

Subíndice de riesgo de seguridad pública

Martha Aguilar Jiménez

04/10/2024

Se instalan las librerías que se necesitarán para realizar análisis de componentes principales y análisis de factores.

```
# Se limpia la memoria
rm(list = ls())
invisible(gc(reset = TRUE))
library(readr) #Para leer archivos
library(corrplot) #Gráficar las correlaciones
library(foreign) # Leer archivos dbf
library(readxl) # Leer archivos xlsx
library(stats) # Cálculos estadísticos
library(psych) # AF por componentes principales
library(ggplot2) # Gráficos basados en grammar of graphics
library(GGally) # Para realizar las gráficas ggpairs
library(kableExtra) # Formato de tablas en R Markdown
library(dplyr) # Manipulación de datos
library(FactoMineR) #Principales métodos de componentes principales
library(factoextra) #Proporciona las funciones para visualizar los resultados del acp

# Se cargan las librerías necesarias para realizar el mapa.
library(sf) # Para manejar datos espaciales
library(tidyr) # Para transformación de datos
#(dplyr) # Para manipulación de datos
#library(ggplot2) # Para la visualización
```

Se leen los datos

```
# Se establece el directorio de trabajo
setwd("C:/Users/User/Documents/ÍNDICE DE RIESGO/Bases de datos")
## Se cargan los datos de interés
datos<-read_excel("BASES DE DATOS.xls",sheet ="SEG_2023")
datos<-as.data.frame(datos)
datos$cve<-c(01,02,03,04,05,06,07,08,09,10,11,12,13,14,15,
           16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32)
rownames(datos)<- datos[,"cve"]
seg <- datos[, -c(1,11,20)]# Es la base con todas las variables de interés.
```

1.1 Análisis Exploratorio

Se observan las características del data frame *seg*.

```
str(seg) #32 observaciones y 17 variables (falta el indicador 10)
```

```
## 'data.frame': 32 obs. of 17 variables:
## $ seg_1_2023 : num 6.8 84.81 8.58 9.26 4.97 ...
## $ seg_2_2023 : num 29584 31690 21756 18610 24418 ...
## $ seg_3_2023 : num 27295 27208 19003 16891 20075 ...
## $ seg_4_2023 : num 89.5 91.2 90.7 90.8 93.2 ...
## $ seg_5_2023 : num 2827 3076 2599 2480 1992 ...
## $ seg_6_2023 : num 93.9 82.6 94.6 93.8 89.2 ...
## $ seg_7_2023 : num 11.5 79.5 30.6 36 6.1 84.1 10.6 67.6 11.2 9.5 ...
## $ seg_8_2023 : num 5894 7642 12053 7166 4437 ...
## $ seg_9_2023 : num 39738 95631 37118 33007 37049 ...
## $ seg_11_2023: num 54.2 76.1 34.9 54.1 50.4 ...
## $ seg_12_2023: num 27.1 24.5 17.6 22.5 22.9 ...
## $ seg_13_2023: num 191 189 249 203 141 ...
## $ seg_14_2023: num 38 51.4 37.5 35.6 39.3 ...
## $ seg_15_2023: num 34.7 33.9 34.1 33.1 32.2 ...
## $ seg_16_2023: num 33.4 36.6 35 37.4 37.1 ...
## $ seg_17_2023: num 2.21 1.26 4.57 5.24 1.79 ...
## $ seg_18_2023: num 97.1 79.2 97.1 94.8 88.7 94.4 79.5 96.9 98.4 91.5 ...
```

```
class(seg) # Todas las variables son numéricas y son data frame.
```

```
## [1] "data.frame"
```

Como se observa se tiene un data frame con 17 variables numéricas y con 32 observaciones.

Se continua con el análisis exploratorio de los datos.

```
summary(seg) #Solo la variable país es de tipo carácter, las demás son numéricas
```

```
##    seg_1_2023      seg_2_2023      seg_3_2023      seg_4_2023
## Min. : 2.338  Min. :16386  Min. :14082  Min. :89.54
## 1st Qu.: 9.888  1st Qu.:21029  1st Qu.:17304  1st Qu.:91.95
## Median :18.280  Median :27641  Median :21179  Median :92.96
## Mean   :30.265  Mean   :27333  Mean   :21650  Mean   :93.09
## 3rd Qu.:43.878  3rd Qu.:31722  3rd Qu.:25329  3rd Qu.:94.35
## Max.  :108.417  Max. :45501   Max. :38253  Max. :96.70
##    seg_5_2023      seg_6_2023      seg_7_2023      seg_8_2023      seg_9_2023
## Min. :1288   Min. :75.79   Min. : 3.30   Min. :4437   Min. :22542
## 1st Qu.:2099   1st Qu.:90.53   1st Qu.:11.43   1st Qu.:5737   1st Qu.:36782
## Median :2478   Median :92.83   Median :36.15   Median :7009   Median :43919
## Mean   :2534   Mean   :91.80   Mean   :41.60   Mean   :7119   Mean   :49302
## 3rd Qu.:3013   3rd Qu.:94.51   3rd Qu.:66.00   3rd Qu.:8039   3rd Qu.:58461
## Max.  :3687   Max. :97.27   Max. :85.50   Max. :12053  Max. :95631
##    seg_11_2023     seg_12_2023     seg_13_2023     seg_14_2023
## Min. :30.59   Min. :13.75   Min. :128.1   Min. :27.57
## 1st Qu.:56.99  1st Qu.:24.35  1st Qu.:166.8  1st Qu.:37.15
## Median :74.40  Median :29.20  Median :189.8  Median :41.74
## Mean   :70.17  Mean   :29.26  Mean   :231.9  Mean   :42.02
## 3rd Qu.:80.94  3rd Qu.:34.18  3rd Qu.:236.4  3rd Qu.:47.86
## Max.  :90.85  Max. :46.58  Max. :976.5  Max. :57.08
```

```

##   seg_15_2023    seg_16_2023    seg_17_2023    seg_18_2023
## Min.   :25.71   Min.   :20.23   Min.   :0.5074   Min.   :69.50
## 1st Qu.:33.59  1st Qu.:34.40  1st Qu.:1.7264  1st Qu.:88.35
## Median :36.27  Median :36.76  Median :2.4076  Median :92.05
## Mean    :36.78  Mean    :38.08  Mean    :2.5965  Mean    :90.43
## 3rd Qu.:39.38  3rd Qu.:41.81  3rd Qu.:3.3193  3rd Qu.:94.50
## Max.    :59.63  Max.    :63.95  Max.    :5.2419  Max.    :98.40

```

Se observa que las 17 variables analizadas tienen diferentes valores muy dispersos, además, se observa que en ningún caso hay valores faltantes.

1. La variable “*seg_1_2023: Tasa de homicidios por cada cien mil habitantes*” en 2021 tuvo un valor mínimo de 2.338, un valor máximo de 108.417 y una media de 30.265, es decir, en promedio hubieron 30 homicidios por cada cien mil hab.
2. La variable “*seg_2_2023: Incidencia delictiva por cada cien mil habitantes*” durante 2021 registró un valor mínimo de 16386, un valor máximo de 45501 y en promedio hubieron 27333 delitos por cada cien mil hab.
3. La variable “*seg_3_2023: Prevalencia delictiva por cada cien mil habitantes*” durante 2021 registró un valor mínimo de 14082, un valor máximo de 38253 y en promedio hubieron 21650 víctimas por cada cien mil hab.
4. La variable “*seg_4_2023: Cifra negra de delitos en la entidad federativa*” durante 2021 registró un valor mínimo de 89.54%, un valor máximo de 96.70% y en promedio se registró una cifra negra de 93.09%, es decir, es el porcentaje de delitos no captados por la autoridad debido a que la víctima no denunció el delito o lo denunció pero no se inició una averiguación previa.
5. La variable “*seg_5_2023: Prevalencia delictiva por cada diez mil unidades económicas*” durante 2021 registró un valor mínimo de 1288, un valor máximo de 3687 y en promedio hubieron 2534 unidades económicas que fueron víctimas de algún delito por cada diez mil unidades económicas.
6. La variable “*seg_6_2023: Cifra negra de delitos en unidades económicas*” durante 2021 registró un valor mínimo de 75.79%, un valor máximo de 97.27% y en promedio se registró una cifra negra de 91.80%, es decir, es el porcentaje de delitos no captados por la autoridad debido a que la unidad económica víctima no denunció el delito o lo denunció pero no se inició una averiguación previa a nivel nacional.
7. La variable “*seg_7_2023: Índice de Riesgo por Crimen Organizado*” durante el segundo trimestre registró un valor mínimo de 3.30, un valor máximo de 85.50 y en promedio el índice se registró 41.60.
8. La variable “*seg_8_2023: Promedio de costos del delito contra personas, por entidad federativa*” durante 2021 registró un valor mínimo de 4437, un valor máximo de 12053 y en promedio los costos del delito contra personas fue de 7119.
9. La variable “*seg_9_2023: Promedio de costos del delito en unidades económicas, por entidad federativa*” durante 2021 registró un valor mínimo de 22542, un valor máximo de 95631 y en promedio los costos del delito en unidades económicas fue de 49302.
10. La variable “*seg_10_2023: Porcentaje del gasto público destinado a seguridad pública en la entidad federativa*” no se calculó.
11. La variable “*seg_11_2023: Porcentaje de personas que consideran insegura su entidad federativa*” durante 2022 registró un valor mínimo de 30.59%, un valor máximo de 90.85% y en promedio el porcentaje fue de 70.17%.
12. La variable “*seg_12_2023: Porcentaje de personas que cambió algún hábito por temor a sufrir algún delito*” durante 2022 registró un valor mínimo de 13.75%, un valor máximo de 46.58% y en promedio el porcentaje fue de 29.26%.
13. La variable “*seg_13_2023: Tasa de policías por cada cien mil habitantes*” durante 2021 registró un valor mínimo de 128.1, un valor máximo de 976.5 en promedio hubieron 231.9 policías estatales y municipales por cada cien mil hab.
14. La variable “*seg_14_2023: Desconfianza en la policía estatal y municipal*” durante 2022 registró un valor mínimo de 27.57, un valor máximo de 57.08 y en promedio hubo una desconfianza en la policía estatal y municipal de 42.02.

15. La variable “*seg_15_2023: Desconfianza en jueces*” durante 2022 registró un valor mínimo de 25.71, un valor máximo de 59.63 y en promedio hubo una desconfianza en los jueces de 36.78.
16. La variable “*seg_16_2023: Desconfianza en ministerios públicos*” durante 2022 registró un valor mínimo de 20.23, un valor máximo de 63.95 y en promedio hubo una desconfianza en ministerios públicos de 38.08.
17. La variable “*seg_17_2023: Variación porcentual del índice de conflictividad en la entidad federativa*” durante el trimestre de 2023 registró un valor mínimo de 0.5074, un valor máximo de 5.2419 y en promedio la variación fue de 2.5965.
18. La variable “*seg_18_2023: Índice de Impunidad estatal en materia penal en entidades federativas*” durante el 2021 registró un valor mínimo de 69.50, un valor máximo de 98.40 y en promedio el índice de impunidad estatal fue de 90.43.

1.2 Visualización de los datos

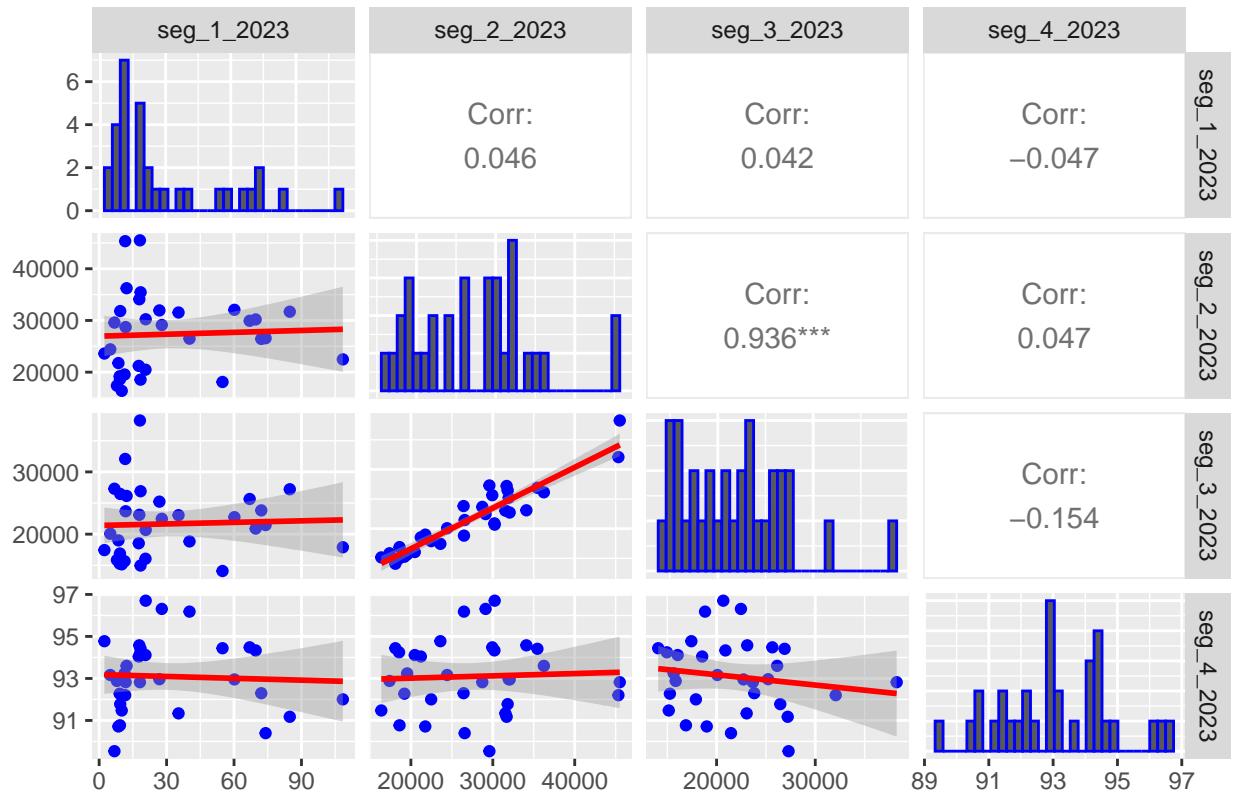
Podemos, además apoyarnos de una gráfica, en donde nos mostrará 3 elementos importantes: 1. Los gráficos de dispersión en los cuales podremos ver si existen relaciones lineales entre las variables. 2. Mostrará también las densidades de las variables. 3. Y finalmente, tendremos las correlaciones de las variables.

```
# Función para gráficos de dispersión
lowerFn <- function(data, mapping, method = "lm", ...) {
  p <- ggplot(data = data, mapping = mapping) +
    geom_point(colour = "blue") +
    geom_smooth(method = method, color = "red", ...)
  p
}

# Matriz de gráficos de dispersión parte 1
ggpairs(seg[, 1:4], lower = list(continuous = wrap(lowerFn, method = "lm")),
         diag = list(continuous = wrap("barDiag", colour = "blue")),
         upper = list(continuous = wrap("cor", size = 4)),
         title = "Matriz de gráficos de dispersión - parte 1")

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

Matriz de gráficos de dispersión – parte 1

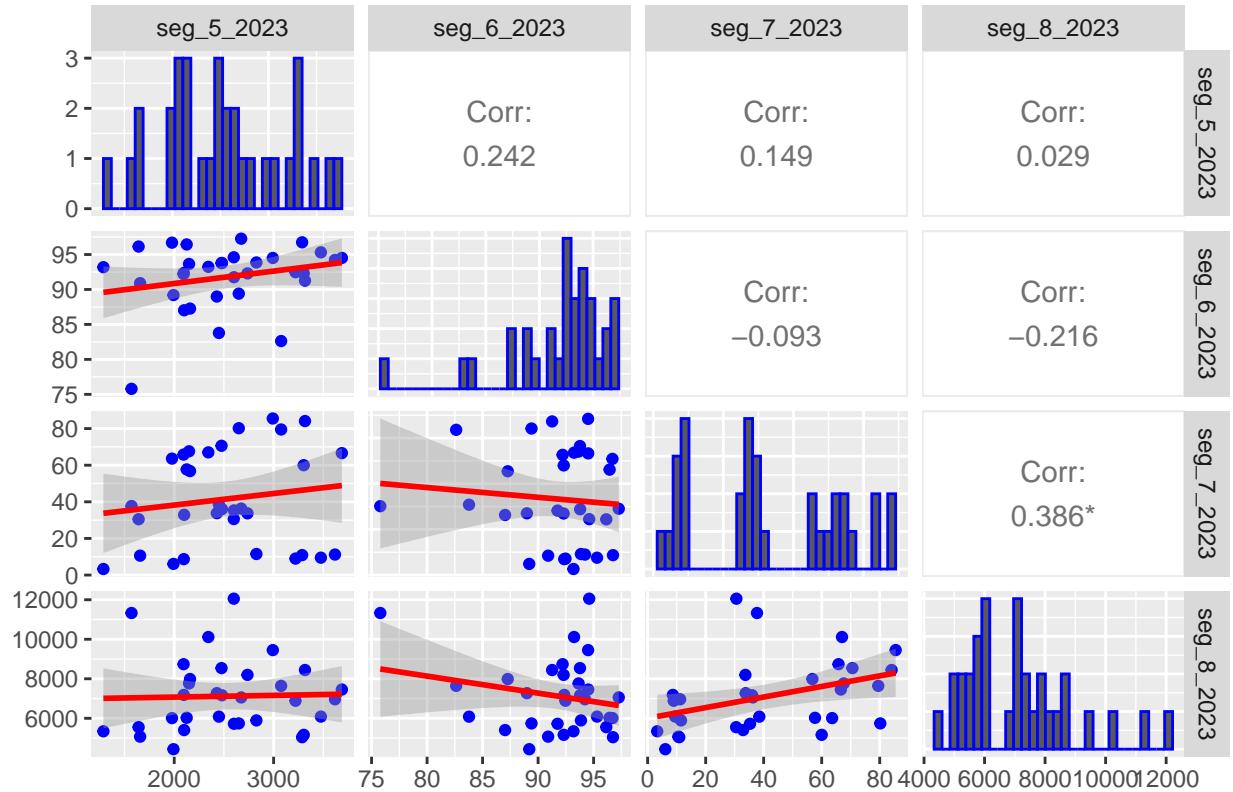


En la matriz de gráficos anterior se aprecia que, la dispersión es variada entre variables, sin embargo, la que completamente sigue la tendencia lineal es la del par de variables seg_2_2023 y seg_3_2023 la cual tiene la mayor correlación lineal de 0.936, en contraste con el par de variables seg_1_2023 y seg_3_2023, cuyo coeficiente de correlación es el menor con un valor de 0.042.

```
# Matriz de gráficos de dispersión parte 2
ggpairs(seg[, 5:8], lower = list(continuous = wrap(lowerFn, method = "lm")),
         diag = list(continuous = wrap("barDiag", colour = "blue")),
         upper = list(continuous = wrap("cor", size = 4)),
         title = "Matriz de gráficos de dispersión - parte 2")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

Matriz de gráficos de dispersión – parte 2

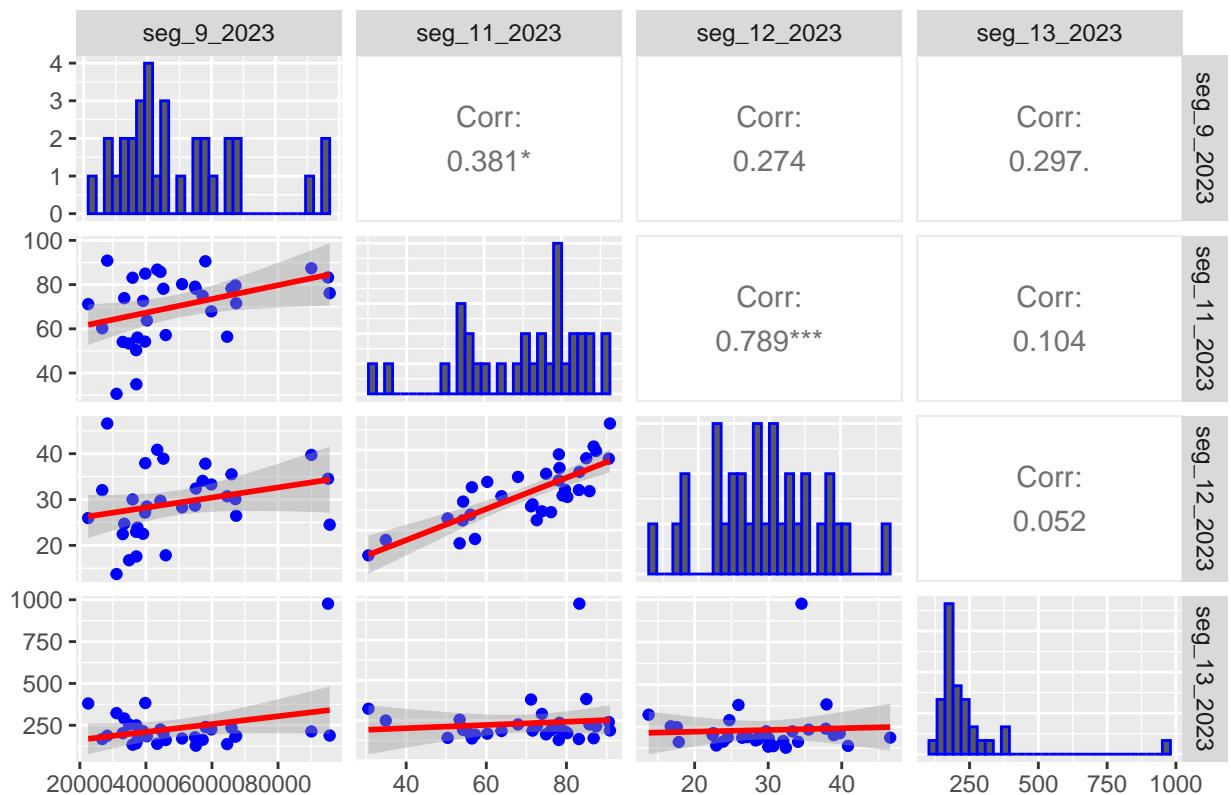


De acuerdo con la matriz de gráficos previa, se observa que la mayoría de las gráficas muestran gran dispersión entre variables, con estas variables, la que mayor correlación lineal tiene es de 0.386, que corresponde al par de variables seg_7_2023 y seg_8_2023 mientras que, la de menor correlación corresponde al par de variables seg_5_2023 y seg_8_2023 con un valor de 0.029. Con respecto a la densidad la variable seg_7_2023 es la que menos se parece a la normal.

```
# Matriz de gráficos de dispersión parte 3
ggpairs(seg[, 9:12], lower = list(continuous = wrap(lowerFn, method = "lm")),
        diag = list(continuous = wrap("barDiag", colour = "blue")),
        upper = list(continuous = wrap("cor", size = 4)),
        title = "Matriz de gráficos de dispersión - parte 3")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

Matriz de gráficos de dispersión – parte 3

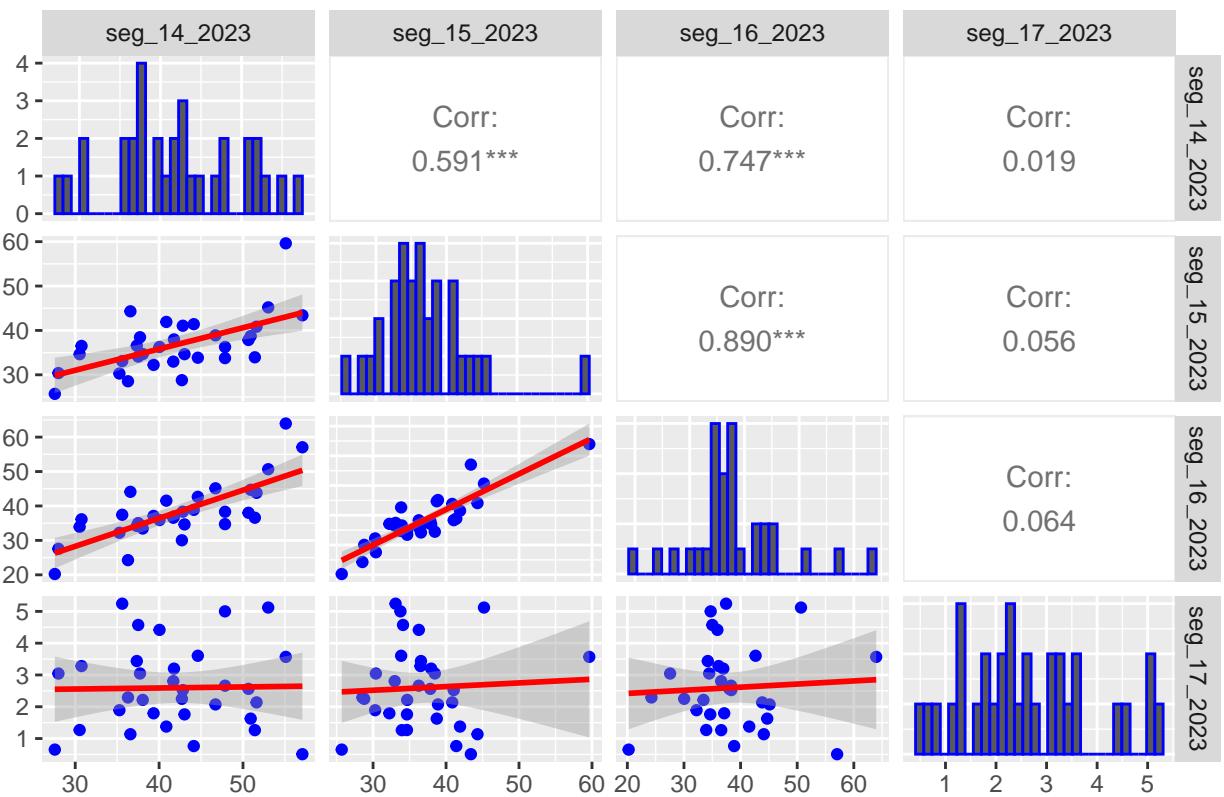


De acuerdo con la matriz de gráficos previa, la que completamente sigue la tendencia lineal es la del par de variables seg_11_2023 y seg_12_2023, con una correlación lineal de 0.789, mientras que, la de menor correlación corresponde al par de variables seg_12_2023 y seg_13_2023 con un valor de 0.052. Con respecto a las densidades, se aprecian que siguen una tendencia normal con diferentes características.

```
# Matriz de gráficos de dispersión parte 4
ggpairs(seg[, 13:16], lower = list(continuous = wrap(lowerFn, method = "lm")),
        diag = list(continuous = wrap("barDiag", colour = "blue")),
        upper = list(continuous = wrap("cor", size = 4)),
        title = "Matriz de gráficos de dispersión - parte 4")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

Matriz de gráficos de dispersión – parte 4



De acuerdo con la matriz de gráficos previa, se observa que varios pares siguen una tendencia lineal, el par de variables con mayor tendencia lineal es seg_15_2023 y seg_16_2023, con una correlación lineal de 0.89, mientras que, la de menor correlación corresponde al par de variables seg_14_2023 y seg_17_2023 con un valor de 0.019. Con respecto a las densidades de cada variable, se observa un parecido con la distribución gaussiana.

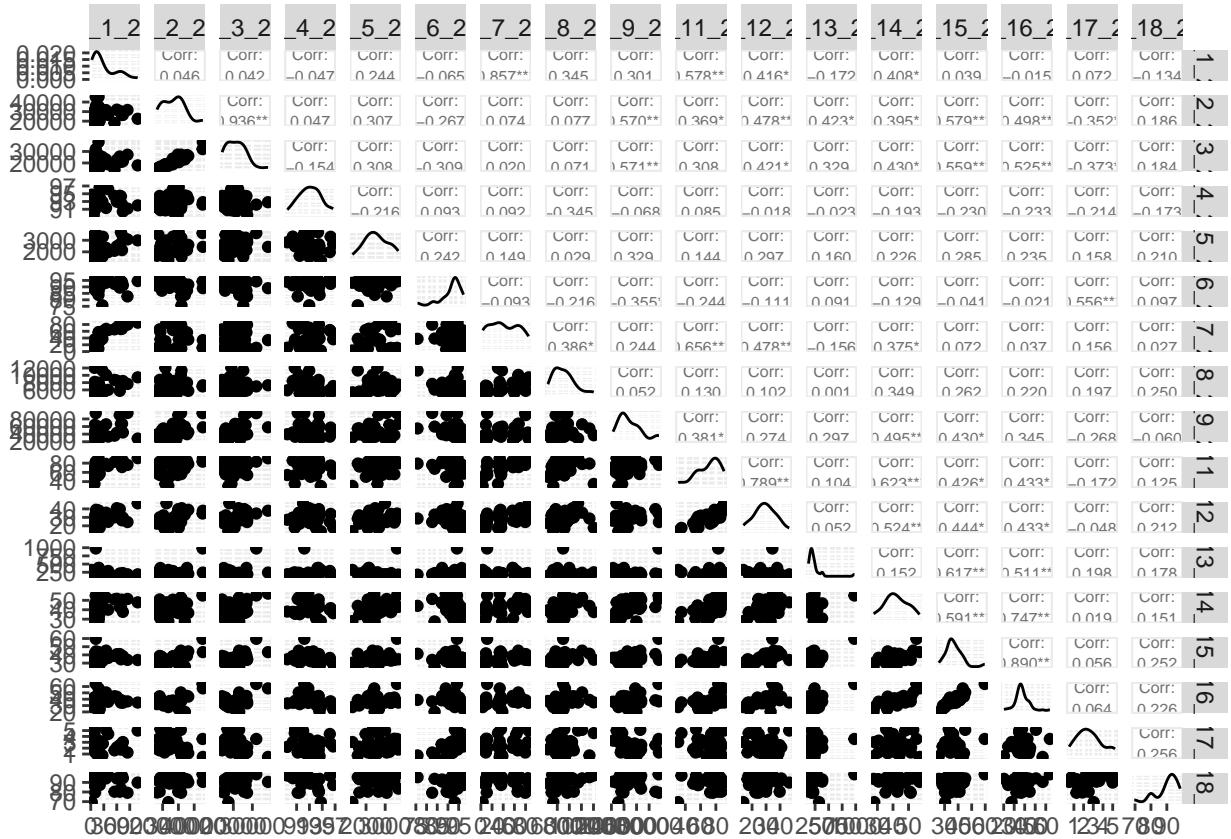
```
# Matriz de gráficos de dispersión parte 5
ggpairs(seg[,17:18], lower = list(continuous = wrap(lowerFn, method = "lm")),
        diag = list(continuous = wrap("barDiag", colour = "blue")),
        upper = list(continuous = wrap("cor", size = 4)),
        title = "Matriz de gráficos de dispersión - parte 5")
```

```
## Error in '[.data.frame'(seg, , 17:18): undefined columns selected
```

#La siguiente salida marcó error, se soluciona en cuanto se tenga información de la variable faltante.

```
#Este error se soluciona en cuanto se tenga información de la variable "seg_10_2023".
```

```
ggpairs(seg,diag=list(continuous="densityDiag"),upper = list(continuous = wrap("cor", size = 2)))
```



Con respecto a las densidades se espera que tengan una distribución gaussiana, vemos en la gráfica que 15 variables tienen un parecido a una campana de gauss, sin embargo, las variables *seg_7_2023* y *seg_13_2023* no tienen esa aproximación, sin embargo, se van a seguir considerando. Con respecto a las relaciones lineales entre variables, en la gráfica se puede apreciar que si hay variables relacionadas y finalmente, con respecto a la correlación entre variables, también se observan algunas correlaciones interesantes.

