nitrógeno
- Componente Principal 4
  - WD: Dirección del viento, PRS: Presión atmosférica
- Componente Principal 5
  - RAINF: Precipitación, PRS: Presión atmosférica
- Componente Principal 6
  - RH: Humedad relativa, WD: Dirección del viento
- Componente Principal 7
  - SO2: Dióxido de Azufre, SR: Radiación solar

## VII. Conclusiones

A través del reto integrador fue posible aplicar el uso de la Estadística Multivariada para obtener conocimiento de un proceso que presenta varias características, registradas por mediciones numéricas y categóricas interrelacionadas tal y como lo fue el análisis de componentes principales sobre la Contaminación del Aire en el Área Metropolitana de Monterrey, en el que se llevó a cabo un proceso aplicado de ciencia de datos, en el que se utilizó una base de datos, para emplear técnicas de limpieza y tratamiento, estandarización, modelación y proyección final de los datos, a partir del objetivo principal de reducir la dimensionalidad de los datos. Los resultados fueron sumamente satisfactorios, no sólo por el proceso educativo y académico previo a la implementación final, sino también porque se consiguió el prometido con éxito representado cerca del 80 % de la varianza total de la matriz de datos originales.

## Referencias

[1] Rodrigo, J. (2021). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. Cienciadedatos.net.

[2] Sagaró del Campo (2021). Técnicas estadísticas multivariadas para el estudio de la causalidad en Medicina. Scielo.sld.cu.

[3] XLSTAT, Análisis de Componentes Principales (ACP). (2020). XLSTAT, Your data analysis solution.