1.3 Estandarización de los datos

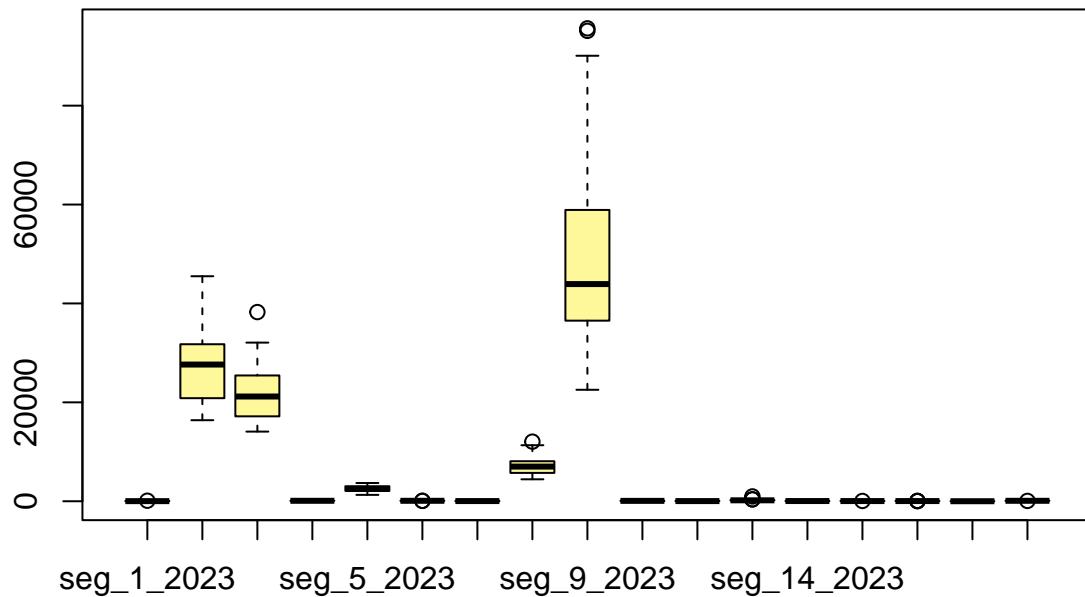
Antes de continuar es importante mencionar que cuando se estandarizan los datos la matriz de covarianzas y de correlaciones son prácticamente las mismas. La estandarización evita el problema de tener una variable con gran varianza.

Se realiza la gráfica de boxplot con los datos estandarizados.

```
#Se estandarizan los datos los datos
est_seg<- scale(seg, center = TRUE, scale = TRUE)

#Se realizan los diagramas de cajas
boxplot(seg, col ="#FFF89A", main= "Datos sin estandarizar")
```

Datos sin estandarizar

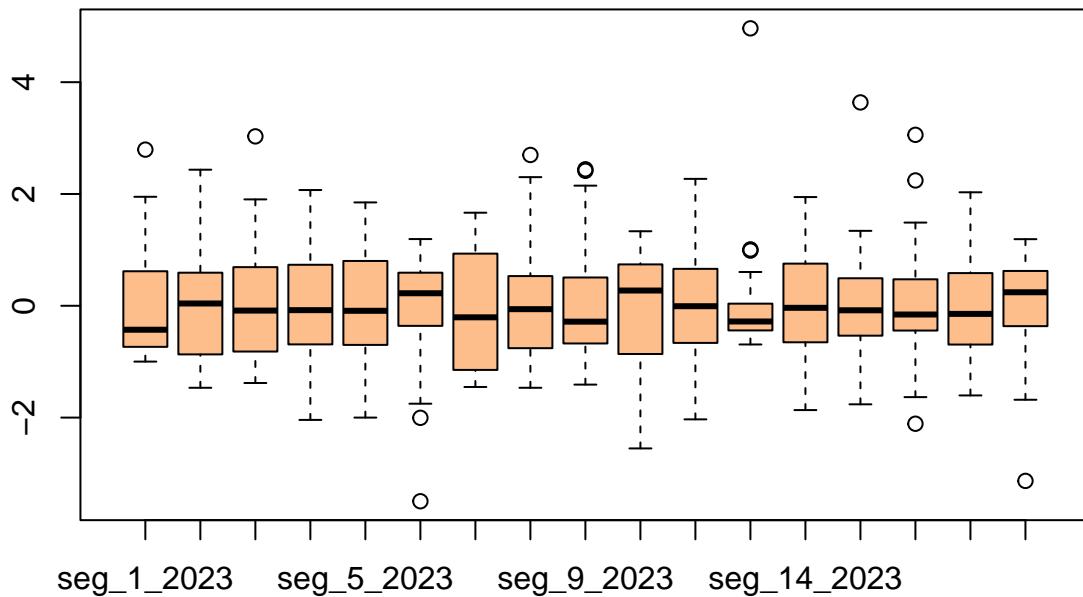


De acuerdo con el gráfico de caja y bigotes, la variable seg_9_2023 tiene valores muy altos lo que implica que las demás variables casi no se visualicen, además, se observan que algunas variables registran algunos valores atípicos aunque son pocos. Entonces, por lo anterior las escalas estén muy desfasadas, por lo tanto se recomienda estandarizar los datos para que sean comparables.

Se realiza el gráfico estandarizando.

```
boxplot(est_seg, col ="#FEE8E8", main="Datos estandarizados")
```

Datos estandarizados



Como se aprecia en el gráfico anterior, ya las variables resultan más parejas, pues ya todas oscilan alrededor del cero, aunque se observa que hay 9 variables que reportan uno o dos valores atípicos.

2. Análisis de componentes principales.

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá, pues, que estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

En el caso del aprendizaje no supervisado, la variable respuesta no se tiene en cuenta ya que el objetivo no es predecir, sino extraer información empleando los predictores, por ejemplo, para identificar sub-grupos. El principal problema al que se enfrentan este tipo de métodos es la dificultad para validar los resultados, dado que no se dispone de una variable respuesta que permita contrastarlos.

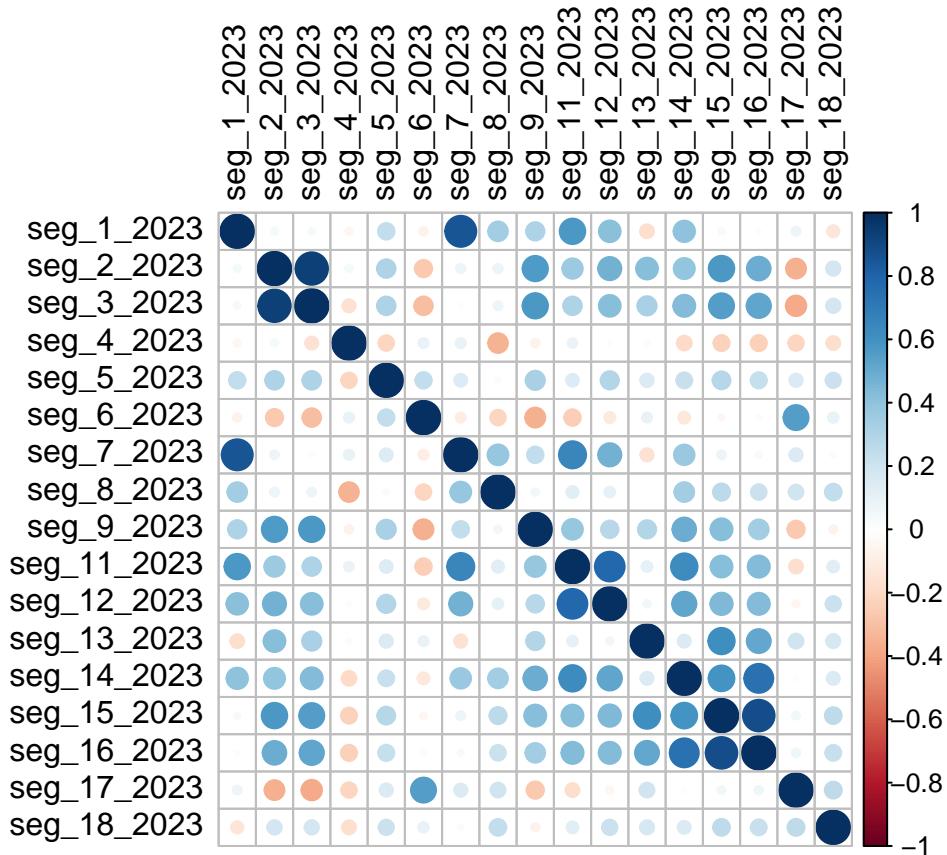
El método de PCA permite, por lo tanto, “condensar” la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering, etc. Aun así no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes.

2.1 Técnica de Análisis de Componentes Principales.

Se aplicará un análisis de componentes principales para reducir la dimensionalidad de los datos, esto para una mejor interpretación de qué es lo que explica al fenómeno de interés, es decir, se busca una relación entre las variables estudiadas de seguridad pública.

Para que un PCA tenga sentido, es necesario que exista una alta correlación entre los predictores, por lo que se explorará a continuación.

```
#seleccionamos los datos numéricos
cor_seg <- cor(est_seg)
#Gráfica de correlaciones entre variables
corrplot(cor_seg, tl.cex = 1, tl.col = "black", mar = c(0, 0, 0, 0))
```



Derivado de la gráfica previa, se observa que en efecto sí existen correlaciones altas entre algunas variables. La correlación mide la fuerza y dirección de la interacción entre dos variables, suponiendo relación lineal. Por ejemplo, la variable `seg_1_2023` tiene correlación positiva fuerte con la variable `seg_7_2023`, le sigue la `seg_11_2023`, `seg_12_2023` y `seg_14_2023`. En general algunas variables tiene correlaciones fuertes positivas, pero también se observan ciertas correlaciones negativas entre algunas variables. Finalmente, se calcula el determinante, se busca que el determinante se cercano a cero, pues implicaría un alta multicolinalidad entre variables.

```
det_seg<-det(cor_seg) #Se calcula el determinante, como es cercano a cero, quiere decir que hay correlacion
det_seg #2.013597e-07
```

```
## [1] 2.013597e-07
```

Como el determinante es 2.013597e-07, cercano a cero, quiere decir que hay correlación. Derivado de lo anterior, el análisis de componentes principales es factible.

2.3 Varianza explicada.

```
### PCA

seg.pc <- princomp(est_seg)
summary(seg.pc)

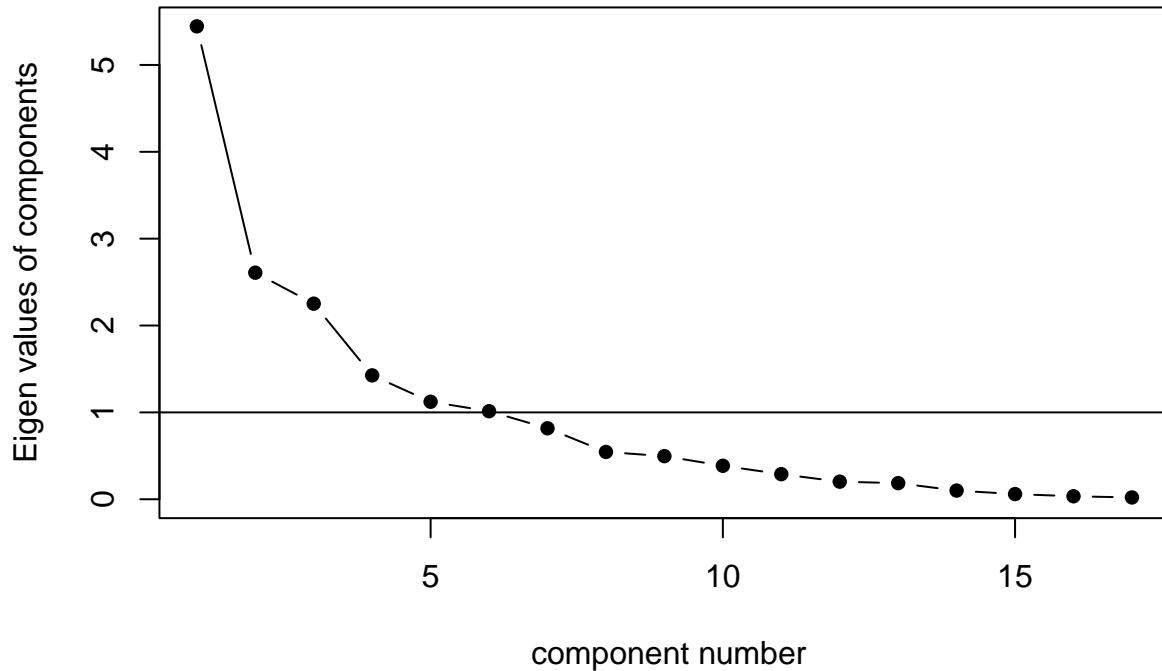
## Importance of components:
##           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation 2.2965971 1.5896394 1.4768955 1.17538938 1.04275931
## Proportion of Variance 0.3202646 0.1534393 0.1324460 0.08388859 0.06602486
## Cumulative Proportion 0.3202646 0.4737039 0.6061499 0.69003853 0.75606339
##           Comp.6    Comp.7    Comp.8    Comp.9    Comp.10
## Standard deviation 0.99076319 0.88952722 0.72634839 0.6933762 0.61098218
## Proportion of Variance 0.05960451 0.04804607 0.03203534 0.0291929 0.02266713
## Cumulative Proportion 0.81566790 0.86371397 0.89574931 0.9249422 0.94760933
##           Comp.11   Comp.12   Comp.13   Comp.14   Comp.15
## Standard deviation 0.52900933 0.4428392 0.42373495 0.310662630 0.239710500
## Proportion of Variance 0.01699284 0.0119078 0.01090255 0.005860267 0.003489101
## Cumulative Proportion 0.96460217 0.9765100 0.98741252 0.993272788 0.996761888
##           Comp.16   Comp.17
## Standard deviation 0.182098541 0.142013284
## Proportion of Variance 0.002013503 0.001224609
## Cumulative Proportion 0.998775391 1.000000000
```

Uno de los criterios para escoger el número adecuado de componentes principales, es cuidar que la varianza acumulada que expliquen, sea de por lo menos 70%, y de acuerdo con la salida del modelo, se necesitan cinco componentes, en el ejemplo se acumula el 75.8% de la varianza del modelo. Aunque también se podría proponer 4 componentes, ya que se acumula el 69% de la varianza, que es muy cercano a 70%.

Otros criterios son el del codo, que a partir del gráfico de sedimentación, se tiene que ubicar el cambio de trayectoria más abrupta, y en ese punto, será el número de componentes adecuado, otro es el de Kaiser, que indica conservar aquellas componentes cuyos valores propios superen a la unidad.

```
scree(est_seg,main = "Gráfico de sedimentación", pc=TRUE, fa=FALSE)
```

Gráfico de sedimentación

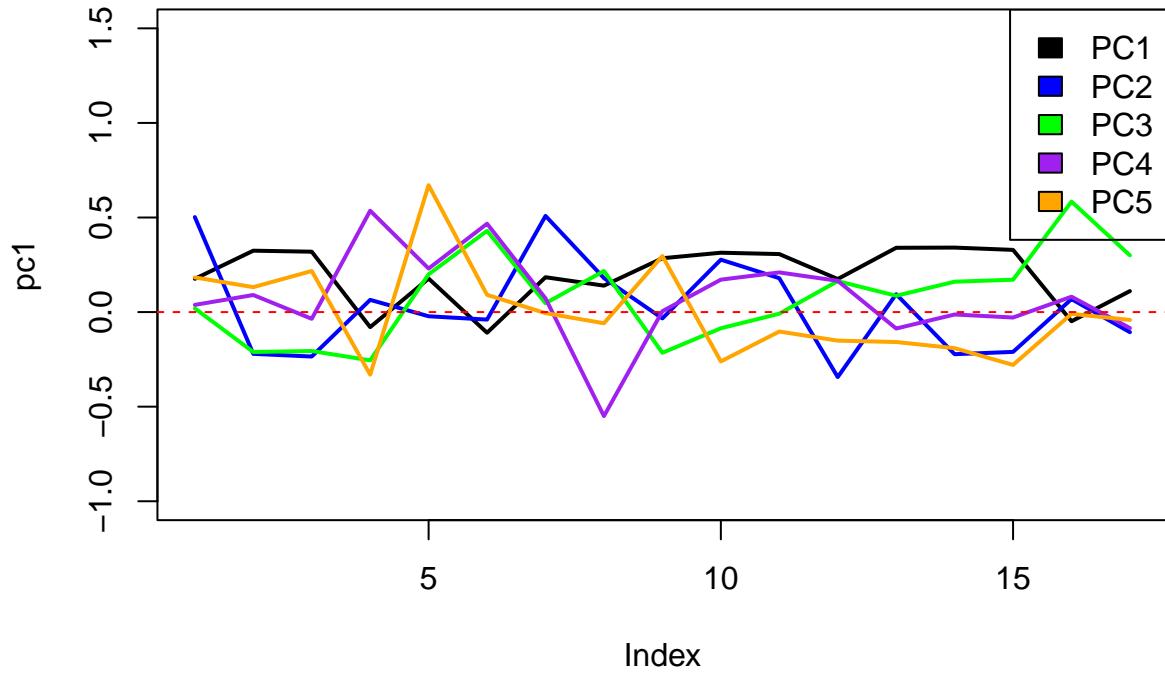


Según el gráfico anterior, el criterio del codo sugiere dos componentes, mientras que la regla de Kaiser sugiere cinco. En este caso, los criterios probados indican un número diferente de componentes, sin embargo, se va a priorizar la eficiencia del modelo, es decir, se van a elegir cinco componentes en favor de explicar lo más que se pueda de la varianza acumulada, esto pese a la idea de la técnica, que es reducir la dimensión.

2.4 Extracción de cargas por componentes principales

```
# Elegimos 5 comp
#Se cachan los vectores propios
pc1 <- seg.pc$loadings[,1]
pc2 <- seg.pc$loadings[,2]
pc3 <- seg.pc$loadings[,3]
pc4 <- seg.pc$loadings[,4]
pc5 <- seg.pc$loadings[,5]

#Se realizan las gráficas
plot(pc1, ylim = c(-1, 1.5), type = "l", lwd=2)
lines(pc2, col="blue", lwd=2)
lines(pc3, col="green", lwd=2)
lines(pc4, col="purple", lwd=2)
lines(pc5, col="orange", lwd=2)
abline(h=0, lty=2, col="red")
legend(x="topright", legend = c("PC1", "PC2", "PC3", "PC4", "PC5"),
       fill = c("black", "blue", "green", "purple", "orange"))
```



En la gráfica se muestra el comportamiento de los primeros 5 componentes principales respecto y la relación con las variables originales.

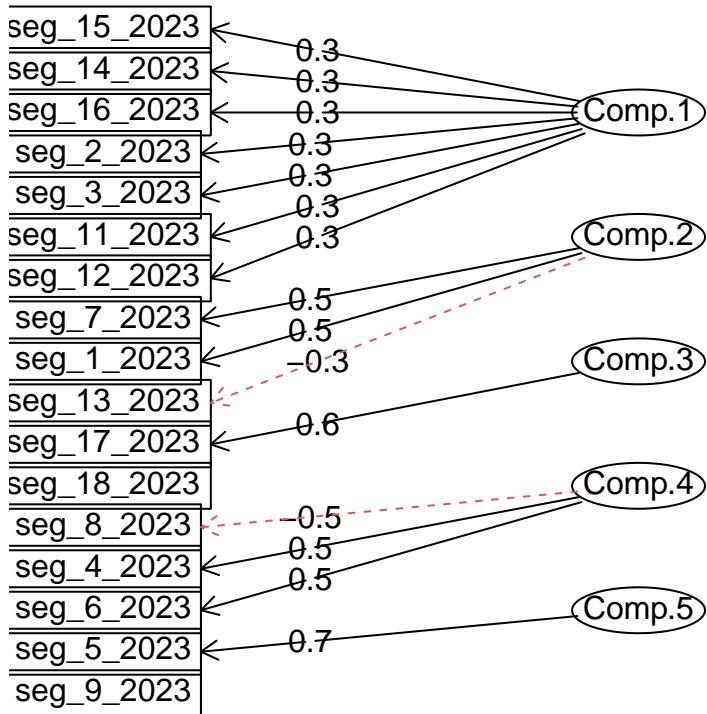
Según las cargas de los componentes, por lo menos visualmente hay algunas que destacan en alguna variable, por ejemplo las cargas del cuarto componente (morado), tiene un pico muy marcado hacia abajo en la variable *seg_8_2023*, que corresponde al *promedio de costos del delito contra personas, por entidad federativa*, además de que tiene un pico pronunciado marcado hacia arriba en la variable *seg_4_2023* correspondiente a la *Cifra negra de delitos en la entidad federativa*, por lo que se podría pensar que las dos variables mencionadas son las que define al componente cuatro.

1. En el primer componente hay 7 variables que tienen pesos similares y son los más altos, rondando en valore de 0.3.
2. Para el segundo componente hay 2 variables con cargas positivas elevadas las cuales son *seg_7_2023*, *seg_1_2023* con cargas de 0.508 y 0.502 respectivamente, además, en tercer lugar quedó la variable *seg_13_2023* con carga de -0.343 y ésta última hace contraste con los componentes antes mencionados.
3. En el componente 3, se observa que la variable con mayor carga la tiene la variable *seg_7_2023* con un valor de 0.584, y luego, la carga de la variable *seg_6_2023* con el siguiente valor mayor es 0.430.
4. En el componente 4, se puede considerar como un componente de contraste, pues se observa que la variable con mayor carga (negativa) la tiene la variable *seg_8_2023* con un valor de -0.55, y la variable *seg_4_2023* hace contraste con un valor de 0.536, le sigue la variable *seg_6_2023* con un peso de 0.467.
5. Finalmente, en el componente 5 la variable que mayor carga es *seg_5_2023* con un peso de 0.671 y las demás variables, tienen pesos muy pequeños en este componente.

Para aterrizar lo anterior, se dibuja el diagrama donde se indica qué variables son las que explican a cada uno de los componentes.

```
fa.diagram(seg.pc$loadings[,1:5], main = "Compontes principales")
```

Compontes principales



Según el diagrama anterior, el componente uno está definido por las variables: seg_14_2023, seg_15_2023, seg_16_2023, seg_2_2023, seg_3_2023, seg_11_2023 y seg_12_2023 todas en sentido positivo y con una carga muy similar, este componente puede estar explicando el desempeño institucional, percepción de inseguridad y cambio de rutinas y Nivel de criminalidad (incidencia delictiva). El componente dos se define por seg_7_2023, seg_1_2023 y seg_13_2023, éste último con carga negativa, que se refiere a temas con el crimen organizado, Nivel de criminalidad (Tasa de homicidios) y la tasa de policías. El componente tres se define por la variable seg_17_2023 con carga positiva, relacionados con el desempeño institucional en particular con la variación porcentual del índice de conflictividad. El componente cuatro se define por las variables seg_8_2023 (carga negativa), seg_4_2023 y seg_6_2023 referente al costo del delito y nivel de criminalidad. Y finalmente, el componente cinco esta definido por las variables seg_5_2023 referente al nivel de criminalidad en particular con el indicador: prevalencia delictiva por cada diez mil unidades económicas.

Ahora, se pueden graficar los scores (a nivel dato) por pares de los componentes principales. Por temas de interpretación, solo se graficarán los scores de los dos primeros componentes principales.

2.5 Gráficas de los scores

Ahora se graficaran los scores de los pares (comp1 vs comp2), (comp1 vs comp3) y (comp2 vs comp3).

Biplot componente 1 vs componente 2

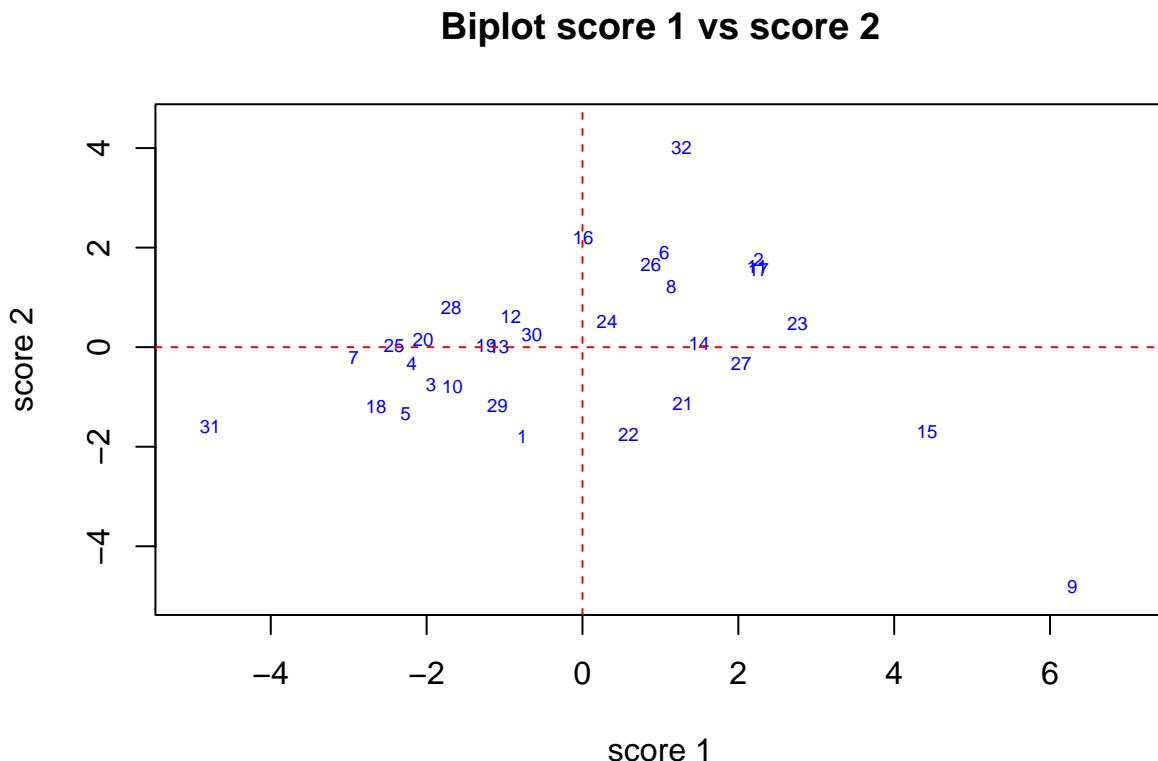
Primero se extraen los puntajes de las observaciones (scores).

```
#Se extraen los scores
score1<-seg.pc$scores[,1]
score2<-seg.pc$scores[,2]
score3<-seg.pc$scores[,3]
score4<-seg.pc$scores[,4]
score5<-seg.pc$scores[,5]
```

Se realiza el gráfico score 1 vs score 2

```
#Se realiza el gráfico

plot(score1,score2,main="Biplot score 1 vs score 2",xlim=c(-5,7),ylim=c(-5,4.5),
  xlab="score 1",ylab="score 2",type="n",lwd=2)
text(score1,score2,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty =2, col = "red")
```



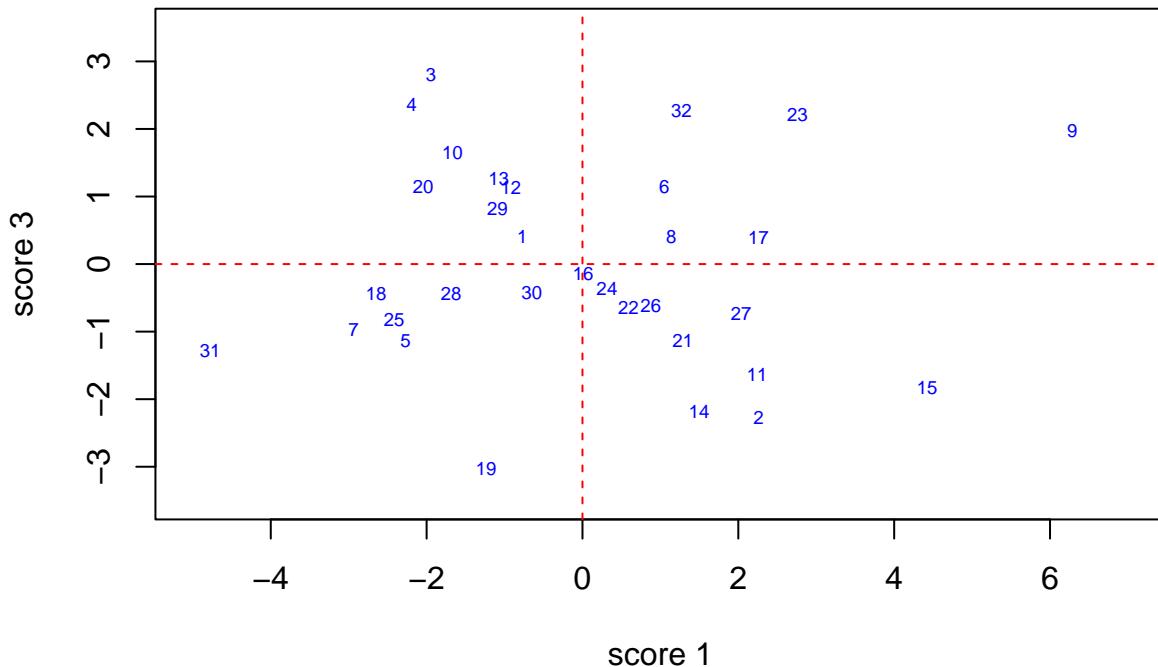
Por ejemplo, la entidad 9 (CDMX), para el componente 1, registra un valor score muy alto (mayores niveles de incidencia, mayor percepción de inseguridad y finalmente mucha desconfianza en las autoridades), sin embargo, en el componente 2 relacionado con una baja tasa de homicidios, bajos niveles en el índice por crimen organizado y una alta tasa de policías, podría concluirse que a pesar de que hay altos niveles delictivos, y la desconfianza en la pob es alta, uno de los delitos más dañinos en la sociedad tiene registros muy bajos. Y en el otro extremo está la entidad 31 (Yucatán) que registra menores tasas de incidencia se podría catalogar como la entidad con menos problemas de inseguridad.

Se realiza el gráfico score 1 vs score 3

```
#Se realiza el gráfico

plot(score1,score3,main="Biplot score 1 vs score 3",xlim=c(-5,7),ylim=c(-3.5,3.5),
  xlab="score 1",ylab="score 3",type="n",lwd=2)
text(score1,score3,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

Biplot score 1 vs score 3



Para esta combinación de scores, se tiene que la entidad 9 (CDMX) para el componente 1, registra un valor score muy alto (mayores niveles de incidencia, mayor percepción de inseguridad y finalmente mucha desconfianza en las autoridades), además con respecto al valor del score 3 es un valor medianamente alto, el cual está relacionado con la variación porcentual del índice de conflictividad en la entidad federativa. En el caso de la entidad 15 (EDOMEX) para el componente 1, registra un valor score muy alto (mayores niveles de incidencia, mayor percepción de inseguridad y finalmente mucha desconfianza en las autoridades), sin embargo, el score 3 es un valor muy bajo, el cual se puede decir que la variación del índice de conflictividad no cambió en esa entidad. En el otro extremo está la entidad 31 (Yucatán) que registra menores tasas de incidencia se podría catalogar sin embargo, con lo que respecta al score 3 está en el promedio, es decir, si hubo variación en su índice. Finalmente, la entidad 19 (Nuevo León) para el componente 1, registra un valor score ni muy baos ni muy altos, sin embargo, el score 3 es un valor muy bajo, el cual se puede decir que la variación del índice de conflictividad no cambió en esa entidad.

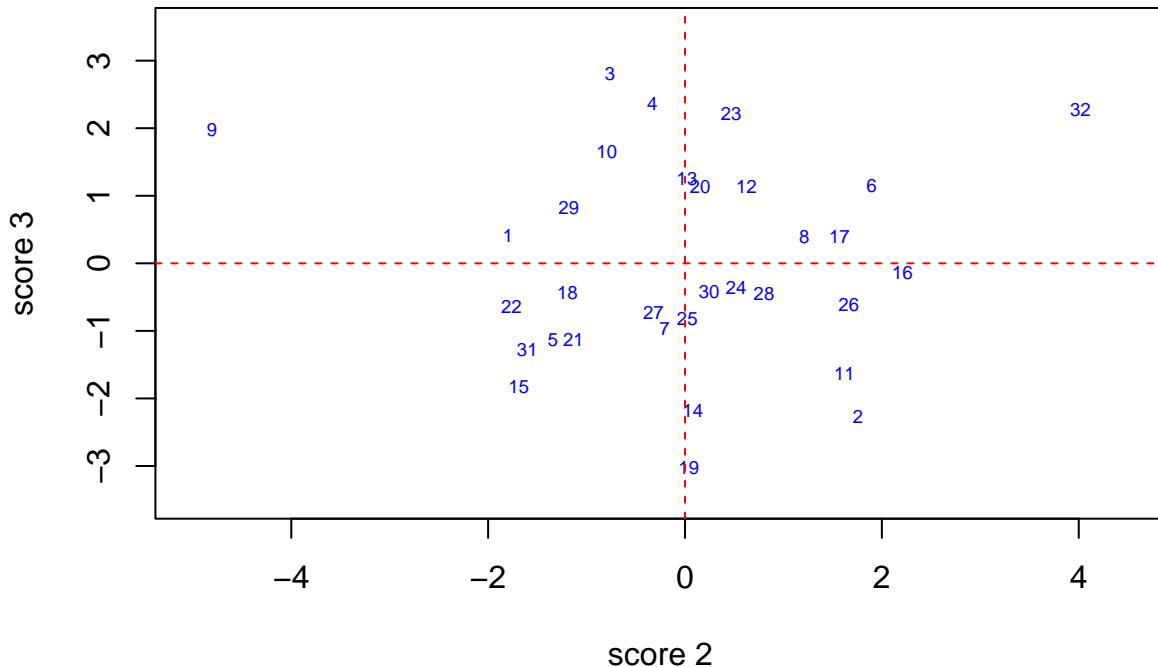
Se realiza el gráfico score 2 vs score 3

```
#Se realiza el gráfico

plot(score2,score3,main="Biplot score 2 vs score 3",xlim=c(-5,4.5),ylim=c(-3.5,3.5),
  xlab="score 2",ylab="score 3",type="n",lwd=2)
```

```
text(score2,score3,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty =2, col ="red")
```

Biplot score 2 vs score 3



Con respecto a las interpretaciones del score 2 vs el score 3 se tiene lo siguiente: En el primer cuadrante se encuentra la entidad 9 (CDMX) con los menores puntajes del score 2 y que tienen que ver con la tasa de homicidios baja, el Índice de Riesgo por Crimen Organizado bajo y una tasa de policías muy alta y por otro lado, el score 3 que representa la cifra negra personal y de empresa medianamente alta y con un costo del delito alto. Para la entidad 19 (Nuevo León) el score 2 se puede interpretar como un valor promedio, sin embargo el score 3, registra un valor bajo y esto porque la cifra negra de delitos en unidades económicas registra el valor más bajo entre todas las entidades federativas. Finalmente, la entidad 32 (Zacatecas) registra valores muy altos en el score 2 y también registra valores muy altos en el score 3 que tienen que ver con un alto porcentaje de delitos no denunciados, así como un costo del delito muy alto, es decir, es una entidad con altos niveles de violencia y alto porcentaje de no denuncia por parte de los ciudadanos y también les genera al costos por la inseguridad.

2.6 Cálculo del subíndice de seguridad

Ahora, se obtienen los scores para crear un data frame para hacer la escala del 0-100.

Opción 1. Se obtiene el promedio de los 5 scores, y sobre ese promedio se hace el escalamiento de 0-100. A continuación, se detalla el cálculo.

NOTA 1: El cálculo del subíndice de seguridad *ind_seg* se realizó de la siguiente manera:

- a. A partir de los scores elegidos se obtiene la media de los mismos por entidad federativa y se nombra *media_seg*.
- b. Se obtiene el valor mínimo de la nueva variable calculada *media_seg* y a ese valor se le asigna el nombre de *val_min*.
- c. Se obtiene la diferencia de la media (*media_seg*) y el valor mínimo de la media (*val_min*) y se nombra *dif*.
- d. Se obtiene el valor máximo de la diferencia *dif* y a ese valor se le asigna el nombre de *val_max*.
- e. Se calcula el subíndice de seguridad, que es dividir la diferencia *dif* entre el valor máximo de la diferencia *val_max*, multiplicado por 100. De esta manera, se obtiene el valor del subíndice para cada una de las entidades federativas.

```

bd_score<-cbind(score1,score2,score3,score4,score5)
#Se convierte en data frame
bd_score<- as.data.frame(bd_score)
bd_score$media_seg<- apply(bd_score,1,mean)
val_min<- min(bd_score$media_seg)# -1.57714
bd_score$dif<-(bd_score$media_seg-val_min)
val_max<- max(bd_score$dif)#3.192246
bd_score$ind_seg<-(bd_score$dif/val_max)*100
#bd_score #Se muestra el data frame construido, la última columna "ind_seg" es el índice de seguridad.
#Finalmente se agrega la variable entidad a la base final
bd_score$ID<-c("01","02","03","04","05","06","07","08","09",
               "10","11","12","13","14","15","16","17","18",
               "19","20","21","22","23","24","25","26","27",
               "28","29","30","31","32")

```

Opción 2. Se realiza el escalamiento por score y sobre eso se realiza el promedio. ¡Ojo en este no se obtendrá un escalamiento del 0-100 por el promedio.

NOTA 2: La segunda opción de cálculo del subíndice de seguridad *ind_seg_f* se realizó de la siguiente manera:

- a. Se obtiene el valor mínimo de cada uno de los scores, de esta manera, se obtendrán n valores mínimos, donde n es el número de scores elegidos, en este ejemplo son 5 scores. A cada valor mínimo se le asigna un nombre, val_{min_i} y $i = 1, 2, 3...n$
- b. Se obtiene la diferencia del $score_i$ y el valor mínimo de cada score val_{min_i} y se nombra $diff_i$
- c. Se obtiene el valor máximo para cada una de las diferencias $diff_i$ y a ese valor se le asigna el nombre de val_{max_i} .
- d. Se calcula el subíndice de seguridad ind_{seg_i} que es dividir la diferencia $diff_i$ entre el valor máximo de la diferencia val_{max_i} , multiplicado por 100. De esta manera, se obtiene el valor del ind_{seg_i} para cada una de las entidades federativas.
- e. Finalmente, se promedian los ind_{seg_i} y se obtiene el subíndice de seguridad *ind_seg_f*

```

bd_score1<-cbind(score1,score2,score3,score4,score5)
#Se convierte en data frame
bd_score1<- as.data.frame(bd_score1)
val_min1<- min(bd_score1$score1)
val_min2<- min(bd_score1$score2)

```

```

val_min3<- min(bd_score1$score3)
val_min4<- min(bd_score1$score4)
val_min5<- min(bd_score1$score5)
bd_score1$dif1<-(bd_score1$score1-val_min1)
bd_score1$dif2<-(bd_score1$score2-val_min2)
bd_score1$dif3<-(bd_score1$score3-val_min3)
bd_score1$dif4<-(bd_score1$score4-val_min4)
bd_score1$dif5<-(bd_score1$score5-val_min5)
val_max1<- max(bd_score1$dif1)
val_max2<- max(bd_score1$dif2)
val_max3<- max(bd_score1$dif3)
val_max4<- max(bd_score1$dif4)
val_max5<- max(bd_score1$dif5)
bd_score1$ind_seg1<-(bd_score1$dif1/val_max1)*100
bd_score1$ind_seg2<-(bd_score1$dif2/val_max2)*100
bd_score1$ind_seg3<-(bd_score1$dif3/val_max3)*100
bd_score1$ind_seg4<-(bd_score1$dif4/val_max4)*100
bd_score1$ind_seg5<-(bd_score1$dif5/val_max5)*100
bd_score1$ind_seg_f<-apply(bd_score1[,c(11:15)],1,mean)
#Finalmente se agrega la variable entidad a la base final
bd_score1$ID<-c("01","02","03","04","05","06","07","08","09",
                 "10","11","12","13","14","15","16","17","18",
                 "19","20","21","22","23","24","25","26","27",
                 "28","29","30","31","32")

```

3. ANÁLISIS DE FACTORES

Análisis factorial confirmatorio

Es una técnica basada en el análisis de estructuras de covarianzas que tiene como objetivo determinar si un modelo de medida especificado por el investigador, basándose en hipótesis teóricas o en un AFE previo, es consistente con la realidad. Para llegar a obtener una conclusión al respecto es preciso abordar una serie de fases comunes al conjunto de los procedimientos que operan con ecuaciones estructurales.

Especificación del modelo

Se establece un modelo tomando en cuenta lo siguiente:

- Número de factores latentes comunes
- Número de variables observables
- Relación entre factores comunes
- Relación entre variables observables y factores comunes
- Relación entre factores únicos y variables observables
- Relación entre factores únicos.

Para el establecimiento formal de la estructura del modelo, el investigador se basará en el AFE y/o en un sólido sustento teórico. A partir de esta base, deberá reflejar mediante ecuaciones la estructura relacional del modelo de medida, es decir, deberá especificar los parámetros del modelo que determinarán las relaciones existentes entre las variables observables y las latentes.

Identificación del modelo

A la hora de determinar si un modelo está o no identificado, caben 3 soluciones posibles:

1. Que el modelo esté exactamente identificado, en cuyo caso se podrá estimar cada parámetro estructural a partir de una única combinación de los elementos de la matriz de covarianzas.
2. Que el modelo esté sobre identificado, es decir, al menos un parámetro podrá obtenerse a partir de dos o más ecuaciones diferentes.
3. Que el modelo esté infra identificado, en cuyo caso no será posible establecer ecuaciones de covarianza para alguno de los parámetros, por lo que no todos podrán ser estimados.

La forma más efectiva de comprobar si un modelo está identificado es demostrando algebraicamente que es posible igualar cada parámetro estructural a una combinación de los elementos de la matriz de covarianza; como en la práctica esta tarea es tediosa y compleja, se han propuesto algunas reglas o condiciones que permiten determinar más fácilmente el estatus de la identificación del modelo. En este sentido, la regla t , según la cual el número de parámetros a estimar ha de ser igual o inferior al número de momentos no redundantes de la matriz de covarianzas:

$$t \leq p(p+1)/2$$

Donde t es el número de parámetros libres y p es el número de variables observables.

Estimación de los parámetros del modelo

Este proceso tiene como objetivo encontrar los valores que generen una matriz de covarianzas estimada tan próxima como sea posible a la matriz de covarianzas muestrales, que, supuestamente, será un estimador consistente de la matriz de covarianzas poblacional.

El más común es el Método de Máxima Verosimilitud, que proporciona estimaciones consistentes, eficientes y no sesgadas cuando se cumple el supuesto de normalidad multivariante. Otros métodos son el de mínimos cuadrados generalizados o el de distribución asintótica libre.

Número de factores a seleccionar

Cuando se tienen algunas expectativas sobre la manera en que se agruparán las variables, se puede utilizar un criterio a priori, es decir, solicitar la extracción de un número de factores determinado.

Otro criterio, probablemente el más utilizado, es la regla de Gutman-Kaiser (o regla K1), esto es, conservar aquellos factores con valores propios mayores que 1. En este caso, resultaría útil consultar el gráfico de sedimentación (scree plot) y, además de aquellos factores que se encuentren antes del punto de inflexión del gráfico, retener algún factor más, siempre que su varianza se pueda considerar relativamente alta o próxima al límite establecido.

Otras pautas para considerar son el análisis paralelo, que es muy similar al gráfico de sedimentación, y el comparativo de la matriz reproducida a partir del número de factores seleccionados, con la matriz original, a lo que se le denomina matriz de residuales. El porcentaje de la varianza total explicada es otro criterio que se puede utilizar para decidir el número de factores que serán retenidos, como umbral para la extracción de los factores se suele establecer un mínimo de 70%.

Generación de los scores

Habiendo determinado el número de factores seleccionados, se calculan las matrices de puntuaciones factoriales F , esto para conocer qué sujetos son los más raros o extremos, identificar dónde se ubican ciertos grupos o subcolectivos de la muestra, o simplemente para determinar qué factor sobresale uno del otro.

Los métodos de estimación más utilizados son por Regresión y Barlett.

- Método de Regresión: Estima F por el método de los mínimos cuadrados: $\hat{F} = (A'A)^{-1}A'X$
- Método de Barlett: Utiliza el método de los mínimos cuadrados generalizados estimando las puntuaciones factoriales mediante: $\hat{F} = (A'\psi^{-1}A)^{-1}A'\psi^{-1}X$

Interpretación de resultados

La matriz de cargas factoriales (A) tiene un papel fundamental en la interpretación. Por otra parte, las cargas factoriales al cuadrado (a_{il}^2) indican si los factores son ortogonales, qué porcentaje de la variable original (X_i) es explicado por el factor F_l .

A efectos prácticos, en la interpretación de los factores se sugiere:

- Identificar las variables cuyas correlaciones con el factor son las más elevadas en valor absoluto.
- Intentar dar un nombre a los factores. El nombre se asigna de acuerdo con la estructura de las correlaciones.
- Una ayuda en la interpretación de los factores puede ser la representación gráfica de los resultados obtenidos, sobre los ejes factoriales se proyectan las variables originales.
 - Las variables al final de un eje son aquellas que tienen correlaciones altas solo en ese factor y en consecuencia lo describen.
 - Las variables que están cerca del origen tienen correlaciones reducidas en ambos factores.
 - Las variables que no están cerca de ninguno de los ejes se relacionan con ambos factores.
- Eliminar las cargas factoriales bajas y de este modo suprimir información redundante. El investigador decide a partir de qué valor deben eliminarse las cargas factoriales, generalmente, se toma como significativas las cargas superiores a 0.5 en valor absoluto.

3.1 Análisis de factibilidad

Antes de dar inicio con la implementación del modelo de factores como tal, es importante analizar la factibilidad de los datos, para ello se analiza la matriz de correlaciones, el cálculo del determinante de la matriz anterior y se generan las pruebas de Kaiser Meyer Olkin (KMO) y de esfericidad de Bartlet.

3.1.1 Prueba de Kaiser Meyer Olkin

La prueba de Kaiser-Meyer-Olkin (KMO), es una medida de la idoneidad de los datos para el análisis factorial. Mide la adecuación del muestreo para cada variable en el modelo y para el modelo completo. La estadística es una medida de la proporción de varianza entre variables que podrían ser varianza común.

La prueba de KMO devuelve valores entre 0 y 1. Una regla general para interpretar la estadística es la siguiente:

- Los valores de KMO entre 0.8 y 1 indican que el muestreo es perfecto para realizar un análisis de factores.
- Los valores entre 0.6 y 0.7 señalan que es viable realizar un análisis de factores.
- Los valores de KMO entre 0.4 y 0.5 indican que es aceptable realizar un análisis de factores, no obstante, se sugiere verificar el estado de las variables.
- Los valores de KMO menores a 0.4 significan que existen grandes correlaciones parciales en comparación con la suma de las correlaciones. En otras palabras, existen correlaciones generalizadas que son un gran problema para el análisis factorial, por lo que no se recomienda efectuar un análisis factorial.

El resultado de la prueba se muestra a continuación:

```
# Prueba de KMO
KMO(r = cor(est_seg))
```

```

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(est_seg))
## Overall MSA = 0.57
## MSA for each item =
##   seg_1_2023  seg_2_2023  seg_3_2023  seg_4_2023  seg_5_2023  seg_6_2023
##   0.52        0.61        0.62        0.30        0.59        0.60
##   seg_7_2023  seg_8_2023  seg_9_2023  seg_11_2023 seg_12_2023 seg_13_2023
##   0.51        0.43        0.62        0.64        0.72        0.46
##   seg_14_2023 seg_15_2023 seg_16_2023 seg_17_2023 seg_18_2023
##   0.57        0.70        0.57        0.41        0.43

```

De acuerdo con la salida anterior, para el modelo completo se tiene un valor de 0.57, lo que implica que la tabla de datos sea aceptable para efectuar un análisis de factores (AF), no obstante, se sugiere verificar el estado de las variables. Por otro lado, en el caso de cada variable, por ejemplo, las variables *seg_2_2023*, *seg_3_2023*, *seg_6_2023*, *seg_9_2023*, *seg_11_2023*, *seg_12_2023* y *seg_15_2023* tienen valores entre 0.6-0.7, lo que implica que para el AF estas variables son viables; las variables *seg_1_2023*, *seg_5_2023*, *seg_7_2023*, *seg_8_2023*, *seg_13_2023*, *seg_14_2023*, *seg_16_2023*, *seg_17_2023* y *seg_18_2023* tienen valores entre 0.4-0.5, por lo que para el AF estas variables son aceptables para realizar un análisis de factores, finalmente, la variable *seg_4_2023* tiene un valor de 0.3, por lo que no se recomienda efectuar un análisis factorial.

3.1.2 Prueba de esfericidad de Bartlett

En la prueba de esfericidad de Bartlett se analiza si tiene sentido efectuar un análisis de factores a partir de las siguientes hipótesis: $H_0 : R = I$ vs $H_a : R \neq I$, donde R es la matriz de correlaciones e I es una matriz identidad del mismo tamaño que R . En caso de no rechazar H_0 , no tendría sentido realizar el análisis de factores. La regla establecida es rechazar H_0 si el valor p es menor a 0.05, el resultado de la prueba es el siguiente.

```

# Prueba
corptest.bartlett(est_seg)

```

```

## R was not square, finding R from data

## $chisq
## [1] 377.7452
##
## $p.value
## [1] 1.2688e-24
##
## $df
## [1] 136

```

Derivado de la anterior, ya que el valor p es casi 0 y éste es menor que 0.05 ($0 < 0.05$) se rechaza H_0 , es decir, tiene sentido continuar con el análisis de factores.

En conclusión, el análisis de factibilidad determina que es viable realizar el análisis de factores con la tabla de datos seleccionada.

4 Elección del número de factores.

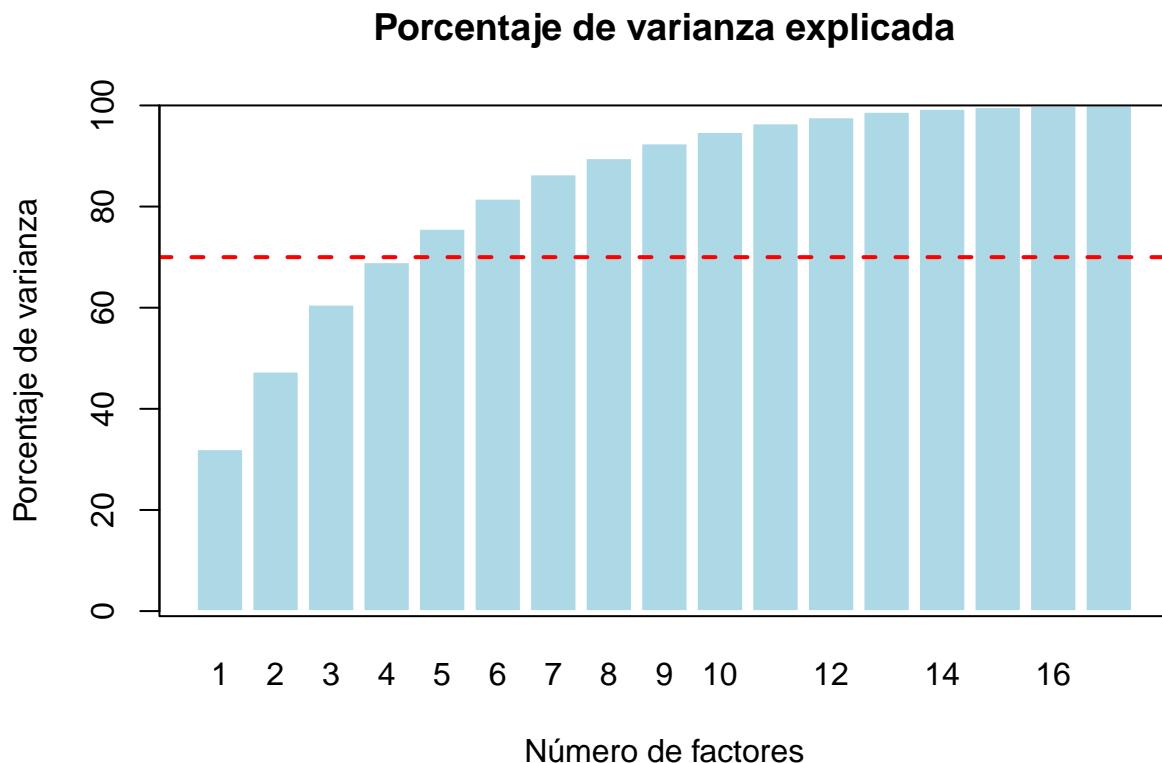
4.1 Porcentaje de varianza explicada

Para este criterio, por lo general se eligen el número de factores que explican de manera acumulada más del 70% de la varianza total. Para ello se obtienen los valores propios de la matriz de correlaciones y se divide entre el total de las variables de estudio. Posteriormente, se efectúa la suma acumulada para determinar el porcentaje de varianza explicada, según el número de factores seleccionados.

```
# Evitar la notación científica
options(scipen = 999)
# Se extraen los valores propios de la matriz de correlaciones
val_propios<-eigen(cor_seg)$values
# Porcentaje de varianza explicada
acumulado<-cumsum(val_propios*100/ncol(est_seg))
round(acumulado, 2)

## [1] 32.03 47.37 60.61 69.00 75.61 81.57 86.37 89.57 92.49 94.76
## [11] 96.46 97.65 98.74 99.33 99.68 99.88 100.00

# Gráfico de la varianza acumulada
barplot(acumulado, col = "lightblue", border = "white", names.arg = 1:17,
        main = "Porcentaje de varianza explicada",
        xlab = "Número de factores", ylab = "Porcentaje de varianza")
abline(h = 70, lty = 2, col = "red", lwd = 2)
box()
```



De acuerdo con las salidas previas, con cinco factores se supera el 70% del total de la varianza explicada (línea roja punteada del gráfico), ya que se obtiene el 75.61%, con seis factores se llega al 81.57%, y así sucesivamente. Aunque si se consideran 4 factores se acumula el 69% de la varianza explicada. Con este criterio cinco factores se deben considerar.

4.2 Criterio de Kaiser

Es uno de los criterios más utilizados por los paquetes estadísticos para determinar el número adecuado de factores. Se aplica cuando se trabaja con la matriz de correlaciones muestrales R . Consiste en elegir el número de valores propios de R que sean mayores a 1. Esta regla busca que cualquier factor retenido contenga al menos la varianza de una de las variables utilizadas en el análisis.

```
# Se imprimen los valores propios a dos decimales
round(val_propios, 2)
```

```
## [1] 5.44 2.61 2.25 1.43 1.12 1.01 0.82 0.54 0.50 0.39 0.29 0.20 0.19 0.10 0.06
## [16] 0.03 0.02
```

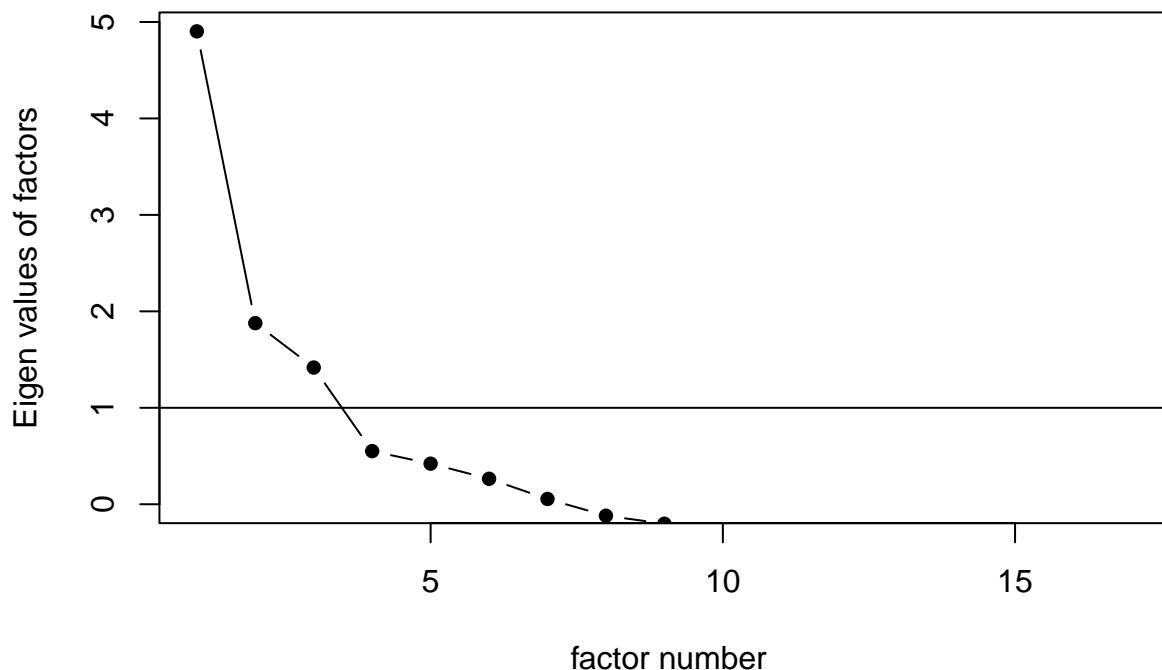
Con base a este criterio y a la salida anterior, se deberían elegir 6 factores, ya que sus valores propios son de 5.44, 2.61, 2.25, 1.43, 1.12 y 1.01, de forma respectiva, sin embargo, el sexto factor es casi 1. Por lo que se puede considerar el tomar solo 5 factores.

4.3 Gráfico de sedimentación

Este criterio complementa al anterior y se basa también en el análisis de la magnitud de los valores propios, pero a partir de la tendencia que se observa en el gráfico. Se procura seleccionar un grupo reducido de factores que tengan valores propios significativamente superiores a los demás, para lo cual se identifica el punto de inflexión en la curva del gráfico de sedimentación (también referido como el codo por su semejanza con un brazo) a partir del cual la curva se transforma a una línea “plana” o relativamente recta. El gráfico para este ejercicio se presenta a continuación.

```
# Gráfico de sedimentación
scree(est_seg, main = "Gráfico de sedimentación", pc=FALSE, fa=TRUE)
```

Gráfico de sedimentación



En la gráfica previa se aprecia un punto de inflexión después de 2 factores para AF, mientras que la regla de Kaiser sugiere tres.

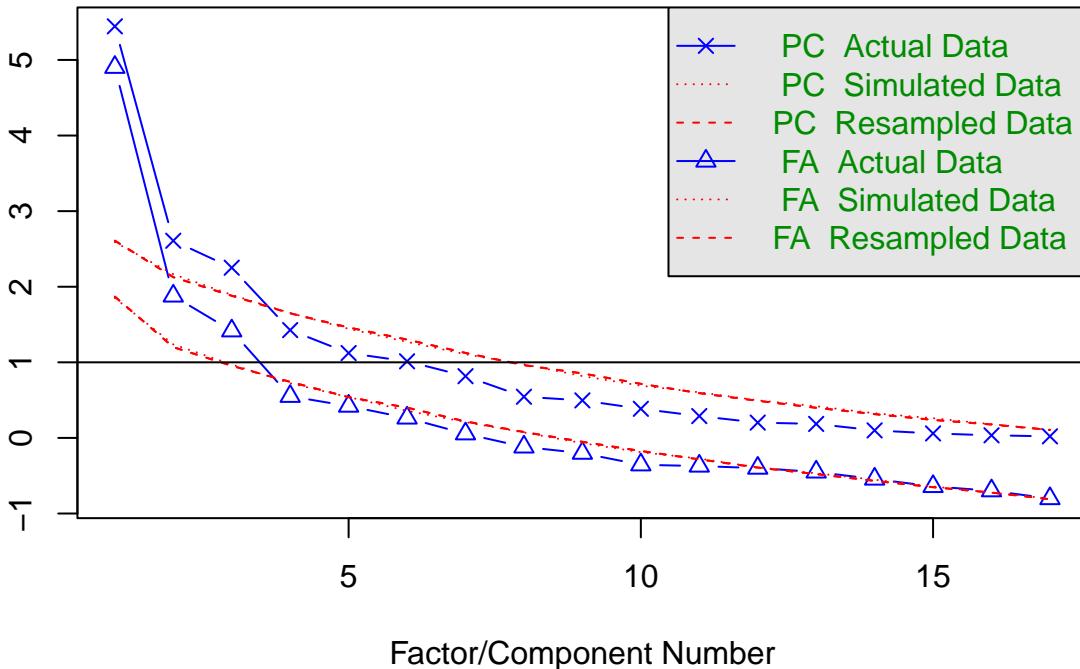
4.4 Análisis paralelo

El análisis paralelo suele complementar los anteriores cuando el número de variables iniciales y factores resultantes es elevado. El procedimiento se basa en el principio de que los factores a extraer deben dar cuenta de más varianza que la esperada de manera aleatoria. El procedimiento reordena las observaciones de manera aleatoria entre cada variable y los valores propios son recalculados a partir de esta nueva base de datos aleatoriamente ordenada. Los factores con valores propios mayores a los valores aleatorios son retenidos para interpretación.

Se comparan gráficamente las dos líneas del gráfico de sedimentación, una para el análisis de los datos reales y otra para el análisis de los datos aleatorios. El número de factores se determina con la intersección de ambas gráficas. El resultado de este criterio se muestra en la próxima gráfica.

```
# Análisis paralelo
fa.parallel(est_seg, main = "Análisis paralelo")
```

Análisis paralelo



```
## Parallel analysis suggests that the number of factors = 3 and the number of components = 3
```

De acuerdo con el análisis paralelo, sugiere conservar 3 factores en el caso de un AF y 3 componentes para un ACP. Hasta el momento los análisis sugieren considerar entre 2 y 5 factores.

5. Análisis factorial confirmatorio

Para el análisis factorial confirmatorio se obtiene la solución por máxima verosimilitud de \mathbf{L} (cargas) y Ψ (varianzas específicas) para 5 y 6 factores, considerando una rotación *varimax* para facilitar la interpretación de los factores.

5.1 Porcentaje de varianza explicada

La rotación de las cargas busca mejorar la interpretación de los factores. Al rotarlos, la suma de los valores propios no se altera, pero podrían cambiar los valores propios, el porcentaje de varianza explicada y los valores de carga de cada factor. Por lo anterior, el porcentaje de varianza explicada por cada uno de los modelos propuestos se muestra a continuación.

```
# Cargas estimadas rotadas para 5 factores por máxima verosimilitud
af.mv5<-factanal(est_seg, factors = 5, rotation = "varimax", n.obs = nrow(est_seg))
# Se almacenan las cargas
cargas_mv5<-af.mv5$loadings
vx5<-colSums(cargas_mv5^2)
```

```
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx5, `Proporción de la varianza` = vx5/nrow(cargas_mv5),
      `Varianza acumulada` = cumsum(vx5/nrow(cargas_mv5)))
```

```
##                               Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings             2.9098849 2.2470516 2.2209281 2.0651285 1.60384582
## Proporción de la varianza 0.1711697 0.1321795 0.1306428 0.1214781 0.09434387
## Varianza acumulada      0.1711697 0.3033492 0.4339920 0.5554702 0.64981406
```

De acuerdo al resultado previo, si se consideran 5 factores el porcentaje de la varianza acumulada es del 64.98%.

Por otra parte, el resultado para el modelo con 6 factores se presenta en seguida.

```
# Cargas estimadas rotadas para 6 factores por máxima verosimilitud
af.mv6<-factanal(est_seg, factors = 6, rotation = "varimax", n.obs = nrow(est_seg))
# Se almacenan las cargas
cargas_mv6<-af.mv6$loadings
vx6<-colSums(cargas_mv6^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx6, `Proporción de la varianza` = vx6/nrow(cargas_mv6),
      `Varianza acumulada` = cumsum(vx6/nrow(cargas_mv6)))
```

```
##                               Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings             2.7302603 2.5786124 2.5057423 1.67182464 1.41475382
## Proporción de la varianza 0.1606035 0.1516831 0.1473966 0.09834263 0.08322081
## Varianza acumulada      0.1606035 0.3122866 0.4596832 0.55802586 0.64124667
##                               Factor6
## SS loadings              1.18766357
## Proporción de la varianza 0.06986256
## Varianza acumulada      0.71110924
```

Cuando se ajusta el modelo con 6 factores, el porcentaje de la varianza acumulada es del 71.11%, por lo que el modelo con 6 factores aumenta aproximadamente el 7% el porcentaje de la varianza acumulada en comparación con el modelo de 5 factores. Con base en el análisis previo para elegir el número de factores, se considera que se trabajarán las comunidades y varianzas específicas con 5 y 6 factores.

5.2 Estimación de cargas mediante máxima verosimilitud

Se realizaron las estimaciones de las cargas rotadas, las **comunidades**, que es la parte de la variación de la variable que es compartida con otras variables, y las **varianzas únicas o específicas**, que es la parte de la variación de la variable que es propia de esa variable, para los modelos de 5 y 6 factores.

```
# Modelo con 5 factores
# Comunidades
comu_mv5<-diag(cargas_mv5[, 1:5] %*% t(cargas_mv5[, 1:5]))
# Unicidades o varianza específica
espe_mv5<-1 - comu_mv5

# Se unen los resultados para el modelo de 5 factores
res_mv5<-data.frame(CARGA5 = cargas_mv5[, 1:5], COM5 = comu_mv5, ESP5 = espe_mv5)
```

```
# Se imprime el resultado de las cargas rotadas
kbl(res_mv5, longtable = T, booktabs = T, digits = 3,
  caption = "Modelo con factores por máxima verosimilitud",
  col.names = c("F_1", "F_2", "F_3", "F_4", "F_5", "Comunalidad", "Varianza única"))
```

Table 1: Modelo con factores por máxima verosimilitud

	F_1	F_2	F_3	F_4	F_5	Comunalidad	Varianza única
seg_1_2023	-0.011	0.921	-0.007	0.267	-0.102	0.930	0.070
seg_2_2023	0.956	-0.015	0.123	0.217	-0.134	0.995	0.005
seg_3_2023	0.893	0.001	0.225	0.113	-0.218	0.909	0.091
seg_4_2023	0.026	-0.125	-0.351	0.207	-0.076	0.188	0.812
seg_5_2023	0.338	0.261	0.125	0.016	0.213	0.244	0.756
seg_6_2023	-0.169	-0.007	0.018	-0.138	0.590	0.396	0.604
seg_7_2023	0.000	0.805	-0.035	0.431	0.032	0.837	0.163
seg_8_2023	0.081	0.431	0.273	-0.081	0.092	0.282	0.718
seg_9_2023	0.504	0.264	0.199	0.149	-0.267	0.456	0.544
seg_11_2023	0.143	0.356	0.195	0.885	-0.165	0.995	0.005
seg_12_2023	0.326	0.255	0.147	0.707	0.021	0.694	0.306
seg_13_2023	0.416	-0.204	0.315	0.125	0.345	0.448	0.552
seg_14_2023	0.237	0.337	0.672	0.367	-0.081	0.763	0.237
seg_15_2023	0.469	-0.027	0.708	0.289	0.152	0.829	0.171
seg_16_2023	0.355	-0.078	0.877	0.290	0.102	0.995	0.005
seg_17_2023	-0.236	0.212	0.125	-0.110	0.851	0.853	0.147
seg_18_2023	0.191	-0.114	0.061	0.210	0.370	0.234	0.766

En la salida anterior se puede apreciar que, la varianza única de *seg_4_2023* (Cifra negra de delitos en la entidad federativa) es la más alta con 0.812, lo que implica que los factores no están explicando de buena forma la variabilidad de esta variable, asimismo, *seg_18_2023* (Índice de Impunidad estatal en materia penal en entidades federativas), *seg_5_2023* (Prevalencia delictiva por cada diez mil unidades económicas), *seg_8_2023* (Promedio de costos del delito contra personas, por entidad federativa), *seg_6_2023* (Cifra negra de delitos en unidades económicas), *seg_13_2023* (Tasa de policías por cada cien mil habitantes) y *seg_9_2023* (Promedio de costos del delito en unidades económicas, por entidad federativa) cuentan con varianzas específicas de 0.766, 0.756, 0.718, 0.604, 0.552 y 0.0544, de forma respectiva.

De manera análoga, el resultado para el modelo con 6 factores es el siguiente.

```
# Modelo con 6 factores
# Comunalidades
comu_mv6<-diag(cargas_mv6[, 1:6] %*% t(cargas_mv6[, 1:6]))
# Unicidades o varianza específica
espe_mv6<-1 - comu_mv6

# Se unen los resultados para el modelo de 6 factores
res_mv6<-data.frame(CARGA6 = cargas_mv6[, 1:6], COM6 = comu_mv6, ESP6 = espe_mv6)
# Se imprime el resultado de las cargas rotadas
kbl(res_mv6, longtable = T, booktabs = T, digits = 3,
  caption = "Modelo con 6 factores por máxima verosimilitud",
  col.names = c("F_1", "F_2", "F_3", "F_4", "F_5", "F_6", "Comunalidad", "Varianza única"))
```

Table 2: Modelo con 6 factores por máxima verosimilitud

	F_1	F_2	F_3	F_4	F_5	F_6	Comunalidad	Varianza única
seg_1_2023	-0.044	0.904	0.012	-0.070	0.164	0.068	0.855	0.145
seg_2_2023	0.285	0.012	0.914	-0.182	0.189	-0.098	0.995	0.005
seg_3_2023	0.311	-0.013	0.857	-0.265	0.145	0.119	0.936	0.064
seg_4_2023	-0.115	0.007	-0.023	-0.040	0.012	-0.925	0.871	0.129
seg_5_2023	0.096	0.159	0.334	0.197	0.155	0.205	0.251	0.749
seg_6_2023	0.000	-0.095	-0.169	0.613	-0.047	-0.101	0.426	0.574
seg_7_2023	0.014	0.922	0.020	0.045	0.210	-0.091	0.905	0.095
seg_8_2023	0.183	0.421	0.089	0.083	-0.107	0.324	0.342	0.658
seg_9_2023	0.260	0.297	0.489	-0.279	-0.007	0.041	0.475	0.525
seg_11_2023	0.370	0.558	0.108	-0.202	0.569	-0.136	0.843	0.157
seg_12_2023	0.240	0.309	0.262	0.007	0.879	-0.004	0.995	0.005
seg_13_2023	0.456	-0.143	0.398	0.301	-0.136	-0.089	0.504	0.496
seg_14_2023	0.699	0.395	0.163	-0.081	0.219	0.146	0.746	0.254
seg_15_2023	0.794	0.031	0.382	0.096	0.162	0.120	0.828	0.172
seg_16_2023	0.943	-0.009	0.242	0.062	0.166	0.129	0.995	0.005
seg_17_2023	0.044	0.160	-0.186	0.903	-0.074	0.191	0.920	0.080
seg_18_2023	0.119	-0.050	0.209	0.306	0.163	0.148	0.202	0.798

Al incluir más factores en el modelo, en algunas variables las comunalidades se incrementan, por ende, las especificidades o varianzas únicas disminuyen. De acuerdo con la tabla anterior, al incluir un factor más en el modelo, la especificidad de *seg_4_2023* pasó de 0.812 a 0.129, lo cual es una disminución importante. Por otra parte, para las variables *seg_18_2023*, *seg_5_2023*, *seg_8_2023*, *seg_6_2023*, *seg_13_2023* y *seg_9_2023*, las varianzas únicas fueron de 0.798 (0.766 para el modelo de 5 factores), 0.749 (0.756 para el modelo de 5 factores), 0.658 (0.718 para el modelo de 5 factores), 0.574 (0.604 para el modelo de 5 factores), 0.496 (0.552 para el modelo de 5 factores) y 0.525 (0.544 para el modelo de 5 factores), de manera correspondiente. Solo en la variable *seg_4_2023* se aprecia una disminución considerable, en las demás variables la varianza única es muy similar que en el modelo de 5 factores.

5.3 Prueba de hipótesis para el número de factores comunes

El supuesto de normalidad permite realizar pruebas sobre la suficiencia del modelo. Es decir, se puede probar si el número de factores m es suficiente para explicar la covarianza observada. Suponiendo que el modelo de m factores comunes tiene buen ajuste, entonces $\Sigma = \mathbf{LL}' + \Psi$, y probar el ajuste del modelo de m factores comunes es equivalente a probar: $H_0: \Sigma = \mathbf{LL}' + \Psi$ (El número de factores en el modelo, es suficiente para capturar la dimensionalidad completa del conjunto de datos).

$H_1: \Sigma \neq \mathbf{LL}' + \Psi$ (El número de factores en el modelo, no es suficiente para capturar la dimensionalidad completa del conjunto de datos).

La regla de decisión es rechazar H_0 si el valor p es menor que 0.05. Tal resultado indica que el número de factores es demasiado pequeño. Por el contrario, no rechazamos H_0 si el valor p excede 0.05. Tal resultado indica que es probable que haya suficientes (o mas que suficientes) factores que capturen la dimensionalidad completa del conjunto de datos.

El resultado de la prueba para el modelo de 5 factores es el siguiente.

Prueba de hipótesis

```
# Valor p para el modelo de 5 factores
af.mv5$PVAL
```

```
## objective
## 0.2230781
```

Como el valor p es mayor a 0.05 ($0.2230781 > 0.05$) no se rechaza H_0 , es decir, 5 factores en el modelo son suficientes para capturar la dimensionalidad completa del conjunto de datos.

```
# Valor p para el modelo de 6 factores
af.mv6$PVAL
```

```
## objective
## 0.2886158
```

Como el valor p es mayor a 0.05 ($0.2886158 > 0.05$) no se rechaza H_0 , es decir, 6 factores en el modelo son suficientes para capturar la dimensionalidad completa del conjunto de datos.

NOTA: Se realizó la prueba para 3 y 4 factores y en ambos casos se rechazó la hipótesis H_0 . Las pruebas están en la sección del *ANEXO*

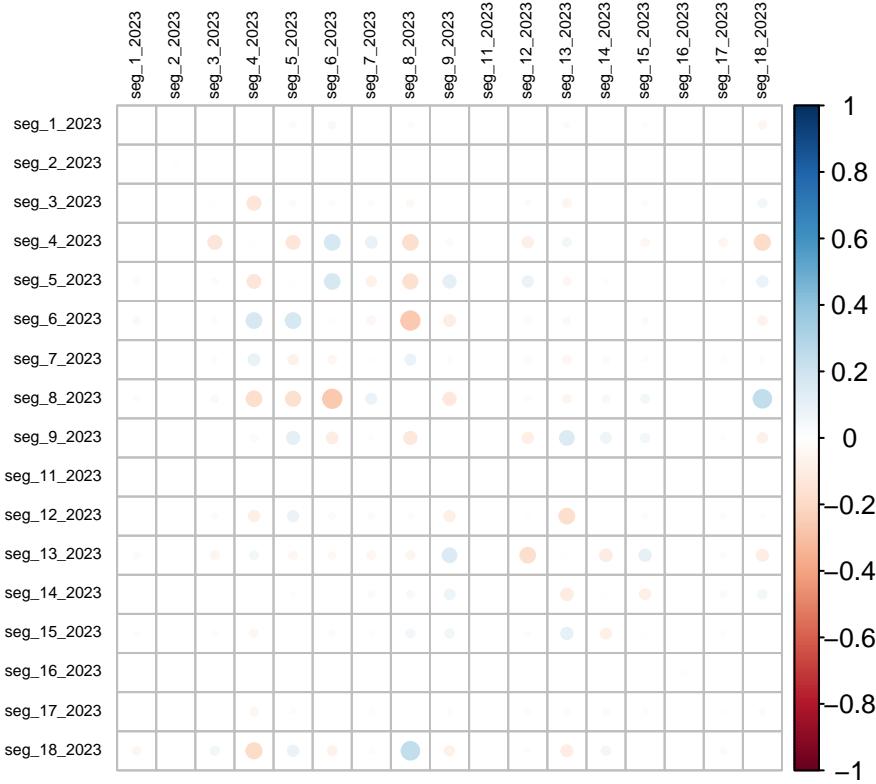
5.4 Matriz de residuales

Se realiza la comparación de las matrices de correlación, en el cual se obtiene la diferencia entre la matriz de correlaciones real y la matriz reproducida mediante m factores, si el número de factores m es adecuado, la diferencia debe ser cercana a 0. Para ello se calculan las correlaciones reproducidas, que es la matriz de correlaciones de las variables originales que resultaría suponiendo que es correcto el número de factores retenidos. Asimismo, se calcula la matriz de correlaciones residuales, que es la matriz de las diferencias entre la matriz de correlaciones reproducida y la matriz de correlaciones reales. Si el número de factores m es adecuado, se esperaría que la matriz de correlaciones residuales tenga todas sus entradas cercanas a 0.

El resultado para el modelo de 5 factores es el siguiente.

```
# Matriz reproducida (5 factores)
mat_rep_mv5<-(cargas_mv5[, 1:5] %*% t(cargas_mv5[, 1:5])) + diag(espe_mv5)
# Matriz de residuales (5 factores)
mat_res_mv5<-cor_seg - mat_rep_mv5
# Gráficos de correlaciones
corrplot(mat_res_mv5, order = "original", tl.col = "black", tl.cex = 0.5,
         mar = c(0, 0, 2, 0), title = "Matriz de residuales para 5 factores por MV")
```

Matriz de residuales para 5 factores por MV

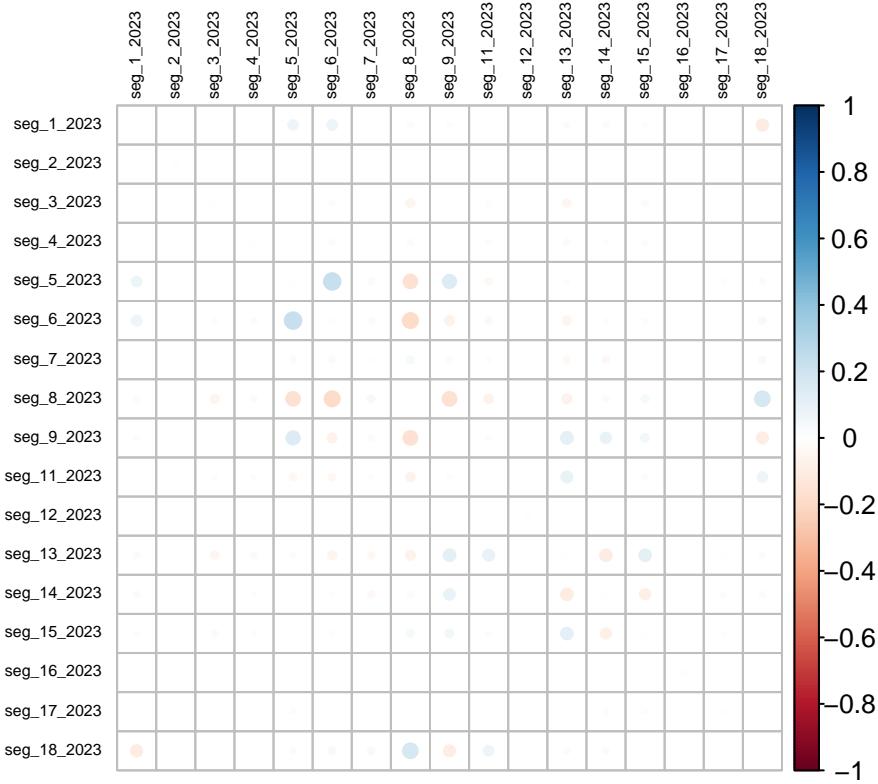


Con base a la gráfica previa, se aprecia que la mayoría de las celdas son cercanas a 0, ya que no se observan círculos de color intenso. Los residuales más grandes que pueden distinguirse son para los pares (*seg_6_2023*, *seg_8_2023*), (*seg_8_2023*, *seg_18_2023*), (*seg_4_2023*, *seg_18_2023*), (*seg_4_2023*, *seg_6_2023*) y (*seg_4_2023*, *seg_6_2023*) cuyos valores son: -0.2692, 0.2494, 0.1859, 0.1756 y 0.1746, de forma respectiva. En conclusión, con cinco factores se realiza un buen ajuste, aunque se tienen algunos residuales con valores mayor a 0 pero menor a 0.3.

A continuación, se muestra el resultado para el modelo de 6 factores.

```
# Matriz reproducida (6 factores)
mat_rep_mv6<-cargas_mv6[, 1:6] %*% t(cargas_mv6[, 1:6]) + diag(espe_mv6)
# Matriz de residuales (6 factores)
mat_res_mv6<-cor_seg - mat_rep_mv6
# Gráficos de correlaciones
corrplot(mat_res_mv6, order = "original", tl.col = "black", tl.cex = 0.5,
         mar = c(0, 0, 2, 0), title = "Matriz de residuales para 6 factores por MV")
```

Matriz de residuales para 6 factores por MV



De acuerdo con la salida anterior, la gran mayoría de las celdas son cercanas a 0, puesto que no se observan círculos de color intenso. Los residuales que más puede distinguirse son para el par de variables (*seg_5_2023*, *seg_6_2023*), (*seg_8_2023*, *seg_18_2023*), (*seg_5_2023*, *seg_8_2023*) y (*seg_5_2023*, *seg_9_2023*), con valores de: 0.2205, 0.1747, -0.1509 y 0.1405, por lo que con 6 factores prácticamente los residuales son cercanos a 0.

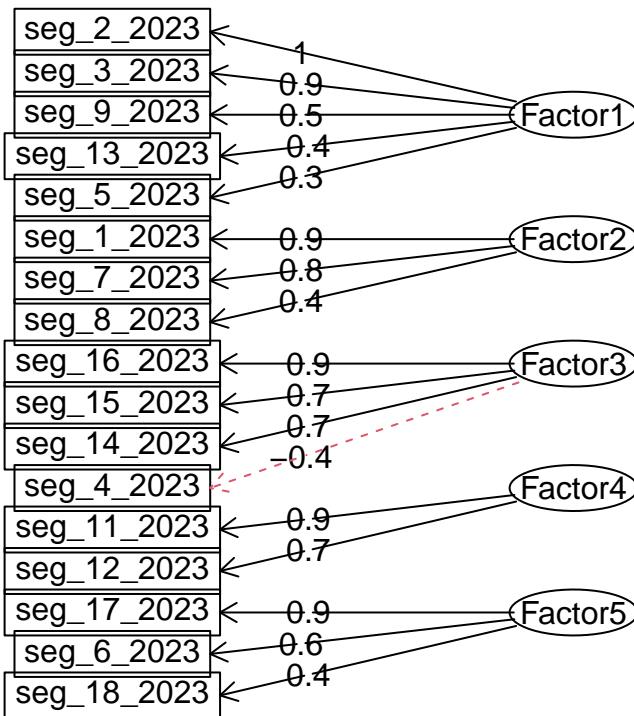
Hasta el momento, con el modelo de 5 factores se explica el 64.98% de la varianza, la prueba de hipótesis para el número de factores comunes determinó que 5 factores son suficientes, y la matriz de residuales arrojó 5 pares de valores considerables. Por otra parte, con el modelo de 6 factores se explica el 71% de la varianza, de igual forma, la prueba de hipótesis sugirió que 6 factores son suficientes, y la matriz de residuales arrojó valores muy cercanos a cero.

5.5 Interpretación de los factores

La interpretación de las cargas de los modelos, generalmente se explican en términos de las cargas principales en cada factor. Para ello se realizó el siguiente gráfico para el modelo de 5 factores.

```
# Gráfico para interpretaciones de los factores
fa.diagram(af.mv5$loadings, main = "Modelo de 5 factores", cex = 1, rsize = 0.7)
```

Modelo de 5 factores



1.- Para el diagrama con 5 factores, vemos que el **factor 1** se asocia con las variables: *seg_2_2023, seg_3_2023, seg_9_2023, seg_13_2023* y *seg_5_2023*, estas variables están relacionadas con: Incidencia delictiva por cada cien mil habitantes, Prevalencia delictiva por cada cien mil habitantes, Promedio de costos del delito en unidades económicas, por entidad federativa, Tasa de policías por cada cien mil habitantes y Prevalencia delictiva por cada diez mil unidades económicas.

2.- El **factor 2** está asociado con las variables: *seg_1_2023, seg_7_2023* y *seg_8_2023*, estas variables son: Tasa de homicidios por cada cien mil habitantes, Índice de Riesgo por Crimen Organizado y Promedio de costos del delito contra personas, por entidad federativa.

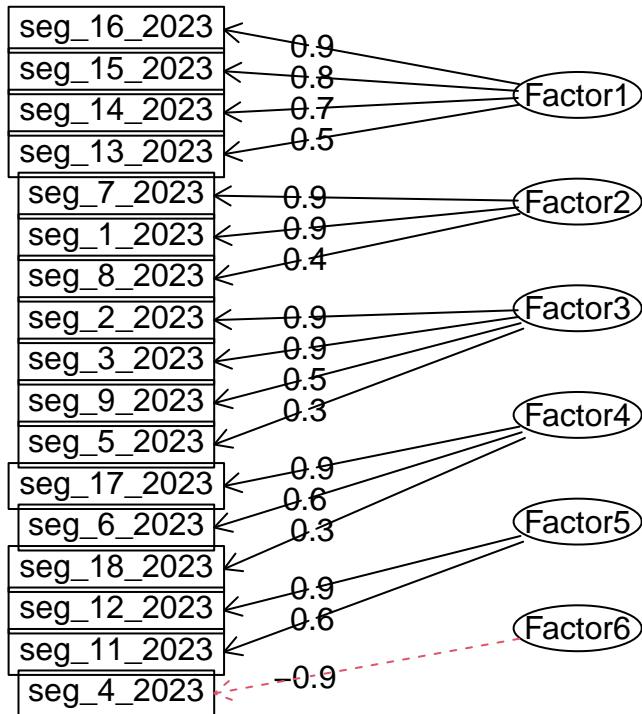
3.- El **factor 3** está asociado con las variables: *seg_16_2023, seg_15_2023, seg_14_2023* y *seg_4_2023*, estas variables son: Desconfianza en ministerios públicos, Desconfianza en jueces, Desconfianza en la policía estatal y municipal y Cifra negra de delitos en la entidad federativa con carga negativa.

4.- El **factor 4** está asociado con las variables: *seg_11_2023* y *seg_12_2023*, estas variables son: Porcentaje de personas que consideran insegura su entidad federativa y Porcentaje de personas que cambió algún hábito por temor a sufrir algún delito.

5.- El **factor 5** está asociado con las variables: *seg_17_2023, seg_6_2023* y *seg_18_2023*, estas variables son: Variación porcentual del índice de conflictividad en la entidad federativa, Cifra negra de delitos en unidades económicas y Índice de Impunidad estatal en materia penal en entidades federativas.

```
# Gráfico para interpretaciones de los factores
fa.diagram(af.mv6$loadings, main = "Modelo de 6 factores", cex = 1, rsize = 0.7)
```

Modelo de 6 factores



1.- Para el diagrama con 6 factores, vemos que el **factor 1** se asocia con las variables: *seg_16_2023, seg_15_2023, seg_14_2023 y seg_13_2023*, estas variables están relacionadas con: Desconfianza en ministerios públicos, Desconfianza en jueces, Desconfianza en la policía estatal y municipal y Tasa de policías por cada cien mil habitantes.

2.- El **factor 2** está asociado con las variables: *seg_7_2023, seg_1_2023 y seg_8_2023*, estas variables son: Índice de Riesgo por Crimen Organizado, Tasa de homicidios por cada cien mil habitantes y Promedio de costos del delito contra personas, por entidad federativa.

3.- El **factor 3** está asociado con las variables: *seg_2_2023, seg_3_2023, seg_9_2023 y seg_5_2023*, estas variables son: Incidencia delictiva por cada cien mil habitantes, Prevalencia delictiva por cada cien mil habitantes, Promedio de costos del delito en unidades económicas, por entidad federativa y Prevalencia delictiva por cada diez mil unidades económicas.

4.- El **factor 4** está asociado con las variables: *seg_17_2023, seg_6_2023 y seg_18_2023*, estas variables son: Variación porcentual del índice de conflictividad en la entidad federativa, Cifra negra de delitos en unidades económicas y Índice de Impunidad estatal en materia penal en entidades federativas.

5.- El **factor 5** está asociado con las variables: *seg_12_2023 y seg_11_2023*, estas variables son: Porcentaje de personas que cambió algún hábito por temor a sufrir algún delito y Porcentaje de personas que consideran insegura su entidad federativa

6.- El **factor 6** está asociado con la variable: *seg_4_2023*, esta variable es: Cifra negra de delitos en la entidad federativa.

Con base en el análisis anterior, se sugiere considerar 5 factores en el análisis, a pesar de que se explica un 64% de la varianza, sin embargo, las demás pruebas consideraron que con 5 factores el análisis es pertinente. Además, siguiendo el principio de parsimonia estadística es más sencillo interpretar 5 factores que 6.

5.6 Estimación de factor scores

Para la estimación de los vectores de puntajes factoriales (factor scores), se implementaron los métodos más consistentes y utilizados, los cuales son el de mínimos cuadrados ponderados (Bartlett) y el método por regresión (Thompson). El método por mínimos cuadrados ponderados supone que el vector de valores de los factores para cada observación es un parámetro a estimar. El método por regresión supone que estos vectores son variables aleatorias. El resultado de la estimación de los factor scores para el modelo de 5 factores se presenta a continuación.

Cálculo de factor scores mediante enfoque de mínimos cuadrados

```
# Cálculo de factor scores mediante enfoque de mínimos cuadrados
scores_min<-factanal(est_seg, factors = 5, scores = "Bartlett")$scores
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score1<-data.frame(ID = paste0(ENT), scores_min)
# Se ordenan los resultados por ID
res_score1<-res_score1[order(res_score1$ID), ]

# Se imprime el resultado por mínimos cuadrados
kbl(res_score1[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3", "Score 4", "Score 5"),
  row.names = FALSE,
  caption = "Factor Scores estimados por Mínimos cuadrados") %>%
add_header_above(c(" ", "Mínimos cuadrados" = 5)) %>%
kable_styling(latex_options = c("repeat_header"))
```

Table 3: Factor Scores estimados por Mínimos cuadrados

Entidad	Mínimos cuadrados				
	Score 1	Score 2	Score 3	Score 4	Score 5
01	0.611	-0.614	-0.594	-0.909	-0.087
02	0.547	2.354	0.411	-1.108	-2.174
03	-0.049	0.319	0.308	-2.501	1.380
04	-0.851	-0.289	0.227	-0.644	1.775
05	-0.320	-0.851	0.441	-1.301	-0.892
06	-0.113	1.495	-0.586	0.730	0.352
07	-1.837	-1.248	-0.308	0.833	-0.996
08	-0.045	1.763	0.524	-0.828	-0.593
09	2.296	-1.018	2.007	0.857	1.727
10	-1.343	-0.657	1.300	-0.766	0.371
11	0.218	1.235	-0.112	0.677	-0.411
12	0.065	0.306	-0.709	0.614	1.727
13	-1.106	-0.585	0.208	-0.003	0.472
14	0.303	-0.301	0.061	0.375	-1.737
15	1.926	-0.765	1.598	0.832	-1.560
16	-1.800	0.669	1.715	0.343	-1.375
17	0.417	0.910	0.480	0.659	-0.170
18	-0.782	-0.619	0.172	-1.084	-1.030
19	1.112	-0.653	-2.987	0.452	-0.720

Table 3: Factor Scores estimados por Mínimos cuadrados (*continued*)

Entidad	Mínimos cuadrados				
	Score 1	Score 2	Score 3	Score 4	Score 5
20	-1.239	-0.597	-0.192	0.743	0.978
21	0.988	-1.059	0.079	0.786	-0.317
22	0.699	-0.565	0.633	-1.181	-0.853
23	0.608	0.599	1.187	0.329	1.811
24	0.274	-0.077	-0.592	0.573	0.507
25	0.805	-0.102	-2.046	-0.545	0.259
26	0.359	1.332	-0.460	0.084	-0.596
27	0.841	-0.626	0.146	1.179	0.036
28	-1.152	-0.764	-1.066	1.261	-0.220
29	0.490	-0.693	-0.846	-0.143	1.063
30	-1.464	-0.990	0.180	1.260	-0.325
31	0.129	-0.527	-0.728	-2.490	0.191
32	-0.589	2.617	-0.450	0.915	1.409

Cálculo de factor scores mediante enfoque de regresión

```
# Cálculo de factor scores mediante enfoque de regresión
scores_reg<-factanal(est_seg, factors = 5, scores = "regression")$scores
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score2<-data.frame(ID = paste0(ENT), scores_reg)
# Se ordenan los resultados por ID
res_score2<-res_score2[order(res_score2$ID), ]

# Se imprime el resultado por regresión
kbl(res_score2[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3", "Score 4", "Score 5"),
  row.names = FALSE,
  caption = "Factor Scores estimados por regresión") %>%
add_header_above(c(" ", "Regresión" = 5)) %>%
kable_styling(latex_options = c("repeat_header"))
```

Table 4: Factor Scores estimados por regresión

Entidad	Regresión				
	Score 1	Score 2	Score 3	Score 4	Score 5
01	0.607	-0.598	-0.589	-0.910	-0.081
02	0.571	2.120	0.262	-0.899	-1.788
03	-0.077	0.211	0.300	-2.450	1.281
04	-0.875	-0.277	0.277	-0.693	1.558
05	-0.293	-0.850	0.384	-1.248	-0.697
06	-0.131	1.429	-0.571	0.736	0.265
07	-1.794	-1.121	-0.314	0.789	-0.865

Table 4: Factor Scores estimados por regresión (*continued*)

Entidad	Regresión				
	Score 1	Score 2	Score 3	Score 4	Score 5
08	-0.041	1.596	0.445	-0.705	-0.440
09	2.260	-0.948	2.077	0.770	1.464
10	-1.326	-0.648	1.274	-0.753	0.414
11	0.216	1.169	-0.131	0.712	-0.373
12	0.026	0.332	-0.620	0.529	1.423
13	-1.100	-0.537	0.224	-0.038	0.427
14	0.336	-0.282	0.007	0.416	-1.498
15	1.953	-0.731	1.544	0.865	-1.340
16	-1.750	0.607	1.610	0.438	-1.078
17	0.414	0.858	0.463	0.685	-0.151
18	-0.750	-0.621	0.113	-1.030	-0.822
19	1.102	-0.554	-2.923	0.381	-0.770
20	-1.243	-0.509	-0.134	0.657	0.818
21	0.992	-0.965	0.104	0.735	-0.323
22	0.715	-0.589	0.577	-1.119	-0.678
23	0.571	0.559	1.230	0.298	1.569
24	0.259	-0.039	-0.546	0.522	0.383
25	0.781	-0.084	-1.995	-0.579	0.146
26	0.356	1.242	-0.491	0.143	-0.525
27	0.838	-0.546	0.182	1.119	-0.028
28	-1.138	-0.641	-1.021	1.178	-0.258
29	0.464	-0.630	-0.776	-0.218	0.863
30	-1.436	-0.871	0.196	1.197	-0.295
31	0.122	-0.567	-0.744	-2.447	0.218
32	-0.628	2.487	-0.415	0.919	1.181

Cálculo de las diferencias

```
# Cálculo de las diferencias
diferencia<-scores_min - scores_reg
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score3<-data.frame(ID = paste0(ENT), diferencia)
# Se ordenan los resultados por ID
res_score3<-res_score3[order(res_score3$ID), ]

# Se imprime el resultado de las diferencias
kbl(res_score3[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3", "Score 4", "Score 5"),
  row.names = FALSE,
  caption = "Diferencias") %>%
add_header_above(c(" ", "Diferencias" = 5)) %>%
kable_styling(latex_options = c("repeat_header"))
```

Table 5: Diferencias

Entidad	Diferencias				
	Score 1	Score 2	Score 3	Score 4	Score 5
01	0.004	-0.016	-0.006	0.000	-0.006
02	-0.024	0.234	0.149	-0.209	-0.386
03	0.027	0.107	0.008	-0.051	0.099
04	0.023	-0.013	-0.050	0.049	0.217
05	-0.027	-0.001	0.057	-0.053	-0.196
06	0.018	0.066	-0.015	-0.006	0.087
07	-0.043	-0.127	0.006	0.044	-0.131
08	-0.004	0.168	0.079	-0.123	-0.153
09	0.036	-0.070	-0.070	0.087	0.262
10	-0.017	-0.008	0.027	-0.013	-0.043
11	0.002	0.066	0.019	-0.035	-0.038
12	0.040	-0.025	-0.089	0.085	0.304
13	-0.006	-0.048	-0.016	0.035	0.045
14	-0.033	-0.018	0.054	-0.041	-0.238
15	-0.027	-0.034	0.054	-0.033	-0.219
16	-0.050	0.063	0.105	-0.095	-0.297
17	0.003	0.052	0.016	-0.027	-0.019
18	-0.031	0.002	0.059	-0.054	-0.209
19	0.011	-0.099	-0.064	0.070	0.050
20	0.004	-0.088	-0.058	0.086	0.160
21	-0.004	-0.094	-0.025	0.051	0.006
22	-0.016	0.024	0.056	-0.062	-0.175
23	0.037	0.039	-0.044	0.031	0.242
24	0.015	-0.038	-0.046	0.052	0.124
25	0.025	-0.018	-0.050	0.034	0.113
26	0.003	0.091	0.031	-0.058	-0.071
27	0.003	-0.080	-0.037	0.060	0.063
28	-0.014	-0.123	-0.045	0.083	0.037
29	0.026	-0.063	-0.071	0.075	0.200
30	-0.028	-0.119	-0.016	0.063	-0.030
31	0.007	0.040	0.016	-0.043	-0.027
32	0.040	0.130	-0.035	-0.005	0.227

```
# Cálculo del promedio de las diferencias
round(apply(diferencia, 2, mean), 3)
```

```
## Factor1 Factor2 Factor3 Factor4 Factor5
##     0     0     0     0     0
```

```
# Cálculo de la desviación estándar de las diferencias
round(apply(diferencia, 2, sd), 3)
```

```
## Factor1 Factor2 Factor3 Factor4 Factor5
## 0.025 0.087 0.056 0.070 0.175
```

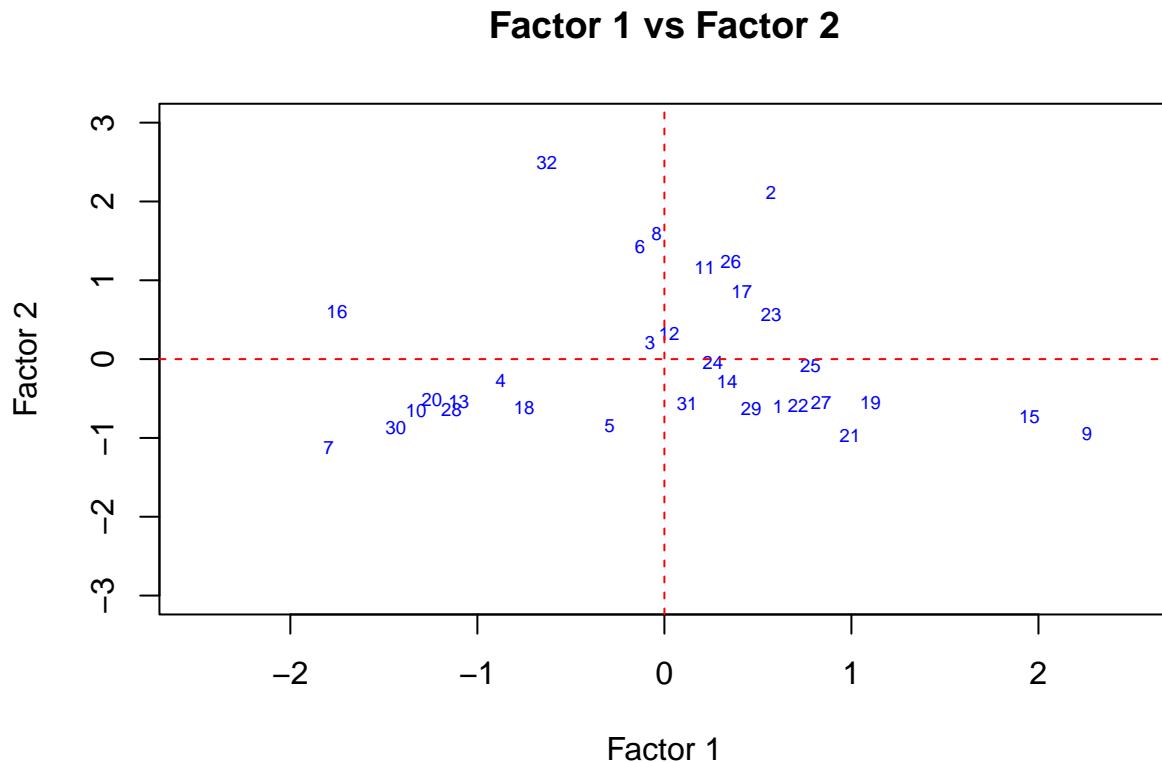
Las tablas anteriores contienen las estimaciones de los factor scores para las 32 entidades. Asimismo, se muestra el promedio de las diferencias (prácticamente 0) y sus desviaciones estándar (cerca de 0) entre los enfoques de mínimos cuadrados y regresión. Se puede apreciar que las diferencias son mínimas, así que las interpretaciones que se den sobre las entidades serán aplicables a ambos enfoques. Por lo anterior, se determinó trabajar con los factor scores estimados mediante regresión.

5.7 Interpretación de resultados

Se considera distinguir a las 32 entidades federativas. Para facilitar la interpretación de los resultados, se realizaron las diferentes parejas posibles de los factor scores, además de dividir la gráfica en 4 cuadrantes. Asimismo, se identificaron los puntos extremos de cada cuadrante (representados con color rojo) para analizar su información. El resultado se muestra a continuación.

Se realiza el gráfico del Factor 1 vs Factor 2 (Opción 2)

```
plot(res_score2[, 2], res_score2[, 3], main="Factor 1 vs Factor 2", xlim=c(-2.5,2.5), ylim=c(-3,3),
  xlab="Factor 1", ylab="Factor 2", type="n", lwd=2)
text(res_score2[, 2], res_score2[, 3],
  labels=abbreviate(row.names(res_score2), minlength=8), cex=0.6, lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

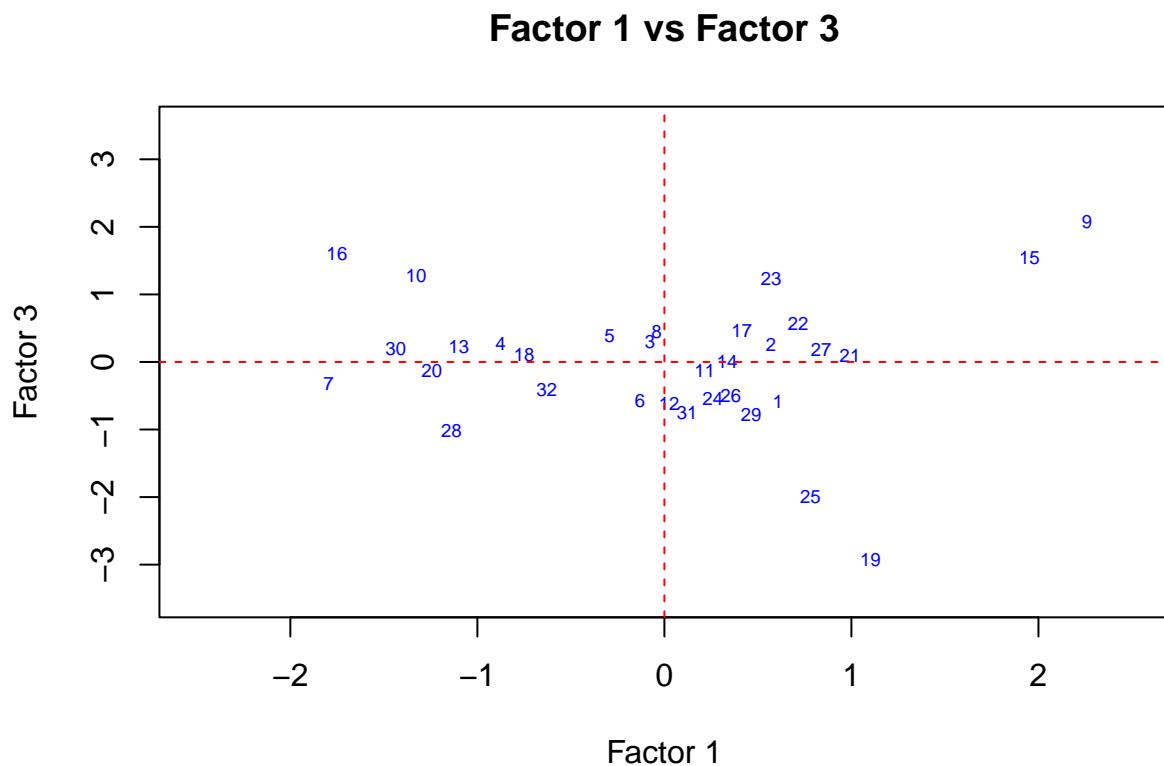


Para la interpretación de las gráficas se recuerda que el factor 1 está relacionado con el nivel de criminalidad, promedio del costo del delito por unidad económica y la tasa de policías, mientras el factor 2 se relaciona con la tasa de homicidios, índice de riesgo por el crimen organizado y el promedio de costos por delitos a nivel persona, finalmente, el factor 3 está relacionado con el desempeño institucional (desconfianza en autoridades) y cifra negra delitos a nivel persona. De esta manera, la entidad 9 (CDMX) es la que tiene el valor más

alto en el factor 1, es decir, tiene alto nivel de criminalidad, en cambio el con respecto al factor 2 tiene valor cercano a la media nacional. La entidad 15 (EDOMEX) tiene un comportamiento muy parecido al de CDMX. La entidad 2 (Baja California) también registra valores altos en el factor 1 y con respecto al factor 2, registra valores un poco más arriba que la media nacional. Para la entidad 32 (Zacatecas) se registra valores por debajo de la media nacional en el factor 1sin embargo, en el factor 2 es la entidad con los valores más altos. Par las entidades 16 (Michoacán) y 7 (Chiapas) registran los valores más pequeños en el factor 1, sin embargo, la primera entidad mencionada registra valores por arriba del promedio en el factor 2, mientras que Chiapas registra valores también muy pequeños en el factor 2. Finalmente, para el caso de la entidad 31 (Yucatán) registra un valor cercano a cero en factor 1, sin embargo, el valor del factor 2 es muy pequeño, casi a la misma altura que la entidad de Chiapas.

Se realiza el gráfico del Factor 1 vs Factor 3

```
plot(res_score2[, 2], res_score2[, 4], main="Factor 1 vs Factor 3", xlim=c(-2.5,2.5), ylim=c(-3.5,3.5),
  xlab="Factor 1", ylab="Factor 3", type="n", lwd=2)
text(res_score2[, 2], res_score2[, 4],
  labels=abbreviate(row.names(res_score2), minlength=8), cex=0.6, lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

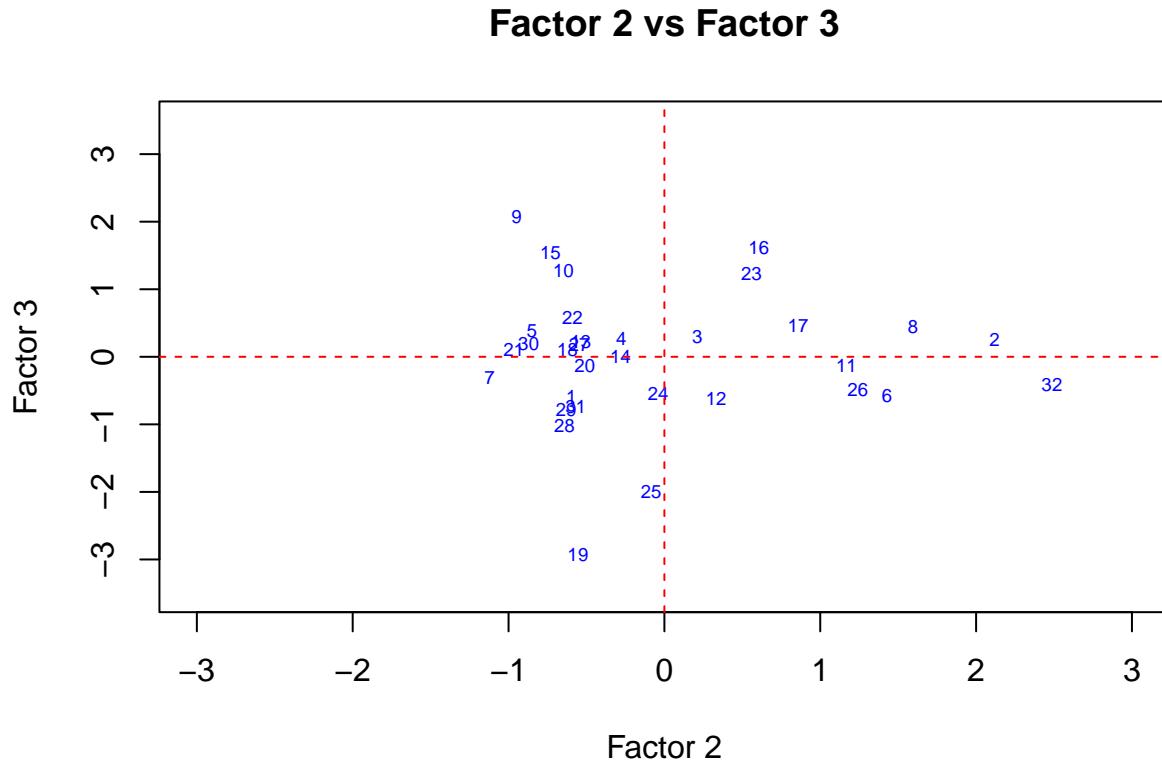


En la siguiente gráfica se muestra que la entidad 9 (CDMX) se encuentra en el cuadrante II reporta los valores más altos es ambos factores, lo que se interpreta como altos niveles de criminalidad y alta desconfianza en las autoridades analizadas, un comportamiento similar lo tiene la entidad 15 (EDOMEX). Por otro lado, la entidad 19 (Nuevo León) que se reporta en el cuadrante IV, registra un valor medianamente alto del factor 1 y un valor bajo del factor 3, es decir, existe hay nivel de criminalidad, sin embargo, la desconfianza en las autoridades analizadas es baja. En el cuadrante 1 de la gráfica se encuentra la entidad de 16 (Michoacán de Ocampo),en donde registra valores bajos en el factor 1 (nivel de criminalidad principalmente) sin embargo, la desconfianza en las autoridades es muy alta muy parecida a la de la CDMX.

En el extremo del cuadrante III, se encuentra la entidad 7 (Chiapas), en donde reporta un valor bajo en el factor 1 y un valor menor al promedio nacional en el factor 3. Las entidades (11, 14 y 2) Guanajuato, Jalisco y Baja California reportan valores considerables (aunque no los máximos) en el factor 1 y en el factor 3.

Se realiza el gráfico del Factor 2 vs Factor 3

```
plot(res_score2[, 3], res_score2[, 4], main="Factor 2 vs Factor 3", xlim=c(-3,3), ylim=c(-3.5,3.5),
  xlab="Factor 2", ylab="Factor 3", type="n", lwd=2)
text(res_score2[, 3], res_score2[, 4],
  labels=abbreviate(row.names(res_score2), minlength=8), cex=0.6, lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```



En el biplot anterior, se muestra la relación del factor 2 y factor 3, por ejemplo, en el cuadrante I, se observa que la entidad 9 (CDMX), en donde se registra un valor pequeño en el factor 2 (Tasa de homicidio, índice de riesgo por crimen organizado y promedio de cotos del delitos a nivel persona), pero un valor muy alto en el factor 3 (desconfianza en las y cifra negra a nivel persona.) En el cuadrante II, la entidad 2 (Baja California) es la que se encuentra en el extremo, registrando valor alto en el factor 2, sin embargo, en el factor 3 registra un valor un poco arriba del promedio del nacional.

En el cuadrante III, se encuentra la entidad 19 (Nuevo León) que registra un nivel por debajo del promedio en el factor 2 y el valor más bajo en el factor 3. Es decir, hay registros de la tasa de homicidios, índice de riesgo por crimen organizado, sin embargo, la desconfianza en las autoridades es baja. Finalmente, en el extremo del cuadrante IV, se encuentra la entidad 32 (Zacatecas), registrando valor alto en el factor 2, sin embargo, en el factor 3 registra un valor un poco por debajo del promedio nacional.

5.8 Cálculo del subíndice de seguridad

Ahora, se obtienen los factor scores para crear un data frame y obtener la primera opción del subíndice de seguridad.

Opción 1. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 1*, solo que en este caso, se obtienen los factor scores por el método de análisis de factores.

```
#Se extraen los 5 factor scores
bd_score2<-res_score2
#Se convierte bd_factor_score en data frame
bd_score2<- as.data.frame(bd_score2)
bd_score2$media_seg<- apply(bd_score2[,c(2:6)],1,mean)
val_min<- min(bd_score2$media_seg) # -0.6837303
bd_score2$dif<-(bd_score2$media_seg-val_min)
val_max<- max(bd_score2$dif) #1.808364
bd_score2$ind_seg<-(bd_score2$dif/val_max)*100 #La variable entidad ya está incluida
```

Opción 2. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 2*, solo que en este caso, se obtienen los factor scores por el método de análisis de factores.

```
#Se extraen los 5 factor scores
bd_score3<-res_score2
#Se convierte en data frame
bd_score3<- as.data.frame(bd_score3)
val_min1<- min(bd_score3$Factor1)
val_min2<- min(bd_score3$Factor2)
val_min3<- min(bd_score3$Factor3)
val_min4<- min(bd_score3$Factor4)
val_min5<- min(bd_score3$Factor5)
bd_score3$dif1<-(bd_score3$Factor1-val_min1)
bd_score3$dif2<-(bd_score3$Factor2-val_min2)
bd_score3$dif3<-(bd_score3$Factor3-val_min3)
bd_score3$dif4<-(bd_score3$Factor4-val_min4)
bd_score3$dif5<-(bd_score3$Factor5-val_min5)
val_max1<- max(bd_score3$dif1)
val_max2<- max(bd_score3$dif2)
val_max3<- max(bd_score3$dif3)
val_max4<- max(bd_score3$dif4)
val_max5<- max(bd_score3$dif5)
bd_score3$ind_seg1<-(bd_score3$dif1/val_max1)*100
bd_score3$ind_seg2<-(bd_score3$dif2/val_max2)*100
bd_score3$ind_seg3<-(bd_score3$dif3/val_max3)*100
bd_score3$ind_seg4<-(bd_score3$dif4/val_max4)*100
bd_score3$ind_seg5<-(bd_score3$dif5/val_max5)*100
bd_score3$ind_seg_f<-apply(bd_score3[,c(12:16)],1,mean)
```

6. Análisis de componentes principales versión 2

A continuación, se realiza el análisis de Componentes principales quitando algunas variables, para ello, se revisó a las variables que no tuviera alta correlación con las demás variables y fueron las que en este nuevo análisis no se consideraron.

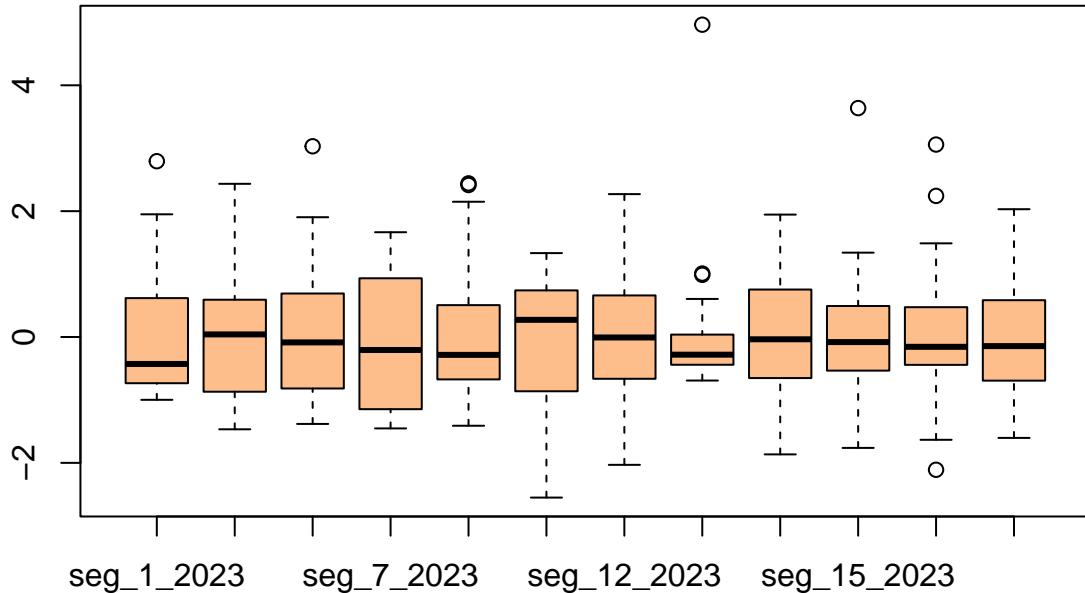
Las variables que se eliminaron fueron:

1. seg_4_2023: Cifra negra de delitos en la entidad federativa
2. seg_5_2023: Prevalencia delictiva por cada diez mil unidades económicas
3. seg_6_2023: Cifra negra de delitos en unidades económicas
4. seg_8_2023: Promedio de costos del delito contra personas, por entidad federativa
5. seg_18_2023: Índice de Impunidad estatal en materia penal en entidades federativas

De la base original “datos”, se crea la base “seg1” la cual no considera a las variables antes mencionadas.

```
seg1 <- datos[, -c(1,5,6,7,9,11,19,20)]
#Se estandarizan los datos
est_seg1<- scale(seg1, center = TRUE, scale = TRUE)
#Se realiza el gráfico de cajas para ver a las variables estandarizadas.
boxplot(est_seg1, col ="#FEBE8C", main="Datos estandarizados")
```

Datos estandarizados

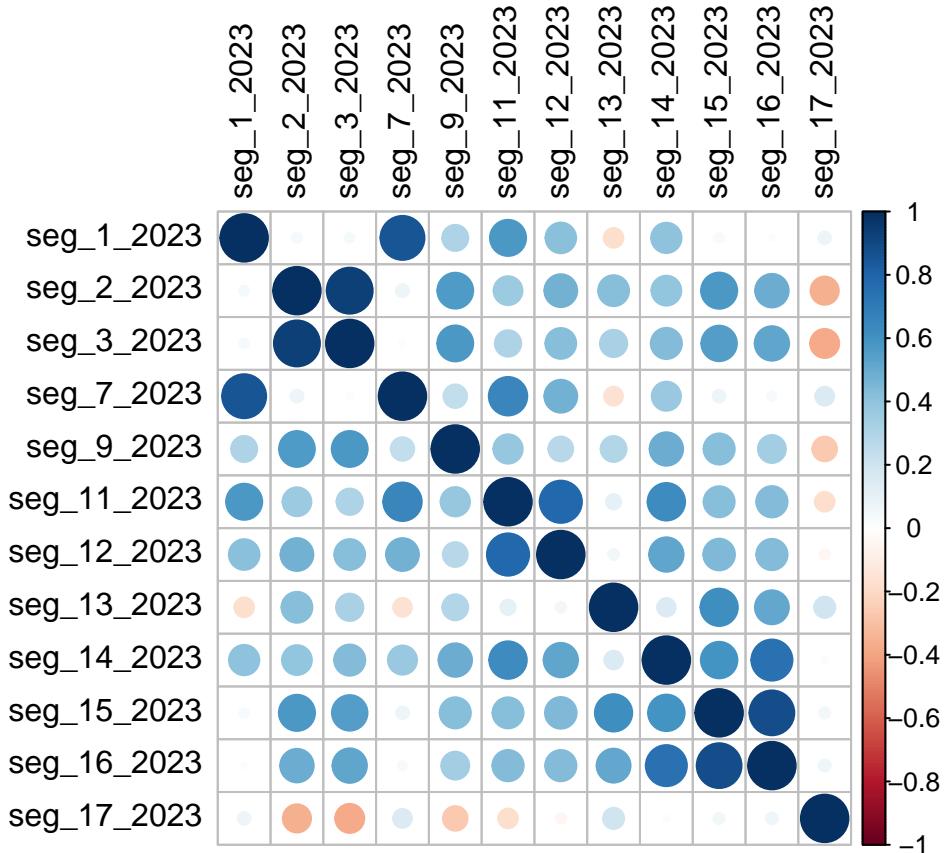


6.1 Técnica de Análisis de Componentes Principales.

Se aplicará un análisis de componentes principales para reducir la dimensionalidad de los datos, esto para una mejor interpretación de qué es lo que explica al fenómeno de interés, es decir, se busca una relación entre las variables estudiadas de seguridad pública.

Para que un PCA tenga sentido, es necesario que exista una alta correlación entre los predictores, por lo que se explorará a continuación.

```
#seleccionamos los datos numericos
cor_seg1 <- cor(est_seg1)
#Gráfica de correlaciones entre variables
corrplot(cor_seg1, tl.cex = 1, tl.col = "black", mar = c(0, 0, 0, 0))
```



Derivado de la gráfica previa, se observa que en efecto sí existen correlaciones altas entre algunas variables. La correlación mide la fuerza y dirección de la interacción entre dos variables, suponiendo relación lineal. Por ejemplo, la variable *seg_1_2023* tiene correlación positiva fuerte con la variable *seg_7_2023*, le sigue la *seg_11_2023*, *seg_12_2023* y *seg_14_2023*. En general algunas variables tiene correlaciones fuertes positivas, pero también se observan ciertas correlaciones negativas entre algunas variables.

Finalmente, se calcula el determinante, se busca que el determinante se cercano a cero, pues implicaría un alta multicolinealidad entre variables.

```
det_seg1<-det(cor_seg1) #Calculamos determinante, como es cercano a cero, quiere decir que hay correlacion
det_seg1 #0.00000606635
```

```
## [1] 0.00000606635
```

Como el determinante es 0.00000606635, cercano a cero, quiere decir que hay correlación. Derivado de lo anterior, el análisis de componentes principales es factible.

```
### PCA

seg.pc1 <- princomp(est_seg1)
summary(seg.pc1)
```

```

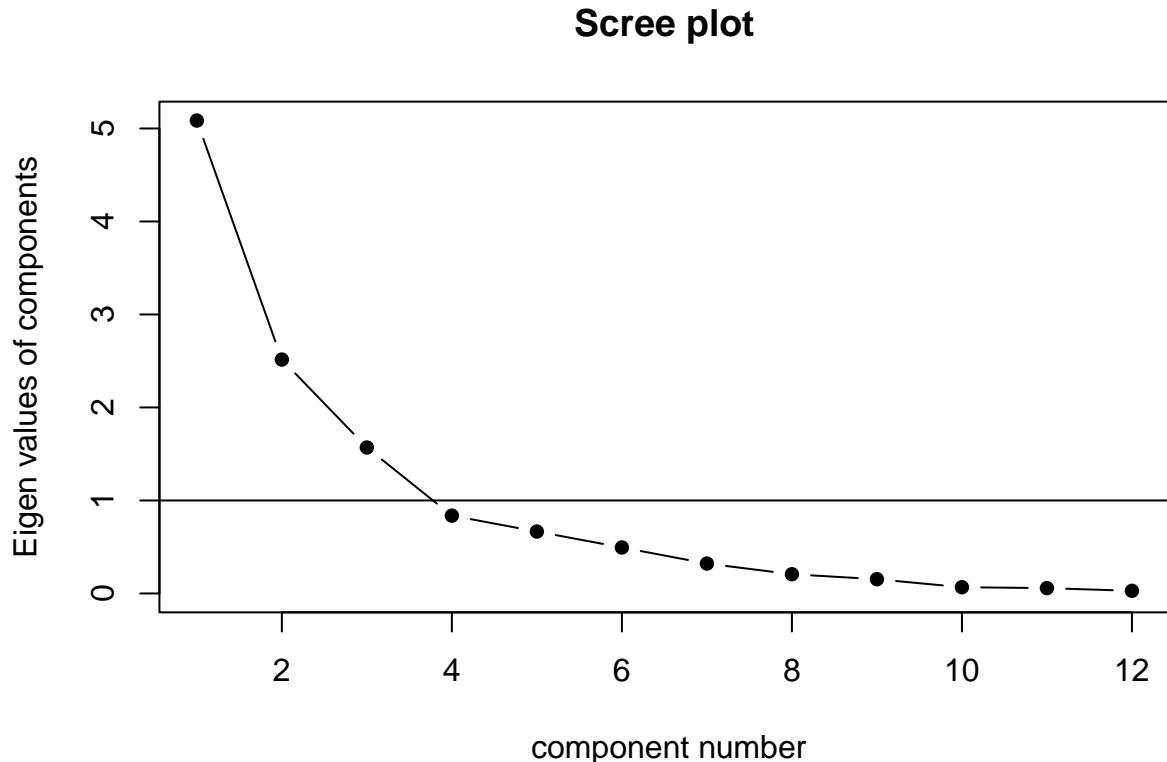
## Importance of components:
##                               Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
## Standard deviation      2.2195494 1.5609258 1.2329976 0.90037421 0.8030692
## Proportion of Variance 0.4237763 0.2095905 0.1307770 0.06973537 0.0554770
## Cumulative Proportion  0.4237763 0.6333668 0.7641438 0.83387920 0.8893562
##                               Comp.6   Comp.7   Comp.8   Comp.9   Comp.10
## Standard deviation      0.6919656 0.55787879 0.44756247 0.38498857 0.254762163
## Proportion of Variance 0.0411885 0.02677236 0.01723115 0.01274978 0.005583119
## Cumulative Proportion  0.9305447 0.95731707 0.97454822 0.98729800 0.992881119
##                               Comp.11  Comp.12
## Standard deviation      0.23558307 0.165098793
## Proportion of Variance 0.00477414 0.002344741
## Cumulative Proportion  0.99765526 1.000000000

```

Uno de los criterios para escoger el número adecuado de componentes principales, es cuidar que la varianza acumulada que expliquen, sea de por lo menos 70%, y de acuerdo con la salida del modelo, se necesitan tres componentes, en el ejemplo se acumula el 76.4% de la varianza del modelo.

Otros criterios son el del codo, que a partir del gráfico de sedimentación, se tiene que ubicar el cambio de trayectoria más abrupta, y en ese punto, será el número de componentes adecuado, otro es el de Kaiser, que indica conservar aquellas componentes cuyos valores propios superen a la unidad.

```
scree(est_seg1, pc=TRUE, fa=FALSE)
```



Según el gráfico anterior, el criterio del codo sugiere 4 componentes, mientras que la regla de Kaiser sugiere tres. En este caso, los criterios probados indican un número diferente de componentes, sin embargo, se va a

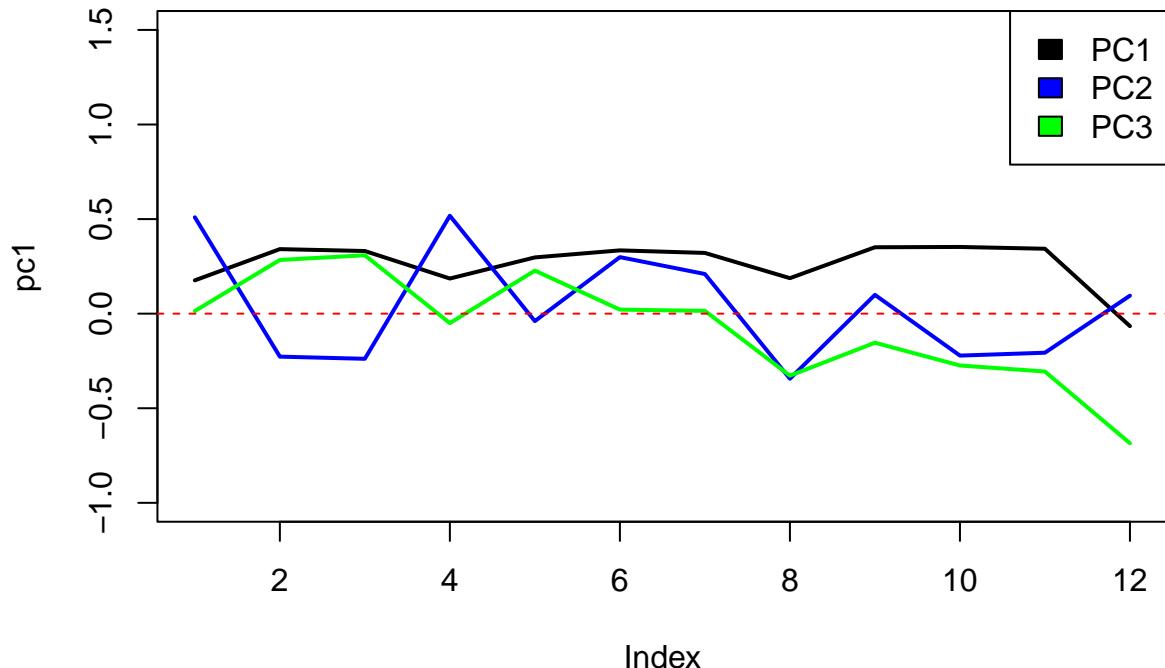
priorizar la parsimonia del modelo, es decir, se van a elegir tres componentes pues el porcentaje de varianza acumulada es muy aceptable.

```
# Se eligen 3 componentes
#Se cachan los vectores propios

pc1 <- seg.pc1$loadings[,1]
pc2 <- seg.pc1$loadings[,2]
pc3 <- seg.pc1$loadings[,3]

#Se realizan las gráficas

plot(pc1, ylim = c(-1, 1.5), type = "l", lwd=2)
lines(pc2, col="blue", lwd=2)
lines(pc3, col="green", lwd=2)
abline(h=0, lty=2, col="red")
legend(x="topright", legend = c("PC1", "PC2", "PC3"),
       fill = c("black", "blue", "green"))
```



En la gráfica se muestra el comportamiento de los primeros 3 componentes principales respecto y la relación con las variables originales.

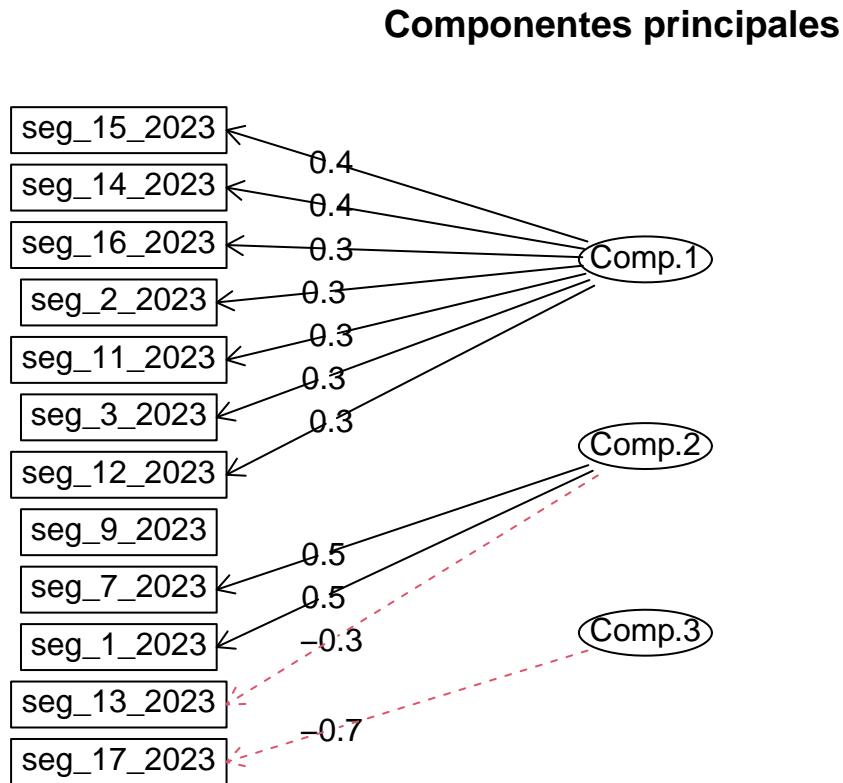
Según las cargas de los componentes, por lo menos visualmente hay algunas que destacan en alguna variable.

1. En el primer componente hay 7 variables que tienen pesos similares y son los más altos, rondando en valores de 0.3. De hecho, las cargas del componente 1 tienen un comportamiento constante.

2. Para el segundo componente hay 2 variables con cargas positivas elevadas las cuales son seg_7_2023, seg_1_2023 con cargas de 0.518 y 0.502 respectivamente, además, en tercer lugar quedó la variable seg_13_2023 con carga de -0.344 y ésta última hace contraste con los componentes antes mencionados.
3. En el componente 3, se observa que la variable con mayor carga la tiene la variable seg_17_2023 con un valor de -0.685, y luego, la carga de la variable seg_13_2023 con el siguiente valor mayor es -0.328. Por otro lado la variable seg_3_2023 registra una carga de 0.309, en general, se considera que el comportamiento de este componente es con una tendencia decreciente.

Para aterrizar lo anterior, se dibuja el diagrama donde se indica qué variables son las que explican a cada uno de los componentes.

```
fa.diagram(seg.pc1$loadings[,1:3], main = "Componentes principales")
```



Según el diagrama anterior, el componente uno está definido por las variables: seg_15_2023, seg_14_2023, seg_16_2023, seg_2_2023, seg_11_2023, seg_3_2023 y seg_12_2023, todas en sentido positivo y con una carga muy similar, este componente puede estar explicando el desempeño institucional, percepción de inseguridad, cambio de rutinas y Nivel de criminalidad (incidencia delictiva). El componente dos se define por las variables seg_7_2023, seg_1_2023 y seg_13_2023, éste último con carga negativa, que se refiere a temas con el crimen organizado, Nivel de criminalidad (Tasa de homicidios) y la tasa de policías. El componente tres se define por las variables seg_17_2023 con carga negativa, relacionados con el desempeño institucional en particular con la variación porcentual del índice de conflictividad. Por otro lado, la variable seg_9_2023 Promedio de costos del delito en unidades económicas, por entidad federativa no fue relacionada con ningún componente.

Ahora, se pueden graficar los scores (a nivel dato) por pares de los componentes principales. Por temas de interpretación, solo se graficarán los scores de los dos primeros componentes principales.

6.2 Gráficas de los scores

Ahora se graficaran los tres scores por pares (comp1 vs comp2), (comp1 vs comp3) y (comp2 vs comp3).

Primero se extraen los puntajes de las observaciones (scores).

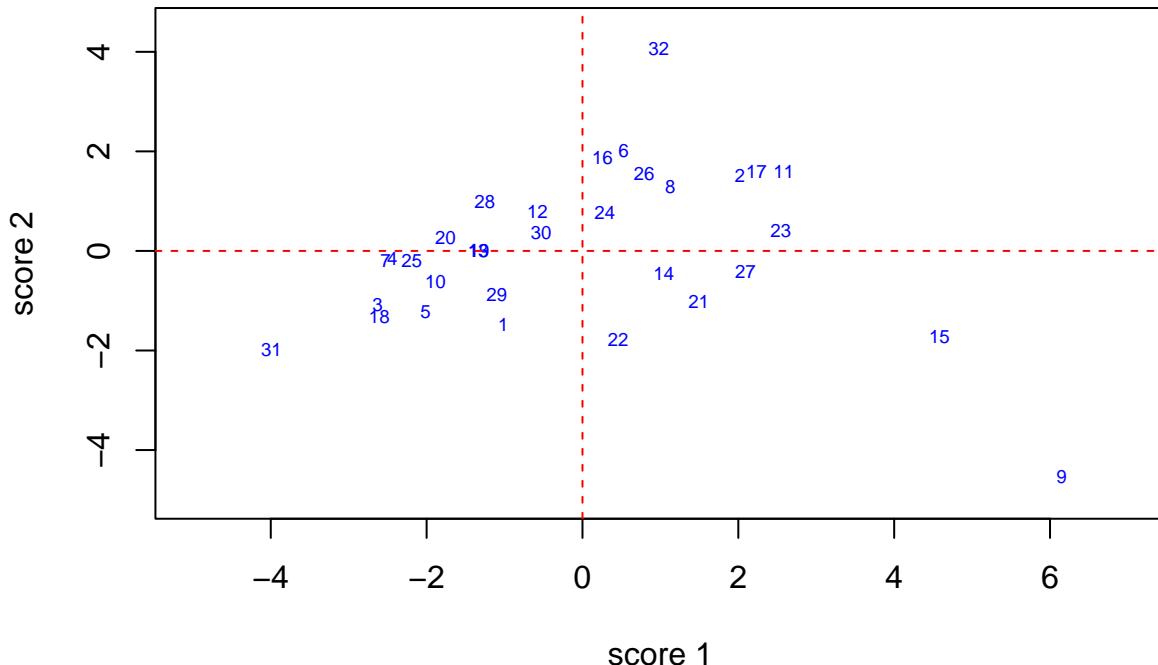
```
#Se extraen los scores
score1_v2<-seg.pc1$scores[,1]
score2_v2<-seg.pc1$scores[,2]
score3_v2<-seg.pc1$scores[,3]
```

Recordando que el componente 1 está relacionado con nivel de criminalidad, percepción de inseguridad y cambio de rutinas y con el desempeño institucional, principalmente relacionado con la desconfianza de la autoridades, el componente 2 con la tasa de homicidio, índice de riesgo por crimen organizado y la tasa de policías por cada cien mil hab. y finalmente, el factor 3 está relacionado con la variable: Variación porcentual del índice de conflictividad en la entidad federativa.

Se realiza el gráfico score 1 vs score 2

```
#Se realiza el gráfico
plot(score1_v2,score2_v2,main="Biplot score 1 vs score 2",xlim=c(-5,7),ylim=c(-5,4.5),
  xlab="score 1",ylab="score 2",type="n",lwd=2)
text(score1_v2,score2_v2,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

Biplot score 1 vs score 2



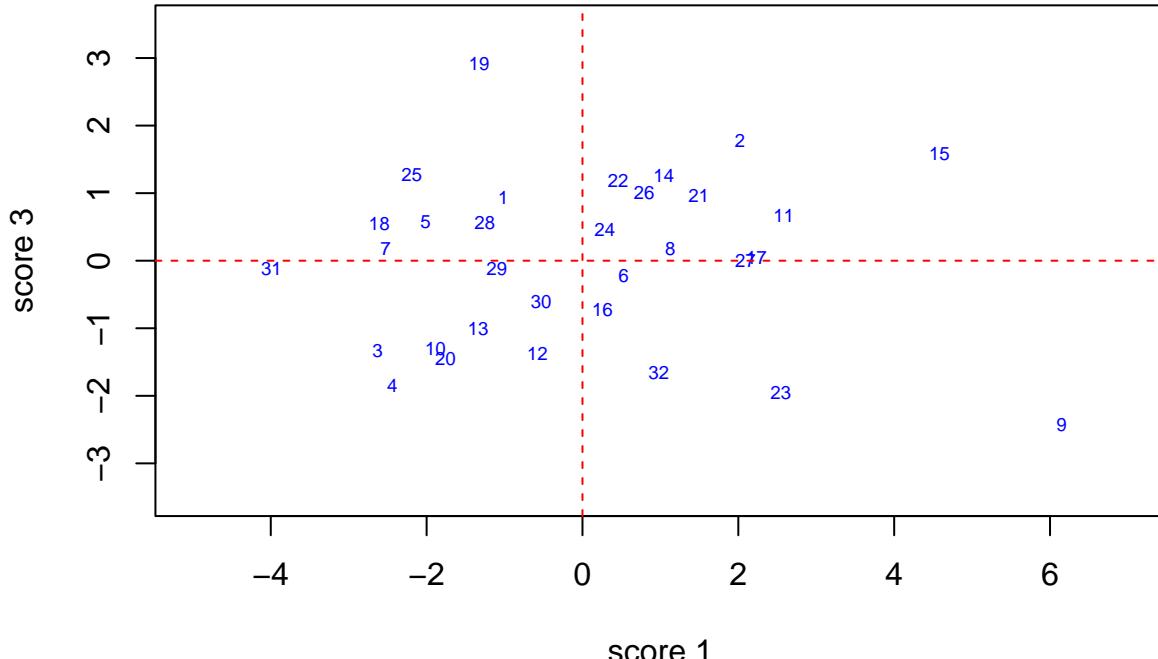
En este ejemplo, el componente 1 es un poco difícil de explicar por todas las variables asociadas a él. La gráfica muestra que en el cuadrante I, están las entidades con valores bajos respecto al componente 1, pero alrededor de la media nacional con respecto al componente 2. En el extremo superior del cuadrante II se encuentra la entidad de 32 (Zacatecas), en donde reporta valores un poco mayores a media nacional en el componente 1, sin embargo, en el componente 2 es la entidad que registra el valor más alto. La entidad 23 (Quinta Roo), registra un valor alto en el componente 1 y un valor por arriba de la media nacional en el componente 2. En el extremo inferior del cuadrante III, cayó la entidad 31 (Yucatán), que es la que registra un valor pequeño en el componente 1 y también registra un valor pequeño en el componente 2. Y finalmente, para el cuadrante IV, en el extremo inferior derecho, se encuentra la entidad 9 con valores altos en el componente 1 y muy bajos en el componente 2, se puede interpretar que es la entidad con mayores tasas de incidencia, mayor desconfianza en las autoridades, en cambio, tiene una tasa de homicidios baja, una índice de riesgo por crimen organizado también bajo y la tasa de policías alta (está en sentido contrario). Un comportamiento similar registra la entidad 15 (EDOMEX).

Se realiza el gráfico score 1 vs score 3

```
#Se realiza el gráfico

plot(score1_v2,score3_v2,main="Biplot score 1 vs score 3",xlim=c(-5,7),ylim=c(-3.5,3.5),
  xlab="score 1",ylab="score 3",type="n",lwd=2)
text(score1_v2,score3_v2,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

Biplot score 1 vs score 3



En el gráfico se muestra la relación de que tienen los componentes 1 y 3 con base en las entidades.

En el extremo superior del cuadrante I, se encuentra la entidad 19 (Nuevo León) que reporta valores por debajo del nivel nacional del componente 1, además es una entidad que registra una variación porcentual

del índice de conflictividad en la entidad federativa casi nula. Con respecto al cuadrante II, en el extremo se encuentra la entidad 15 (EDOMEX), con un valor alto en el componente 1 y un valor similar al de Nuevo León del componente 3.

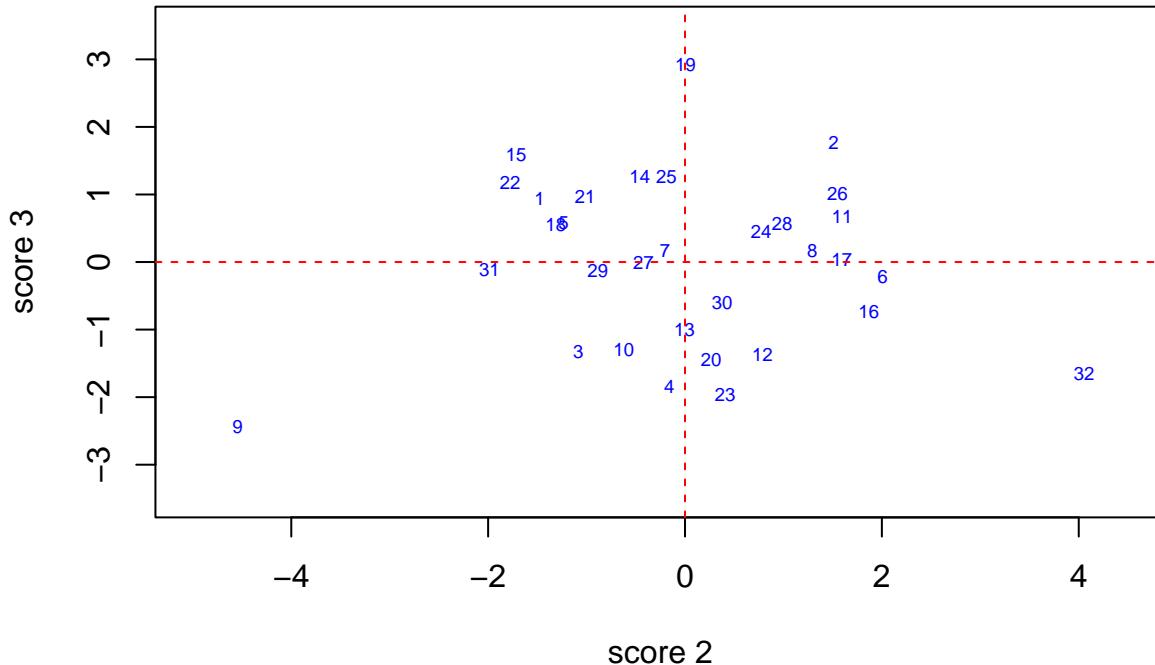
En el cuadrante III, se encuentra la entidad 31 (Yucatán), en donde registra un valor muy pequeño en el componente 1 y un valor parecido al valor nacional en el componente 3, es decir, si hay variación porcentual del índice de conflictividad. En ese mismo cuadrante está la entidad 4 (Campeche), en donde registra un valor por debajo del nacional en el componente 1, sin embargo, registra una variación porcentual del índice de conflictividad considerable. En el cuadrante IV, se encuentra la entidad 32 (Zacatecas), en donde registra un valor por arriba del nacional para el componente 1 y una variación porcentual del índice de conflictividad (componente 3) considerable. Finalmente, la entidad 23 (Quintana Roo) tiene un comportamiento similar al de Zacatecas con respecto al componente 3.

Se realiza el gráfico score 2 vs score 3

```
#Se realiza el gráfico
```

```
plot(score2_v2,score3_v2,main="Biplot score 2 vs score 3",xlim=c(-5,4.5),ylim=c(-3.5,3.5),
  xlab="score 2",ylab="score 3",type="n",lwd=2)
text(score2_v2,score3_v2,
  labels=abbreviate(row.names(seg),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty = 2, col = "red")
```

Biplot score 2 vs score 3



En gráfica anterior se muestra la relación del componente 2 y 3.

Por ejemplo, la entidad (9) CDMX, registra una de los valores más bajos en el componente 2 y y también, una porcentual del índice de conflictividad por arriba de la del valor nacional. Luego, en la parte superior del eje Y, está la entidad 19 (Nuevo León), es decir, el componente 2 está por el que se reporta a nivel nacional,

sin embargo, el componente 3 que tiene que ver con la variación porcentual del índice de conflictividad es una de las entidades en el top con variaciones casi nulas. Finalmente, en el cuadrante IV, en el extremo inferior derecho se encuentra la entidad 32 (Zacatecas), en donde se registra un valor alto en el componente 2 y también una de las entidades que registran mayor variación porcentual del índice de conflictividad referente al componente 3.

6.3 Cálculo del subíndice de seguridad de los scores por componentes principales (versión 2).

Se obtienen los scores para crear un data frame para hacer la escala del 0-100.

Opción 1. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 1*, solo que en este caso se contemplan 3 scores.

```
bd_score4<-cbind(score1_v2,score2_v2,score3_v2)
#Se convierte en data frame
bd_score4<- as.data.frame(bd_score4)
bd_score4$media_seg<- apply(bd_score4,1,mean)
val_min<- min(bd_score4$media_seg) # -2.032756
bd_score4$dif<-(bd_score4$media_seg-val_min)
val_max<- max(bd_score4$dif)
bd_score4$ind_seg<-(bd_score4$dif/val_max)*100
#bd_score4 #Es el data frame construido, la última columna "ind_seg" es el índice de seguridad.
#Finalmente se agrega la variable entidad a la base final
bd_score4$ID<-c("01","02","03","04","05","06","07","08","09",
                 "10","11","12","13","14","15","16","17","18",
                 "19","20","21","22","23","24","25","26","27",
                 "28","29","30","31","32")
```

Opción 2. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 2*, solo que en este caso se contemplan 3 scores.

```
bd_score5<-cbind(score1_v2,score2_v2,score3_v2)
#Se convierte en data frame
bd_score5<- as.data.frame(bd_score5)
val_min1<- min(bd_score5$score1_v2)
val_min2<- min(bd_score5$score2_v2)
val_min3<- min(bd_score5$score3_v2)
bd_score5$dif1<-(bd_score5$score1_v2-val_min1)
bd_score5$dif2<-(bd_score5$score2_v2-val_min2)
bd_score5$dif3<-(bd_score5$score3_v2-val_min3)
val_max1<- max(bd_score5$dif1)
val_max2<- max(bd_score5$dif2)
val_max3<- max(bd_score5$dif3)
bd_score5$ind_seg1<-(bd_score5$dif1/val_max1)*100
bd_score5$ind_seg2<-(bd_score5$dif2/val_max2)*100
bd_score5$ind_seg3<-(bd_score5$dif3/val_max3)*100
bd_score5$ind_seg_f<-apply(bd_score5[,c(7:9)],1,mean)
#Finalmente se agrega la variable entidad a la base final
bd_score5$ID<-c("01","02","03","04","05","06","07","08","09",
                 "10","11","12","13","14","15","16","17","18",
                 "19","20","21","22","23","24","25","26","27",
                 "28","29","30","31","32")
```

7. ANÁLISIS DE FACTORES (Versión 2)

7.1 Análisis de factibilidad

Antes de dar inicio con la implementación del modelo de factores como tal, es importante analizar la factibilidad de los datos, para ello se analiza la matriz de correlaciones, el cálculo del determinante de la matriz anterior y se generan las pruebas de Kaiser Meyer Olkin (KMO) y de esfericidad de Bartlet.

7.1.1 Prueba de Kaiser Meyer Olkin

La prueba de Kaiser-Meyer-Olkin (KMO), es una medida de la idoneidad de los datos para el análisis factorial. Mide la adecuación del muestreo para cada variable en el modelo y para el modelo completo. La estadística es una medida de la proporción de varianza entre variables que podrían ser varianza común.

La prueba de KMO devuelve valores entre 0 y 1. Una regla general para interpretar la estadística es la siguiente:

- Los valores de KMO entre 0.8 y 1 indican que el muestreo es perfecto para realizar un análisis de factores.
- Los valores entre 0.6 y 0.7 señalan que es viable realizar un análisis de factores.
- Los valores de KMO entre 0.4 y 0.5 indican que es aceptable realizar un análisis de factores, no obstante, se sugiere verificar el estado de las variables.
- Los valores de KMO menores a 0.4 significan que existen grandes correlaciones parciales en comparación con la suma de las correlaciones. En otras palabras, existen correlaciones generalizadas que son un gran problema para el análisis factorial, por lo que no se recomienda efectuar un análisis factorial.

El resultado de la prueba se muestra a continuación:

```
# Prueba de KMO
KMO(r = cor(est_seg1))

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(est_seg1))
## Overall MSA =  0.59
## MSA for each item =
##   seg_1_2023  seg_2_2023  seg_3_2023  seg_7_2023  seg_9_2023 seg_11_2023
##   0.54        0.62        0.68        0.50        0.73        0.62
##   seg_12_2023 seg_13_2023 seg_14_2023 seg_15_2023 seg_16_2023 seg_17_2023
##   0.64        0.41        0.59        0.73        0.59        0.23
```

De acuerdo con la salida anterior, para el modelo completo se tiene un valor de 0.59, lo que implica que la tabla de datos sea aceptable para efectuar un análisis de factores (AF), no obstante, se sugiere verificar el estado de las variables. Por otro lado, en el caso de cada variable, por ejemplo, las variables *seg_2_2023*, *seg_3_2023*, *seg_9_2023*, *seg_11_2023*, *seg_12_2023* y *seg_15_2023* tienen valores entre 0.6-0.7, lo que implica que para el AF estas variables son viables; las variables *seg_1_2023*, *seg_7_2023*, *seg_13_2023*, *seg_14_2023* y *seg_16_2023* tienen valores entre 0.4-0.5, por lo que para el AF estas variables son aceptables para realizar un análisis de factores, finalmente, la variable *seg_17_2023* tiene un valor de 0.23, por lo que no se recomienda efectuar un análisis factorial.

7.1.2 Prueba de esfericidad de Bartlett

En la prueba de esfericidad de Bartlett se analiza si tiene sentido efectuar un análisis de factores a partir de las siguientes hipótesis: $H_0 : R = I$ vs $H_a : R \neq I$, donde R es la matriz de correlaciones e I es una matriz identidad del mismo tamaño que R . En caso de no rechazar H_0 , no tendría sentido realizar el análisis de factores. La regla establecida es rechazar H_0 si el valor p es menor a 0.05, el resultado de la prueba es el siguiente.

```
# Prueba
cortest.bartlett(est_seg1)

## R was not square, finding R from data

## $chisq
## [1] 314.3337
##
## $p.value
## [1] 0.000000000000000000000000000000000000000000005066208
##
## $df
## [1] 66
```

Derivado de la anterior, ya que el valor p es casi 0 y éste es menor que 0.05 ($0 < 0.05$) se rechaza H_0 , es decir, tiene sentido continuar con el análisis de factores.

En conclusión, el análisis de factibilidad determina que es viable realizar el análisis de factores con la tabla de datos seleccionada.

8 Elección del número de factores.

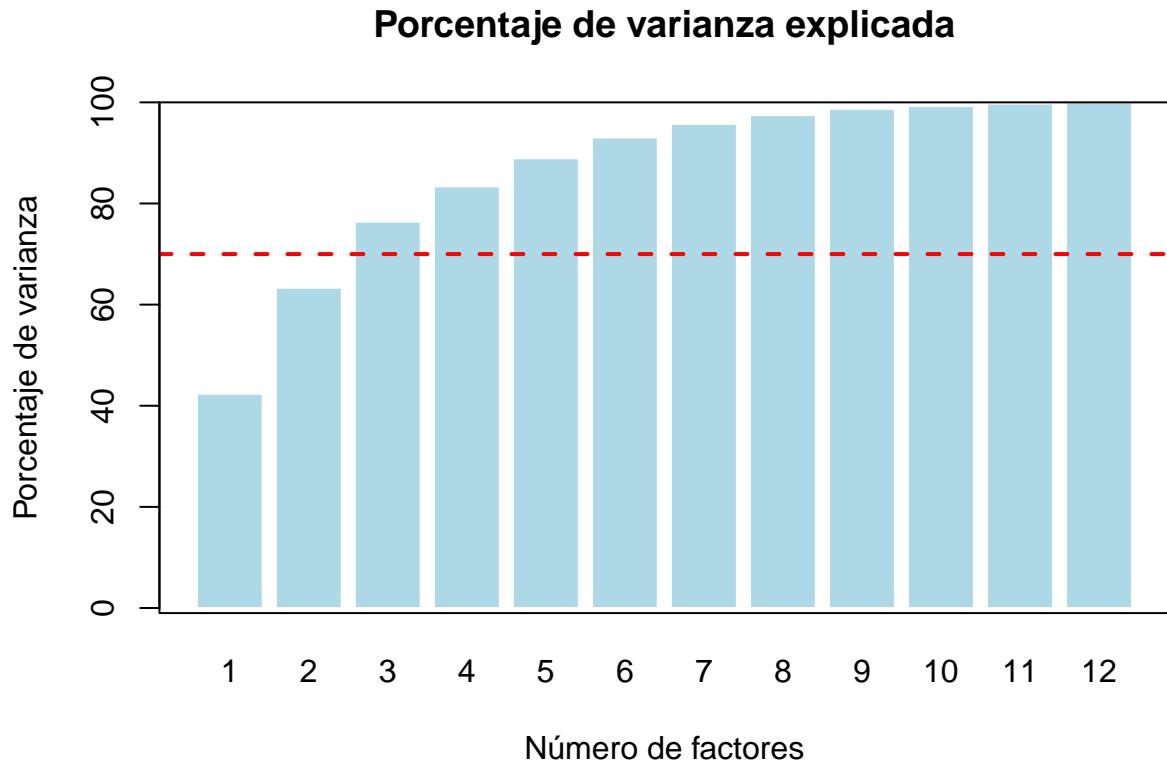
8.1 Porcentaje de varianza explicada

Para este criterio, por lo general se eligen el número de factores que explican de manera acumulada más del 70% de la varianza total. Para ello se obtienen los valores propios de la matriz de correlaciones y se divide entre el total de las variables de estudio. Posteriormente, se efectúa la suma acumulada para determinar el porcentaje de varianza explicada, según el número de factores seleccionados.

```
# Evitar la notación científica
options(scipen = 999)
# Se extraen los valores propios de la matriz de correlaciones
val_propios<-eigen(cor_seg1)$values
# Porcentaje de varianza explicada
acumulado<-cumsum(val_propios*100/ncol(est_seg1))
round(acumulado, 2)

## [1] 42.38 63.34 76.41 83.39 88.94 93.05 95.73 97.45 98.73 99.29
## [11] 99.77 100.00
```

```
# Gráfico de la varianza acumulada
barplot(acumulado, col = "lightblue", border = "white", names.arg = 1:12,
        main = "Porcentaje de varianza explicada",
        xlab = "Número de factores", ylab = "Porcentaje de varianza")
abline(h = 70, lty = 2, col = "red", lwd = 2)
box()
```



De acuerdo con las salidas previas, con tres factores se supera el 70% del total de la varianza explicada (línea roja punteada del gráfico), ya que se obtiene el 76.41%, con cuatro factores se llega al 83.39%, y así sucesivamente.

8.2 Criterio de Kaiser

Es uno de los criterios más utilizados por los paquetes estadísticos para determinar el número adecuado de factores. Se aplica cuando se trabaja con la matriz de correlaciones muestrales R . Consiste en elegir el número de valores propios de R que sean mayores a 1. Esta regla busca que cualquier factor retenido contenga al menos la varianza de una de las variables utilizadas en el análisis.

```
# Se imprimen los valores propios a dos decimales
round(val_propios, 2)
```

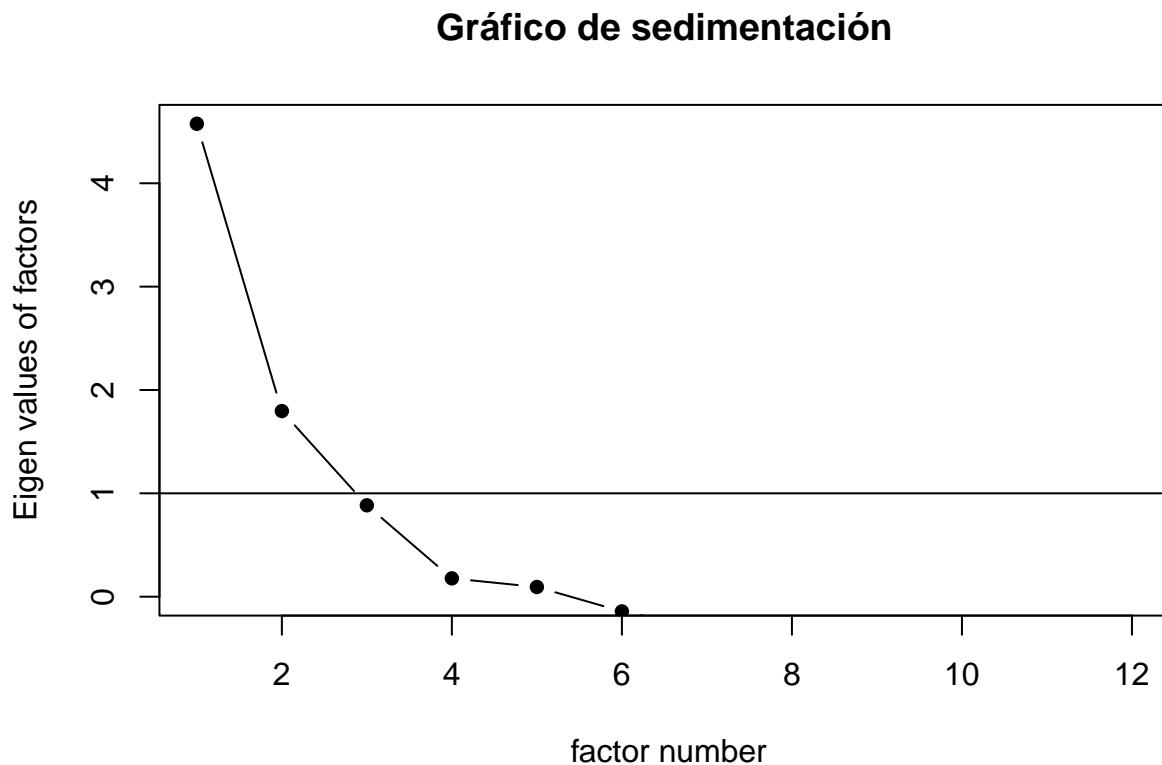
```
## [1] 5.09 2.52 1.57 0.84 0.67 0.49 0.32 0.21 0.15 0.07 0.06 0.03
```

Con base a este criterio y a la salida anterior, se deberían elegir 3 factores, ya que sus valores propios son de 5.09, 2.52, 1.57 respectivamente.

8.3 Gráfico de sedimentación

Este criterio complementa al anterior y se basa también en el análisis de la magnitud de los valores propios, pero a partir de la tendencia que se observa en el gráfico. Se procura seleccionar un grupo reducido de factores que tengan valores propios significativamente superiores a los demás, para lo cual se identifica el punto de inflexión en la curva del gráfico de sedimentación (también referido como el codo por su semejanza con un brazo) a partir del cual la curva se transforma a una línea “plana” o relativamente recta. El gráfico para este ejercicio se presenta a continuación.

```
# Gráfico de sedimentación
scree(est_seg1, main = "Gráfico de sedimentación", pc=FALSE, fa=TRUE)
```



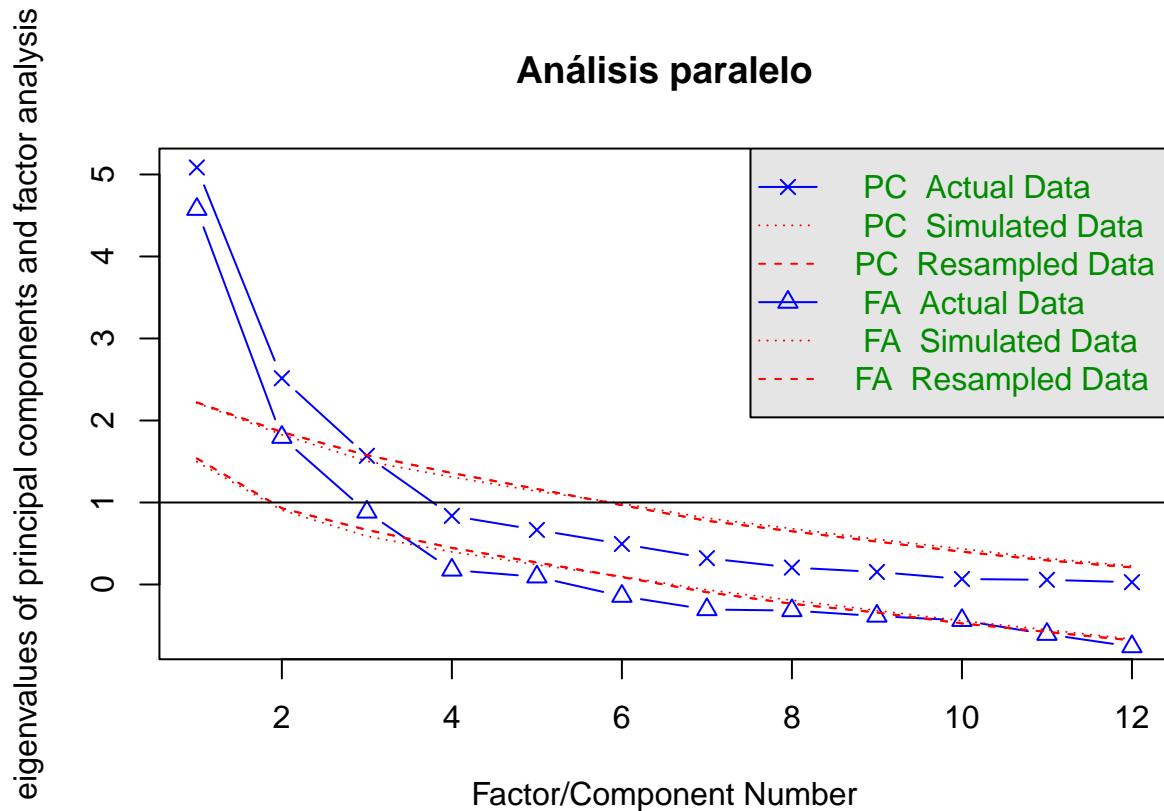
En la gráfica previa se aprecia un punto de inflexión después de 2 factores para AF.

8.4 Análisis paralelo

El análisis paralelo suele complementar los anteriores cuando el número de variables iniciales y factores resultantes es elevado. El procedimiento se basa en el principio de que los factores a extraer deben dar cuenta de más varianza que la esperada de manera aleatoria. El procedimiento reordena las observaciones de manera aleatoria entre cada variable y los valores propios son recalculados a partir de esta nueva base de datos aleatoriamente ordenada. Los factores con valores propios mayores a los valores aleatorios son retenidos para interpretación.

Se comparan gráficamente las dos líneas del gráfico de sedimentación, una para el análisis de los datos reales y otra para el análisis de los datos aleatorios. El número de factores se determina con la intersección de ambas gráficas. El resultado de este criterio se muestra en la próxima gráfica.

```
# Análisis paralelo
fa.parallel(est_seg1, main = "Análisis paralelo")
```



```
## Parallel analysis suggests that the number of factors = 3 and the number of components = 2
```

De acuerdo con el análisis paralelo, sugiere conservar 3 factores en el caso de un AF y 2 componentes para un ACP. Hasta el momento los análisis sugieren considerar 3 factores. Hasta el momento, las pruebas sugieren 3 factores.

9. Análisis factorial confirmatorio

Para el análisis factorial confirmatorio se obtiene la solución por máxima verosimilitud de \mathbf{L} (cargas) y Ψ (varianzas específicas) para 3 factores, considerando una rotación *varimax* para facilitar la interpretación de los factores.

9.1 Porcentaje de varianza explicada

La rotación de las cargas busca mejorar la interpretación de los factores. Al rotarlos, la suma de los valores propios no se altera, pero podrían cambiar los valores propios, el porcentaje de varianza explicada y los valores de carga de cada factor. Por lo anterior, el porcentaje de varianza explicada por cada uno de los modelos propuestos se muestra a continuación.

```

# Cargas estimadas rotadas para 3 factores por máxima verosimilitud
af.mv3<-factanal(est_seg1, factors = 3, rotation = "varimax", n.obs = nrow(est_seg1))
# Se almacenan las cargas
cargas_mv3<-af.mv3$loadings
vx3<-colSums(cargas_mv3^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx3, `Proporción de la varianza` = vx3/nrow(cargas_mv3),
`Varianza acumulada` = cumsum(vx3/nrow(cargas_mv3)))

```

```

##                               Factor1   Factor2   Factor3
## SS loadings            3.2201784 2.9773224 2.0860178
## Proporción de la varianza 0.2683482 0.2481102 0.1738348
## Varianza acumulada      0.2683482 0.5164584 0.6902932

```

De acuerdo al resultado previo, si se consideran 3 factores el porcentaje de la varianza acumulada es del 69.02%.

9.2 Estimación de cargas mediante máxima verosimilitud

Se realizaron las estimaciones de las cargas rotadas, las **comunalidades**, que es la parte de la variación de la variable que es compartida con otras variables, y las **varianzas únicas o específicas**, que es la parte de la variación de la variable que es propia de esa variable, para el modelo de 3 factores.

```

# Modelo con 3 factores
# Comunalidades
comu_mv3<-diag(cargas_mv3[, 1:3] %*% t(cargas_mv3[, 1:3]))
# Unicidades o varianza específica
espe_mv3<-1 - comu_mv3

# Se unen los resultados para el modelo de 5 factores
res_mv3<-data.frame(CARGA3 = cargas_mv3[, 1:3], COM3 = comu_mv3, ESP3 = espe_mv3)
# Se imprime el resultado de las cargas rotadas
kbl(res_mv3, longtable = T, booktabs = T, digits = 3,
  caption = "Modelo con factores por máxima verosimilitud",
  col.names = c("F_1", "F_2", "F_3", "Comunalidad", "Varianza única"))

```

Table 6: Modelo con factores por máxima verosimilitud

	F_1	F_2	F_3	Comunalidad	Varianza única
seg_1_2023	-0.107	0.906	-0.040	0.834	0.166
seg_2_2023	0.420	0.141	0.894	0.995	0.005
seg_3_2023	0.457	0.105	0.816	0.885	0.115
seg_7_2023	-0.058	0.923	-0.036	0.856	0.144
seg_9_2023	0.281	0.333	0.455	0.397	0.603
seg_11_2023	0.351	0.733	0.132	0.679	0.321
seg_12_2023	0.357	0.571	0.275	0.529	0.471
seg_13_2023	0.514	-0.145	0.250	0.348	0.652
seg_14_2023	0.697	0.496	0.037	0.733	0.267
seg_15_2023	0.868	0.133	0.217	0.818	0.182
seg_16_2023	0.989	0.104	0.075	0.995	0.005

seg_17_2023	0.092	0.063	-0.450	0.215	0.785
-------------	-------	-------	--------	-------	-------

En la salida anterior se puede apreciar que, la varianza única de *seg_17_2023* (Variación porcentual del índice de conflictividad en la entidad federativa) es la más alta con 0.785, lo que implica que los factores no están explicando de buena forma la variabilidad de esta variable, asimismo, *seg_13_2023* (Tasa de policías por cada cien mil habitantes), *seg_9_2023* (Promedio de costos del delito en unidades económicas, por entidad federativa) cuentan con varianzas específicas de 0.652 y 0.603 de forma respectiva.

9.3 Prueba de hipótesis para el número de factores comunes

El supuesto de normalidad permite realizar pruebas sobre la suficiencia del modelo. Es decir, se puede probar si el número de factores m es suficiente para explicar la covarianza observada. Suponiendo que el modelo de m factores comunes tiene buen ajuste, entonces $\Sigma = \mathbf{LL}' + \Psi$, y probar el ajuste del modelo de m factores comunes es equivalente a probar: $H_0: \Sigma = \mathbf{LL}' + \Psi$ (El número de factores en el modelo, es suficiente para capturar la dimensionalidad completa del conjunto de datos).

$H_1: \Sigma \neq \mathbf{LL}' + \Psi$ (El número de factores en el modelo, no es suficiente para capturar la dimensionalidad completa del conjunto de datos).

La regla de decisión es rechazar H_0 si el valor p es menor que 0.05. Tal resultado indica que el número de factores es demasiado pequeño. Por el contrario, no rechazamos H_0 si el valor p excede 0.05. Tal resultado indica que es probable que haya suficientes (o mas que suficientes) factores que capturen la dimensionalidad completa del conjunto de datos.

El resultado de la prueba para el modelo de 3 factores es el siguiente.

Prueba de hipótesis

```
# Valor p para el modelo de 3 factores
af.mv3$PVAL
```

```
## objective
## 0.00263155
```

Como el valor p es menor a 0.05 ($0.00263155 < 0.05$) se rechaza H_0 , es decir, la prueba sugiere que 3 factores en el modelo no son suficientes para capturar la dimensionalidad completa del conjunto de datos.

Nota: Se realizó la prueba con distintos números de factores (5 y 6) y resultó que la con ese número de factores se acepta la prueba. Ver en el anexo Por el momento, se trabajará con 3 factores, aunque la prueba de hipótesis no se haya aprobado, sin embargo, las demás pruebas para elegir el número de factores sugieren 3 factores.

9.4 Matriz de residuales

Se realiza la comparación de las matrices de correlación, en el cual se obtiene la diferencia entre la matriz de correlaciones real y la matriz reproducida mediante m factores, si el número de factores m es adecuado, la diferencia debe ser cercana a 0. Para ello se calculan las correlaciones reproducidas, que es la matriz de correlaciones de las variables originales que resultaría suponiendo que es correcto el número de factores retenidos. Asimismo, se calcula la matriz de correlaciones residuales, que es la matriz de las diferencias entre la matriz de correlaciones reproducida y la matriz de correlaciones reales. Si el número de factores m es adecuado, se esperaría que la matriz de correlaciones residuales tenga todas sus entradas cercanas a 0.

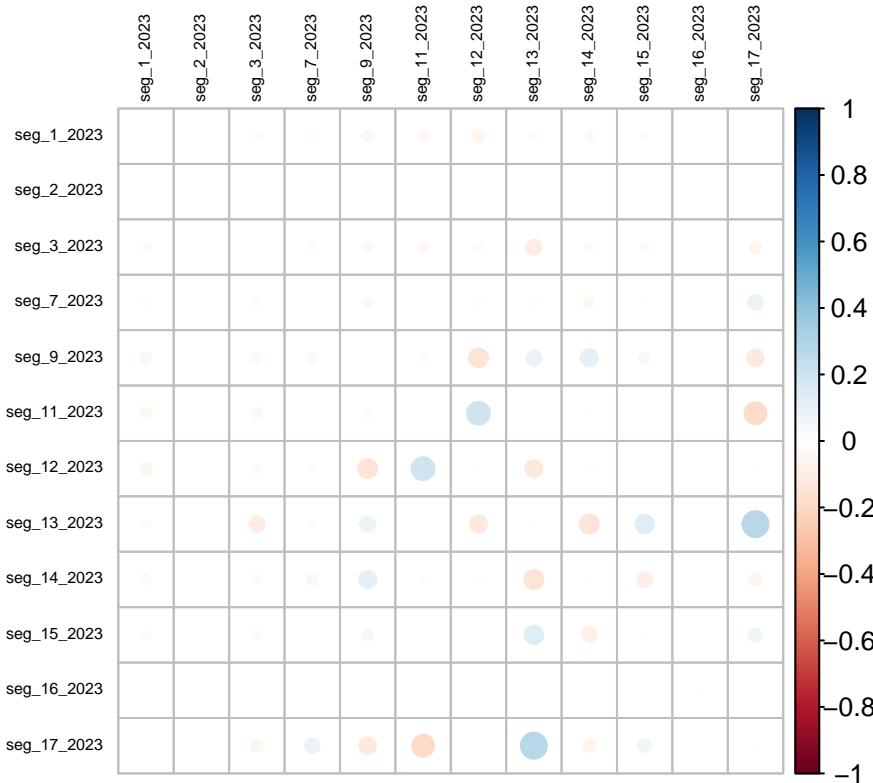
El resultado para el modelo de 3 factores es el siguiente.

```

# Matriz reproducida (3 factores)
mat_rep_mv3<-cargas_mv3[, 1:3] %*% t(cargas_mv3[, 1:3]) + diag(espe_mv3)
# Matriz de residuales (3 factores)
mat_res_mv3<-cor_seg1 - mat_rep_mv3
# Gráficos de correlaciones
corrplot(mat_res_mv3, order = "original", tl.col = "black", tl.cex = 0.5,
         mar = c(0, 0, 2, 0), title = "Matriz de residuales para 3 factores por MV")

```

Matriz de residuales para 3 factores por MV



Con base a la gráfica previa, se aprecia que la mayoría de las celdas son cercanas a 0, ya que no se observan círculos de color intenso. Los residuales más grandes que pueden distinguirse son para los pares $(seg_{13_2023}, seg_{17_2023})$, $(seg_{11_2023}, seg_{12_2023})$ y $(seg_{11_2023}, seg_{17_2023})$ cuyos valores son: 0.2723, 0.2089 y -0.1910, de forma respectiva. En la gráfica se observa que el residual mayor tiene un valor de 0.2723671, y de manera general se observan algunos pares de variables con valores mayores a cero.

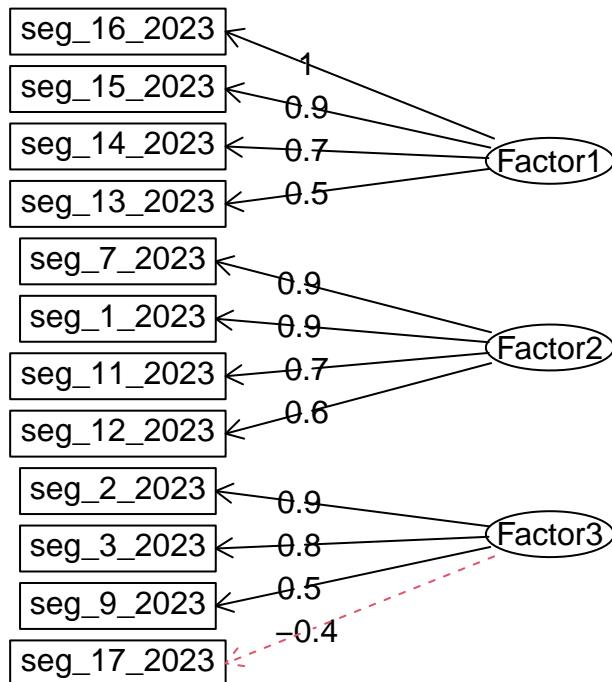
Hasta el momento, con el modelo de 3 factores se explica el 69% de la varianza, las demás pruebas para elegir el número de factores también sugieren tener 3 factores, además, la matriz de residuales arrojó valores cercanos de cero, sin embargo, la prueba de hipótesis para el número de factores comunes determinó que 3 factores no son suficientes, peso a esto, se seguirán considerando 3 factores por dos razones, la primera es por seguir el principio de parsimonia estadística y la segunda para comparar ese mismo número de factores respecto a componentes principales.

9.5 Interpretación de los factores

La interpretación de las cargas de los modelos, generalmente se explican en términos de las cargas principales en cada factor. Para ello se realizó el siguiente gráfico para el modelo de 3 factores.

```
# Gráfico para interpretaciones de los factores
fa.diagram(af.mv3$loadings, main = "Modelo de 3 factores", cex = 1, rsize = 0.7)
```

Modelo de 3 factores



- 1.- Para el diagrama con 3 factores, el **factor 1** se asocia con las variables: *seg_16_2023, seg_15_2023, seg_14_2023* y *seg_13_2023*, éstas variables son: Desconfianza en ministerios públicos, Desconfianza en jueces, Desconfianza en la policía estatal y municipal y Tasa de policías por cada cien mil habitantes.
- 2.- El **factor 2** está asociado con las variables: *seg_7_2023, seg_1_2023, seg_11_2023* y *seg_12_2023*, estas variables son: Índice de Riesgo por Crimen Organizado, Tasa de homicidios por cada cien mil habitantes, Porcentaje de personas que consideran insegura su entidad federativa y Porcentaje de personas que cambió algún hábito por temor a sufrir algún delito.
- 3.- El **factor 3** está asociado con las variables: *seg_2_2023, seg_3_2023, seg_9_2023* y *seg_17_2023*, estas variables son: Incidencia delictiva por cada cien mil habitantes, Prevalencia delictiva por cada cien mil habitantes, Promedio de costos del delito en unidades económicas, por entidad federativa y Variación porcentual del índice de conflictividad en la entidad federativa.

9.6 Estimación de factor scores

Para la estimación de los vectores de punajes factoriales (factor scores), se implementaron los métodos más consistentes y utilizados, los cuales son el de mínimos cuadrados ponderados (Bartlett) y el método por regresión (Thompson). El método por mínimos cuadrados ponderados supone que el vector de valores de los factores para cada observación es un parámetro a estimar. El método por regresión supone que estos vectores son variables aleatorias. El resultado de la estimación de los factor scores para el modelo de 3 factores se presenta a continuación.

Cálculo de factor scores mediante enfoque de mínimos cuadrados

```

# Cálculo de factor scores mediante enfoque de mínimos cuadrados
scores_min_v2<-factanal(est_seg1, factors = 3, scores = "Bartlett")$scores
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score1_v2<-data.frame(ID = paste0(ENT), scores_min_v2)
# Se ordenan los resultados por ID
res_score1_v2<-res_score1_v2[order(res_score1_v2$ID), ]

# Se imprime el resultado por mínimos cuadrados
kbl(res_score1_v2[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3"), row.names = FALSE,
  caption = "Factor Scores estimados por Mínimos cuadrados") %>%
add_header_above(c(" ", "Mínimos cuadrados" = 3)) %>%
kable_styling(latex_options = c("repeat_header"))

```

Table 7: Factor Scores estimados por Mínimos cuadrados

Entidad	Mínimos cuadrados		
	Score 1	Score 2	Score 3
01	-0.491	-1.078	0.769
02	-0.408	1.649	0.610
03	-0.221	-1.000	-0.577
04	0.079	-0.702	-1.236
05	0.024	-1.278	-0.249
06	-0.583	1.495	-0.072
07	-0.507	-0.922	-1.241
08	-0.107	1.188	-0.249
09	3.091	-0.733	1.348
10	0.776	-1.071	-1.688
11	-0.139	1.538	0.245
12	-0.305	0.499	-0.088
13	0.018	-0.482	-1.154
14	0.057	-0.013	0.676
15	2.114	-0.040	1.755
16	0.829	0.978	-1.938
17	0.529	1.297	0.235
18	-0.334	-1.201	-0.560
19	-2.269	-0.246	2.111
20	-0.134	-0.095	-1.243
21	0.670	-0.344	1.064
22	0.476	-1.016	0.626
23	1.441	0.741	-0.186
24	-0.260	0.414	0.321
25	-1.702	-0.409	1.275
26	-0.581	1.178	0.501
27	0.744	0.273	0.816
28	-0.896	-0.100	-0.587
29	-0.395	-0.922	0.544
30	0.161	-0.221	-1.208

Table 7: Factor Scores estimados por Mínimos cuadrados (*continued*)

Entidad	Mínimos cuadrados		
	Score 1	Score 2	Score 3
31	-1.095	-1.739	0.212
32	-0.584	2.365	-0.833

Cálculo de factor scores mediante enfoque de regresión

```
# Cálculo de factor scores mediante enfoque de regresión
scores_reg_v2<-factanal(est_seg1, factors = 3, scores = "regression")$scores
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score2_v2<-data.frame(ID = paste0(ENT), scores_reg_v2)
# Se ordenan los resultados por ID
res_score2_v2<-res_score2_v2[order(res_score2_v2$ID), ]

# Se imprime el resultado por regresión
kbl(res_score2_v2[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3"), row.names = FALSE,
  caption = "Factor Scores estimados por regresión") %>%
add_header_above(c(" ", "Regresión" = 3)) %>%
kable_styling(latex_options = c("repeat_header"))
```

Table 8: Factor Scores estimados por regresión

Entidad	Regresión		
	Score 1	Score 2	Score 3
01	-0.494	-1.005	0.754
02	-0.393	1.541	0.616
03	-0.228	-0.939	-0.580
04	0.071	-0.664	-1.230
05	0.015	-1.195	-0.256
06	-0.569	1.391	-0.062
07	-0.513	-0.873	-1.239
08	-0.099	1.106	-0.239
09	3.070	-0.654	1.339
10	0.760	-1.007	-1.679
11	-0.127	1.436	0.254
12	-0.300	0.463	-0.085
13	0.012	-0.458	-1.148
14	0.058	-0.007	0.670
15	2.105	-0.011	1.745
16	0.826	0.904	-1.913
17	0.535	1.216	0.244
18	-0.341	-1.128	-0.565
19	-2.253	-0.230	2.086

Table 8: Factor Scores estimados por regresión (*continued*)

Entidad	Regresión		
	Score 1	Score 2	Score 3
20	-0.136	-0.099	-1.233
21	0.666	-0.309	1.054
22	0.468	-0.941	0.615
23	1.437	0.700	-0.176
24	-0.254	0.387	0.321
25	-1.691	-0.384	1.258
26	-0.568	1.099	0.504
27	0.743	0.266	0.813
28	-0.892	-0.104	-0.585
29	-0.398	-0.860	0.532
30	0.156	-0.214	-1.199
31	-1.100	-1.629	0.196
32	-0.566	2.198	-0.810

Cálculo de las diferencias

```

# Cálculo de las diferencias
diferencia_v2<-scores_min_v2 - scores_reg_v2
# Identificadores de los registros
ENT<-formatC(datos$cve, width = 2, format = "d", flag = "0")
# Se crea el data frame con los resultados
res_score3_v2<-data.frame(ID = paste0(ENT), diferencia_v2)
# Se ordenan los resultados por ID
res_score3_v2<-res_score3_v2[order(res_score3_v2$ID), ]

# Se imprime el resultado de las diferencias
kbl(res_score3_v2[1:32, ], longtable = T, booktabs = T, escape = FALSE, digits = 3,
  col.names = c("Entidad", "Score 1", "Score 2", "Score 3"), row.names = FALSE,
  caption = "Diferencias") %>%
add_header_above(c(" ", "Diferencias" = 3)) %>%
kable_styling(latex_options = c("repeat_header"))

```

Table 9: Diferencias

Entidad	Diferencias		
	Score 1	Score 2	Score 3
01	0.003	-0.074	0.015
02	-0.015	0.108	-0.006
03	0.007	-0.061	0.003
04	0.008	-0.038	-0.006
05	0.009	-0.083	0.007
06	-0.014	0.104	-0.010
07	0.006	-0.049	-0.003
08	-0.008	0.081	-0.010
09	0.021	-0.080	0.010

Table 9: Diferencias (*continued*)

Entidad	Diferencias		
	Score 1	Score 2	Score 3
10	0.016	-0.064	-0.008
11	-0.012	0.101	-0.009
12	-0.005	0.036	-0.004
13	0.006	-0.024	-0.006
14	-0.001	-0.006	0.006
15	0.009	-0.030	0.010
16	0.003	0.073	-0.025
17	-0.006	0.081	-0.009
18	0.007	-0.073	0.005
19	-0.017	-0.016	0.025
20	0.003	0.004	-0.009
21	0.004	-0.035	0.010
22	0.008	-0.075	0.011
23	0.004	0.041	-0.010
24	-0.005	0.027	0.000
25	-0.010	-0.025	0.017
26	-0.013	0.079	-0.003
27	0.001	0.007	0.003
28	-0.003	0.004	-0.002
29	0.003	-0.062	0.012
30	0.005	-0.007	-0.009
31	0.005	-0.110	0.017
32	-0.018	0.167	-0.023

```
# Cálculo del promedio de las diferencias
```

```
round(apply(diferencia_v2, 2, mean), 3)
```

```
## Factor1 Factor2 Factor3
##      0      0      0
```

```
# Cálculo de la desviación estándar de las diferencias
```

```
round(apply(diferencia_v2, 2, sd), 3)
```

```
## Factor1 Factor2 Factor3
##  0.010  0.070  0.012
```

La tablas anteriores contienen las estimaciones de los factor scores para las 32 entidades. Asimismo, se muestra el promedio de las diferencias (prácticamente 0) y sus desviaciones estándar (cercaas a 0) entre los enfoques de mínimos cuadrados y regresión. Se puede apreciar que las diferencias son mínimas, así que las interpretaciones que se den sobre las entidades serán aplicables a ambos enfoques. Por lo anterior, se determinó trabajar con los factor scores estimados mediante regresión.

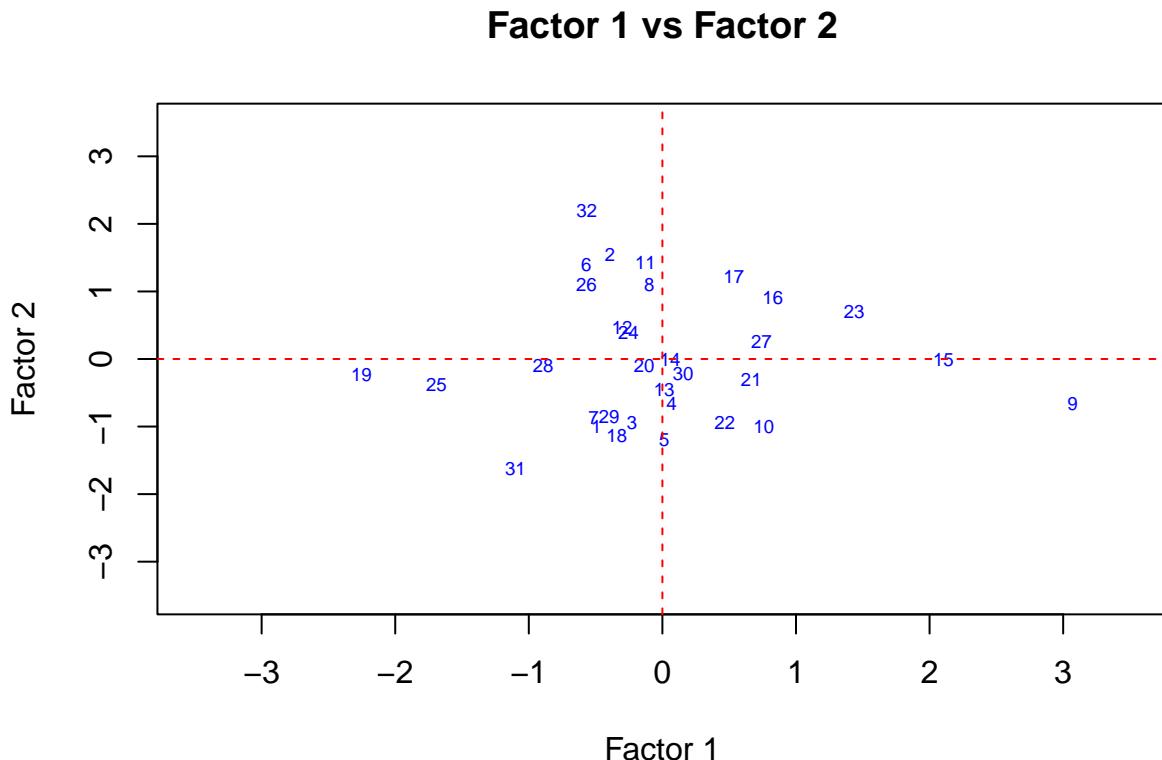
9.7 Interpretación de resultados

Se considera distinguir a las 32 entidades federativas. Para facilitar la interpretación de los resultados, se realizaron las diferentes parejas posibles de los factor scores, además de dividir la gráfica en 4 cuadrantes.

Asimismo, se identificaron los puntos extremos de cada cuadrante (representados con color rojo) para analizar su información. El resultado se muestra a continuación.

Se realiza el gráfico del Factor 1 vs Factor 2

```
plot(res_score2_v2[, 2], res_score2_v2[, 3], main="Factor 1 vs Factor 2", xlim=c(-3.5,3.5), ylim=c(-3.5,3.5),
  xlab="Factor 1", ylab="Factor 2", type="n", lwd=2)
text(res_score2_v2[, 2], res_score2_v2[, 3],
  labels=abbreviate(row.names(res_score2_v2), minlength=8), cex=0.6, lwd=2, col = "blue")
abline(v = 0, h = 0, lty =2, col ="red")
```



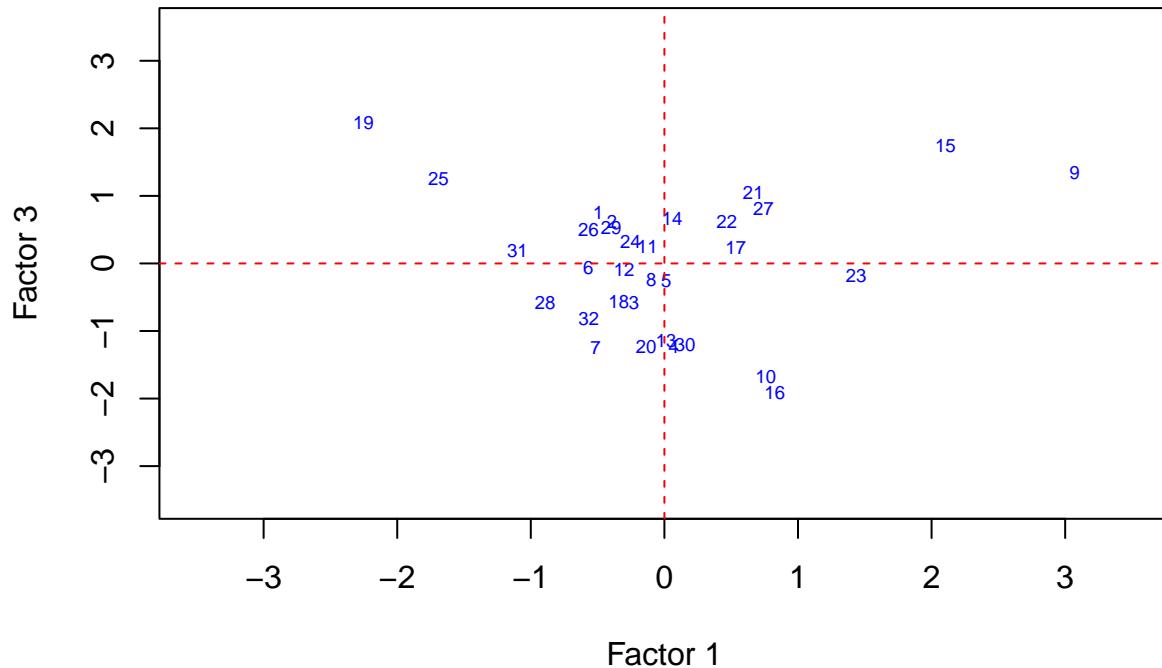
Como se observa en el gráfico, la entidad 9 (CDMX), registra el mayor valor en el factor 1, relacionado a la desconfianza en instituciones y la tasa de policías, en cambio, el factor 2 la CDMX reporta valores por debajo de la media nacional y se puede entender porque el factor 2 explica la percepción de inseguridad, índice de riesgo y la tasa de homicidios. La entidad 15 (EDOMEX), también registra un alto valor en el factor 1, pero en el factor 2 el valor registrado es parecido a la media nacional. Por otro lado, la entidad 19 (Nuevo León), reporta un valor bajó en el factor 1 y un valor parecido al de la media nacional en el factor 2. Por su parte, la entidad 32 (Zacatecas), registra un valor por debajo de la media nacional en el factor 1, sin embargo, es el que registra el valor más alto respecto al factor 2. Finalmente, otra entidad que está en un punto extremo es la entidad 31 (Yucatán), en donde está dentro del top de valores mínimos del factor 1 y también se registra el menor valor del factor 2.

Se realiza el gráfico del Factor 1 vs Factor 3

```
plot(res_score2_v2[, 2], res_score2_v2[, 4], main="Factor 1 vs Factor 3", xlim=c(-3.5,3.5), ylim=c(-3.5,3.5),
  xlab="Factor 1", ylab="Factor 3", type="n", lwd=2)
text(res_score2_v2[, 2], res_score2_v2[, 4],
```

```
labels=abbreviate(row.names(res_score2_v2),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty =2, col ="red")
```

Factor 1 vs Factor 3

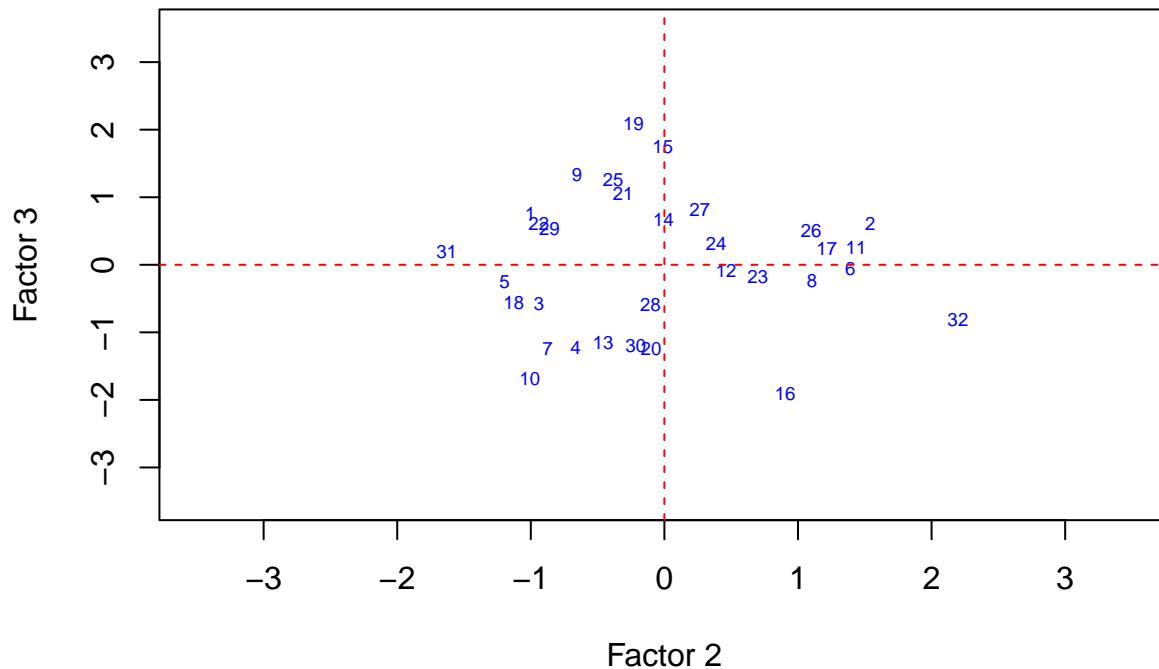


En la gráfica anterior, se muestra como la entidad 9 (CDMX), registra valores máximos para el factor 1 y también está dentro del top de los valores máximos para el factor 3, este factor hace referencia a la incidencia delictiva, la prevalencia delictiva, el promedio del costo del delito y la variación porcentual del índice de conflictividad. Un comportamiento similar tiene la entidad 15 (EDOMEX). En otro cuadrante se encuentra la entidad 19 (Nuevo León), en donde se registran los valores mínimos del factor 1 en contraste con el registro de valores máximos para el factor 3. Por otro lado, la entidad 31 (Yucatán) reporta un valor por debajo del nivel nacional para el factor 1 y para el factor 3 reporta un valor muy parecido al del nacional para el factor 3. En otro extremo del otro cuadrante, se encuentra la entidad 16 (Michoacán de Ocampo), en donde se registran valores por encima de la media nacional para el factor 1 en contraste con el factor 3 que registra uno de los valores mínimos, es decir, hay cierto nivel de desconfianza en las autoridades y bajos índices delictivos , costos por promedio y variación del índice de conflictividad.

Se realiza el gráfico del Factor 2 vs Factor 3

```
plot(res_score2_v2[, 3], res_score2_v2[, 4],main="Factor 2 vs Factor 3",xlim=c(-3.5,3.5),ylim=c(-3.5,3.5)
  xlab="Factor 2",ylab="Factor 3",type="n",lwd=2)
text(res_score2_v2[, 3],res_score2_v2[, 4],
  labels=abbreviate(row.names(res_score2_v2),minlength=8),cex=0.6,lwd=2, col = "blue")
abline(v = 0, h = 0, lty =2, col ="red")
```

Factor 2 vs Factor 3



En la gráfica anterior, se muestra que la entidad 2 (Baja California) reporta valores por encima de la media nacional para el factor 2 y valores cercanos a la media nacional para el factor 3. En otro cuadrante está la entidad 19 (Nuevo León) en donde se reporta un valor cercano a media nacional para el factor 2 y el máximo valor para el factor 3. Por otro lado, está la entidad 31 (Yucatán) en donde se registra el valor mínimo en el factor 2 y un valor cercano a la media nacional para el factor 3. En otro cuadrante, la entidad 10 (Durango), registra un valor mínimo para el factor 2 y también para el factor 3, pues ambos están por debajo de la media nacional. En otro cuadrante está la entidad 16, en donde registra un valor por encima de la media nacional para el factor 2 y un valor mínimo para el factor 3. Finalmente, en otro extremo se encuentra la entidad 32 (Zacatecas), que reporta un valor alto en el factor 2 y un valor por debajo de la media nacional para el factor 3.

9.8 Cálculo del subíndice de seguridad

Opción 1. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 1*, solo que en este caso, se obtienen los factor scores por el método de análisis de factores considerando 3 factores.

```
#Se extraen los 3 factor scores
bd_score6<-res_score2_v2
#Se convierte bd_factor_score en data frame
bd_score6<- as.data.frame(bd_score6)
bd_score6$media_seg<- apply(bd_score6[,c(2:4)],1,mean)
val_min<- min(bd_score6$media_seg) # -0.8749612
bd_score6$dif<-(bd_score6$media_seg-val_min)
val_max<- max(bd_score6$dif) #2.154723
bd_score6$ind_seg<-(bd_score6$dif/val_max)*100
#bd_score6 #La variable entidad ya está incluida
```

Opción 2. Como se observa a continuación, se utiliza forma de cálculo que se describe en la *NOTA 2*, solo que en este caso, se obtienen los factor scores por el método de análisis de factores considerando 3 factores.

```
#Se extraen los 3 factor scores
bd_score7<-res_score2_v2
#Se convierte en data frame
bd_score7<- as.data.frame(bd_score7)
val_min1<- min(bd_score7$Factor1)
val_min2<- min(bd_score7$Factor2)
val_min3<- min(bd_score7$Factor3)
bd_score7$dif1<-(bd_score7$Factor1-val_min1)
bd_score7$dif2<-(bd_score7$Factor2-val_min2)
bd_score7$dif3<-(bd_score7$Factor3-val_min3)
val_max1<- max(bd_score7$dif1)
val_max2<- max(bd_score7$dif2)
val_max3<- max(bd_score7$dif3)
bd_score7$ind_seg1<-(bd_score7$dif1/val_max1)*100
bd_score7$ind_seg2<-(bd_score7$dif2/val_max2)*100
bd_score7$ind_seg3<-(bd_score7$dif3/val_max3)*100
bd_score7$ind_seg_f<-apply(bd_score7[,c(8:10)],1,mean)
```

Con base en lo anterior, se hace una tabla resumen de los índices calculados.

```
tr<-merge(bd_score, bd_score1, by = c("ID"))
tr<-merge(tr, bd_score2, by = c("ID"))
tr<-merge(tr, bd_score3, by = c("ID"))
tr<-merge(tr, bd_score4, by = c("ID"))
tr<-merge(tr, bd_score5, by = c("ID"))
tr<-merge(tr, bd_score6, by = c("ID"))
tr<-merge(tr, bd_score7, by = c("ID"))
tabla<-tr[,c(1,9,25,33,49,55,65,71,81)]
tabla$ent <- c('AGS', 'BC',
              'BCS', 'CAMP', 'COAH', 'COL',
              'CHIS', 'CHIH', 'CDMX', 'DGO', 'GTO',
              'GRO', 'HGO', 'JAL', 'MEX',
              'MICH', 'MOR', 'NAY', 'NL', 'OAX',
              'PUE', 'QRO', 'QR', 'SLP', 'SIN',
              'SON', 'TAB', 'TAMPS', 'TLAX',
              'VER', 'YUC', 'ZAC')

#Cambiando nombre
names (tabla) = c("CVE_ENT", "cp5_opc1", "cp5_opc2", "afc5_opc1",
                  "afc5_opc2", "cp3_opc1", "cp3_opc2", "afc3_opc1",
                  "afc3_opc2", "ent")
tabla$CVE_ENT <- as.numeric(tabla$CVE_ENT)
tabla_final<- tabla[,c(10,2,3,4,5,6,7,8,9)]
```

```

# Formato de tabla
tbl(tabla_final, longtable = T, booktabs = T, digits = 3,
  col.names = c("ent", "cp5_opc1", "cp5_opc2", "afc5_opc1",
  "afc5_opc2", "cp3_opc1", "cp3_opc2", "afc3_opc1",
  "afc3_opc2"), row.names = FALSE,
  caption = "Resultados del subíndice de riesgo de seguridad") %>%
kable_styling(latex_options = c("repeat_header"))

```

Table 10: Resultados del subíndice de riesgo de seguridad

ent	cp5_opc1	cp5_opc2	afc5_opc1	afc5_opc2	cp3_opc1	cp3_opc2	afc3_opc1	afc3_opc2
AGS	40.728	52.741	20.441	42.697	39.828	42.639	29.091	38.681
BC	64.874	57.630	40.749	50.877	100.000	69.421	67.891	60.333
BCS	38.998	50.802	29.688	47.033	9.260	24.753	13.594	29.806
CAMP	44.821	53.392	37.702	51.585	14.491	25.741	12.398	28.646
COAH	16.262	38.270	7.908	35.228	30.079	38.065	18.399	31.795
COL	79.670	66.384	56.923	61.454	73.861	54.070	52.363	52.276
CHIS	16.602	37.555	1.260	33.701	31.076	37.864	0.000	23.099
CHIH	64.213	56.694	47.255	54.777	76.213	55.668	52.495	51.265
CDMX	77.770	62.542	100.000	77.995	46.247	33.333	98.701	68.930
DGO	50.968	57.330	26.301	44.137	20.100	29.182	10.813	26.234
GTO	74.630	60.961	55.432	59.552	96.001	64.720	64.782	58.068
GRO	56.826	54.661	56.488	61.708	43.326	38.512	41.811	45.685
HGO	45.297	50.714	26.499	45.680	32.933	35.227	15.952	30.759
JAL	14.577	26.920	26.507	44.318	69.744	55.429	51.778	50.135
MEX	49.106	45.151	63.141	59.361	92.549	64.183	100.000	71.871
MICH	50.108	46.500	35.889	47.994	65.723	49.424	37.784	41.346
MOR	72.011	57.450	62.917	62.359	87.399	59.658	71.469	60.214
NAY	17.673	39.368	3.417	33.609	23.914	35.629	9.147	27.574
NL	29.796	44.515	7.239	39.018	67.487	59.717	34.476	45.519
OAX	44.365	50.232	33.263	49.833	27.674	32.096	17.890	32.246
PUE	46.649	48.999	43.812	52.911	66.183	52.890	62.429	54.501
QRO	45.041	52.701	25.714	43.239	52.269	47.845	42.798	44.098
QR	72.677	58.042	84.564	72.666	62.181	43.617	70.942	57.869
SLP	68.257	62.137	44.203	54.863	66.787	52.660	47.626	48.694
SIN	42.588	53.373	18.654	43.940	43.740	45.839	27.953	40.783
SON	79.998	68.077	45.822	55.176	82.902	60.798	56.622	54.458
TAB	46.060	44.366	55.119	58.652	68.084	51.091	68.791	57.985
TAMPS	40.602	47.340	17.016	42.518	56.067	49.078	16.143	32.871
TLAX	54.743	59.748	34.533	50.488	34.973	38.054	29.374	38.689
VER	35.063	42.565	24.432	44.522	46.871	41.856	21.165	33.361
YUC	0.000	33.450	0.000	33.201	0.000	24.307	1.415	24.792
ZAC	100.000	72.952	77.011	71.955	83.080	54.484	53.321	53.088

Como se observa en la tabla resumen, se están representando las 8 opciones de subíndices que se construyeron, dado que se está midiendo un riesgo, bajo el supuesto de que el riesgo siempre existe, entonces, éste no puede tener valor de cero, en ese sentido, los subíndices “cp5_opc1”, “afc5_opc1”, “cp3_opc1” y “afc3_opc1” no se considerarán como opción factible, por lo tanto no se considerarían.

10 Mapas de calor por entidad federativa

Se realizan mapas de calor por cada una de las 8 opciones que se realizaron de los subíndices de seguridad, con la finalidad de comparar los resultados. Se lee la capa en el formato shp

```
# Leer la capa de entidad federativa con la librería de sf
mexico <- st_read("C:/Users/User/Documents/ÍNDICE DE RIESGO/Bases de datos/Capas/00ent.shp")

## Reading layer '00ent' from data source
##   'C:/Users/User/Documents/ÍNDICE DE RIESGO/Bases de datos/Capas/00ent.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 32 features and 3 fields
## Geometry type: MULTIPOLYGON
## Dimension:      XY
## Bounding box:  xmin: 911292 ymin: 319149.1 xmax: 4082997 ymax: 2349615
## Projected CRS: MEXICO_ITRF_2008_LCC

# Convertir el objeto sf a un data frame para ggplot
mexico_df <- mexico %>%
  mutate(CVE_ENT = as.numeric(CVE_ENT)) %>%
  st_as_sf()
```

Se trabaja un mapa por subíndice calculado

```
#Tabulado para el mapa 1 correspondiente al subíndice 1:cp5_opc1
sub_cp5_opc1 <- tabla %>%
  dplyr::select(CVE_ENT, cp5_opc1) %>%
  mutate(cp5_opc1 = round(cp5_opc1, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa1 <- mexico_df %>%
  left_join(sub_cp5_opc1, by = "CVE_ENT")

#Tabulado para el mapa 2 correspondiente al subíndice 2:cp5_opc2
sub_cp5_opc2 <- tabla %>%
  dplyr::select(CVE_ENT, cp5_opc2) %>%
  mutate(cp5_opc2 = round(cp5_opc2, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa2 <- mexico_df %>%
  left_join(sub_cp5_opc2, by = "CVE_ENT")

#Tabulado para el mapa 3 correspondiente al subíndice 3:afc5_opc1
sub_afc5_opc1 <- tabla %>%
  dplyr::select(CVE_ENT, afc5_opc1) %>%
  mutate(afc5_opc1 = round(afc5_opc1, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa3 <- mexico_df %>%
```

```

left_join(sub_afc5_opc1, by = "CVE_ENT")

#Tabulado para el mapa 4 correspondiente al subíndice 4:afc5_opc2
sub_afc5_opc2 <- tabla %>%
  dplyr::select(CVE_ENT, afc5_opc2) %>%
  mutate(afc5_opc2 = round(afc5_opc2, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa4 <- mexico_df %>%
  left_join(sub_afc5_opc2, by = "CVE_ENT")

#Tabulado para el mapa 5 correspondiente al subíndice 5:cp3_opc1
sub_cp3_opc1 <- tabla %>%
  dplyr::select(CVE_ENT, cp3_opc1) %>%
  mutate(cp3_opc1 = round(cp3_opc1, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa5 <- mexico_df %>%
  left_join(sub_cp3_opc1, by = "CVE_ENT")

#Tabulado para el mapa 6 correspondiente al subíndice 6:cp3_opc2
sub_cp3_opc2 <- tabla %>%
  dplyr::select(CVE_ENT, cp3_opc2) %>%
  mutate(cp3_opc2 = round(cp3_opc2, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa6 <- mexico_df %>%
  left_join(sub_cp3_opc2, by = "CVE_ENT")

#Tabulado para el mapa 7 correspondiente al subíndice 7:afc3_opc1
sub_afc3_opc1 <- tabla %>%
  dplyr::select(CVE_ENT, afc3_opc1) %>%
  mutate(afc3_opc1 = round(afc3_opc1, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa7 <- mexico_df %>%
  left_join(sub_afc3_opc1, by = "CVE_ENT")

#Tabulado para el mapa 8correspondiente al subíndice 8:afc3_opc2
sub_afc3_opc2 <- tabla %>%
  dplyr::select(CVE_ENT, afc3_opc2) %>%
  mutate(afc3_opc2 = round(afc3_opc2, 1)) # Redondear a un decimal
# Unir los datos espaciales con los valores del subíndice
datos_mapa8 <- mexico_df %>%
  left_join(sub_afc3_opc2, by = "CVE_ENT")

```

Mapa del subíndice 1

```

ggplot() +
  geom_sf(data = datos_mapa1, aes(fill = cp5_opc1), color = "gray") +
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +
  labs(
    title = "Subíndice 1",
    caption = "Elaboración propia.\nSubíndice realizado con 5 componentes principales, forma 1 de cálculo",
    ) +
  theme_minimal() +
  theme(plot.caption = element_text(size = 8))

```

Subíndice 1



Mapa del subíndice 2

```
ggplot() +  
  geom_sf(data = datos_mapa2, aes(fill = cp5_opc2), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 2",  
    caption = "Elaboración propia.\nSubíndice realizado con 5 componentes principales, forma 2 de cálculo") +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

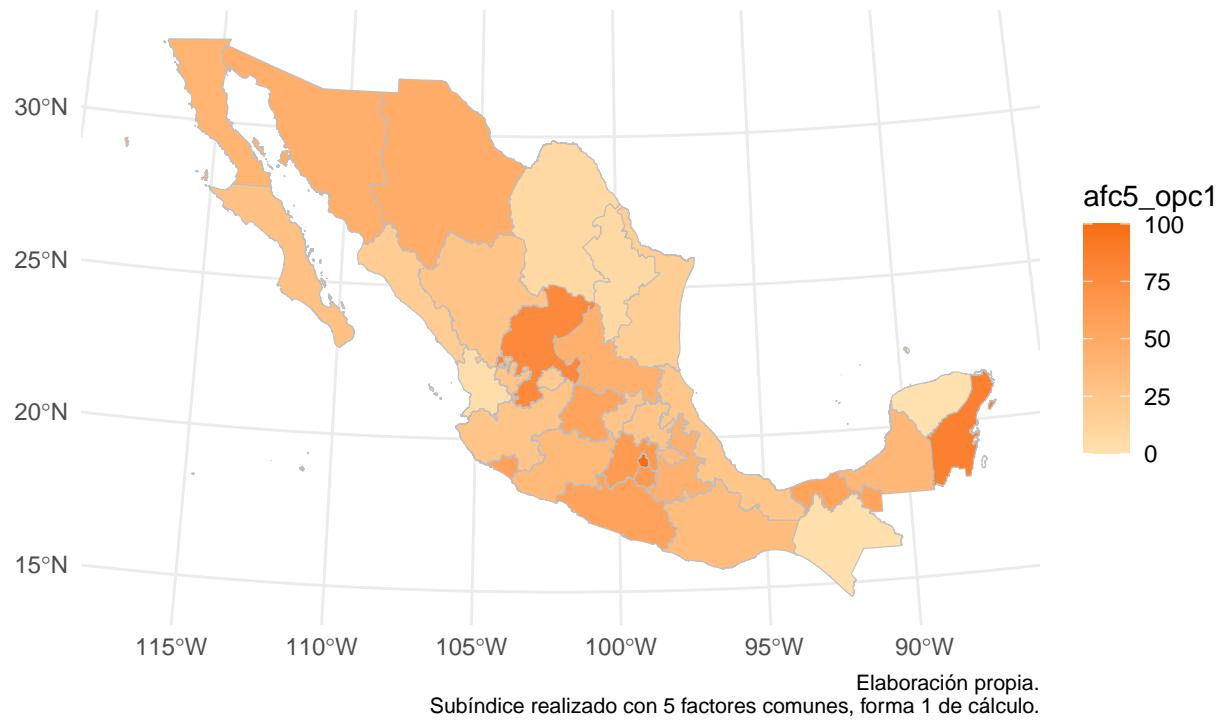
Subíndice 2



Mapa del subíndice 3

```
ggplot() +  
  geom_sf(data = datos_mapa3, aes(fill = afc5_opc1), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 3",  
    caption = "Elaboración propia.\nSubíndice realizado con 5 factores comunes, forma 1 de cálculo."  
  ) +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

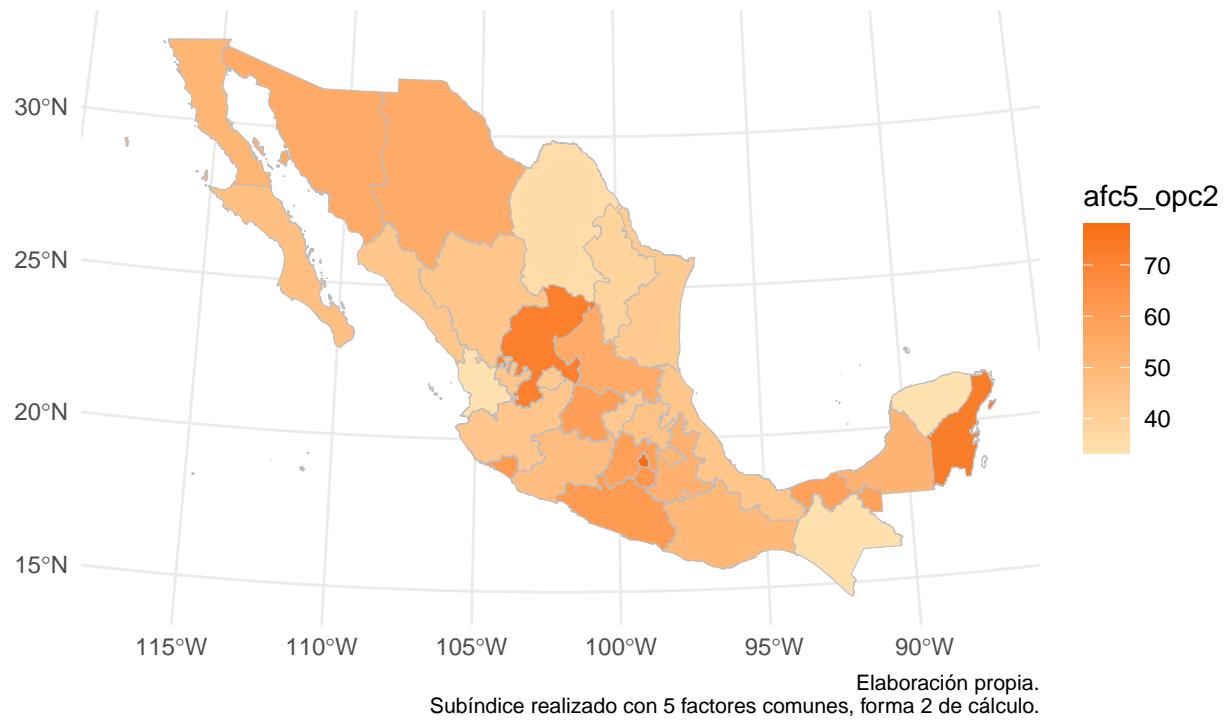
Subíndice 3



Mapa del subíndice 4

```
ggplot() +  
  geom_sf(data = datos_mapa4, aes(fill = afc5_opc2), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 4",  
    caption = "Elaboración propia.\nSubíndice realizado con 5 factores comunes, forma 2 de cálculo."  
  ) +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

Subíndice 4



Mapa del subíndice 5

```
ggplot() +  
  geom_sf(data = datos_mapa5, aes(fill = cp3_opc1), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 5",  
    caption = "Elaboración propia.\nSubíndice realizado con 3 componentes principales, forma 1 de cálculo") +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

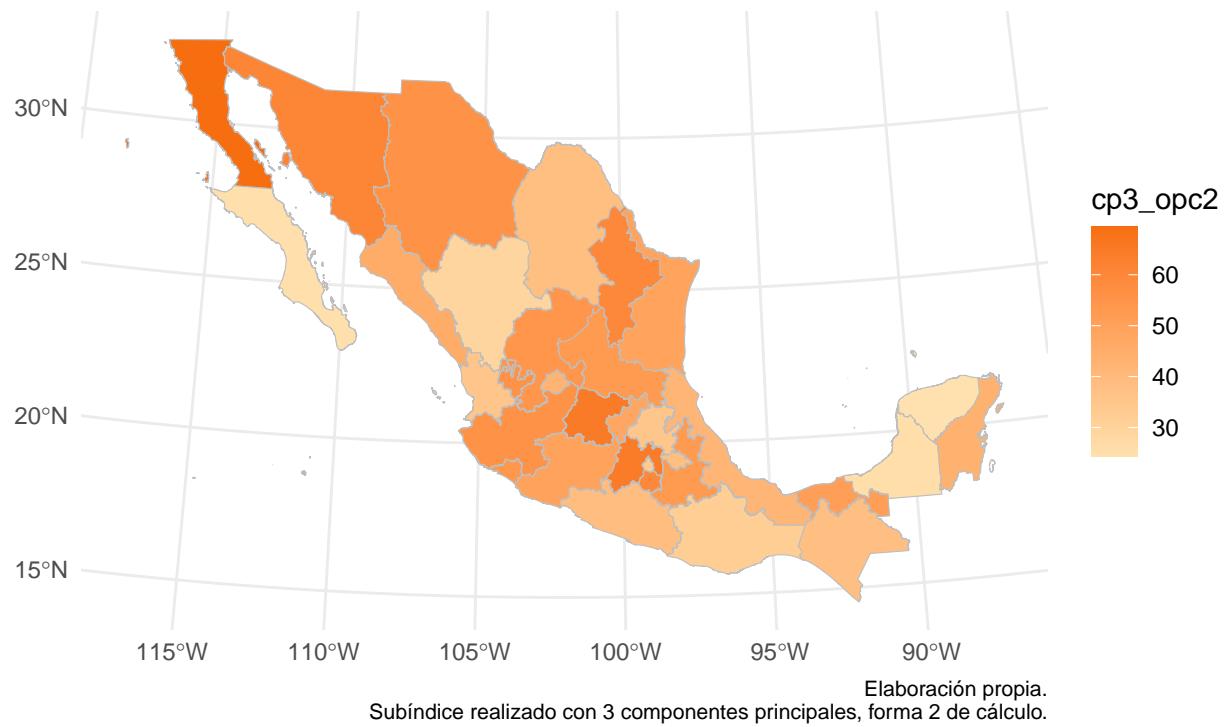
Subíndice 5



Mapa del subíndice 6

```
ggplot() +  
  geom_sf(data = datos_mapa6, aes(fill = cp3_opc2), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 6",  
    caption = "Elaboración propia.\nSubíndice realizado con 3 componentes principales, forma 2 de cálculo") +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

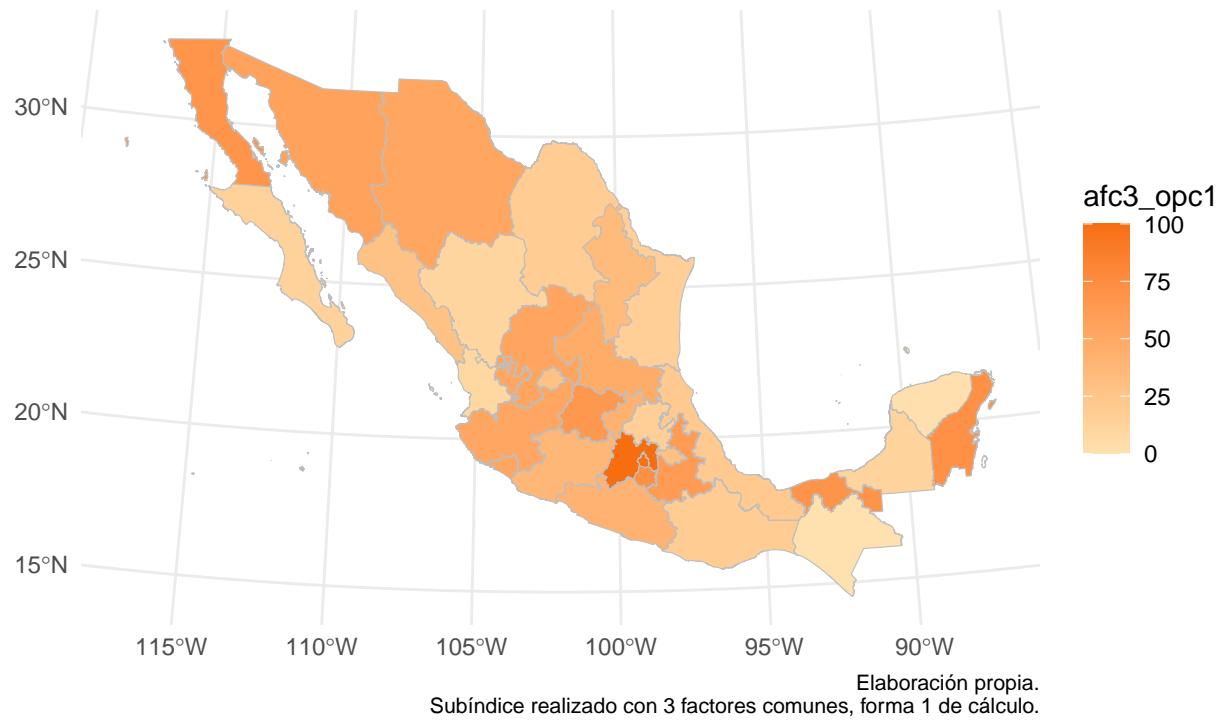
Subíndice 6



Mapa del subíndice 7

```
ggplot() +  
  geom_sf(data = datos_mapa7, aes(fill = afc3_opc1), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 7",  
    caption = "Elaboración propia.\nSubíndice realizado con 3 factores comunes, forma 1 de cálculo."  
  ) +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

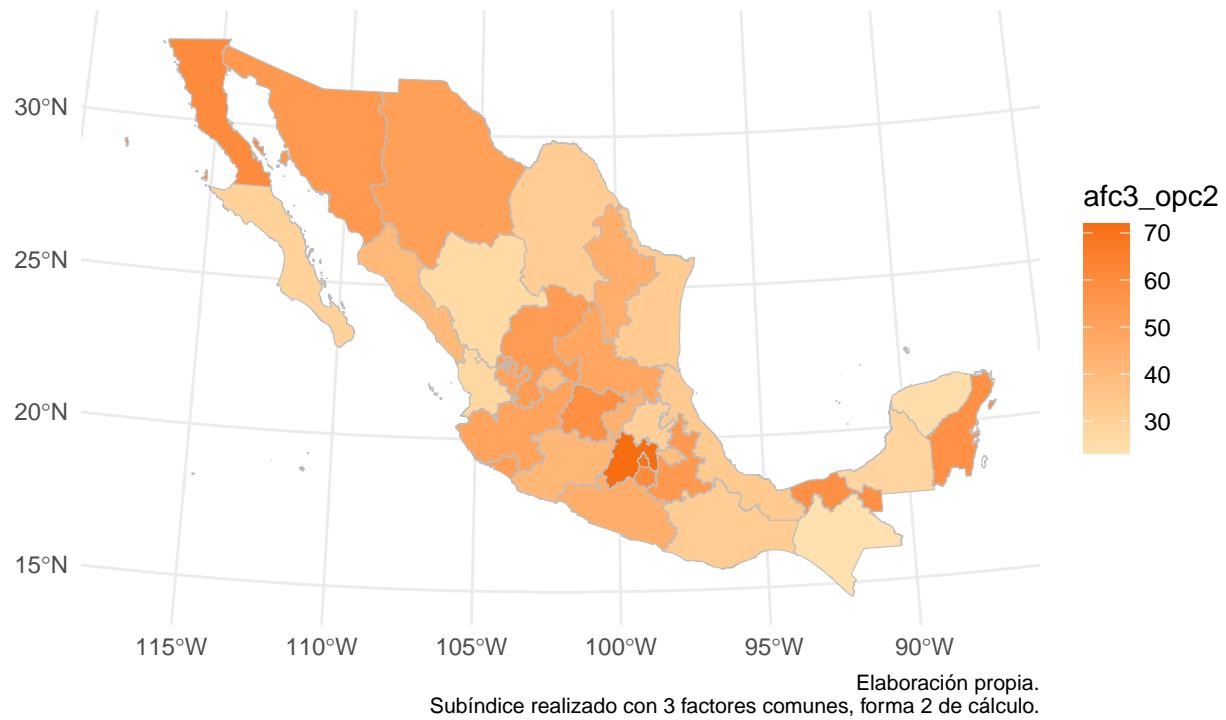
Subíndice 7



Mapa del subíndice 8

```
ggplot() +  
  geom_sf(data = datos_mapa8, aes(fill = afc3_opc2), color = "gray") +  
  scale_fill_gradient(low = "#FFE1AF", high = "#F76E11") +  
  labs(  
    title = "Subíndice 8",  
    caption = "Elaboración propia.\nSubíndice realizado con 3 factores comunes, forma 2 de cálculo."  
  ) +  
  theme_minimal() +  
  theme(plot.caption = element_text(size = 8))
```

Subíndice 8



Conclusiones:

Como se describió en el documento, el análisis realizado estuvo enfocado en ver qué opción sería la más factible, para calcular el subíndice de riesgo, tomando como base los resultados obtenidos por entidad federativa, se sugiere que el subíndice 2: “cp5_opc2”, realizado por componentes principales, es una buena opción, porque se están considerando a todas las variables, los resultados arrojados empantan con la realidad en términos de seguridad/inseguridad a nivel entidad federativa y finalmente, por cuestiones de parsimonia estadística, se sugiere considerar al subíndice 2, como la opción más factible.

Simbología:

1. seg: Es la base con todas las variables de interés y con la que se trabaja en la primera parte.
2. est_seg: Es la base de datos “seg” estandarizada
3. cor_seg: Es la matriz de correlación de las variables de la base “seg”.
4. det_seg: Es el valor del determinante de la matriz de correlaciones “cor_seg”
5. seg.pc: Es el objeto en donde se realiza el análisis de componentes principales usando la base “est_seg”, devolviendo los resultados como un objeto de clase princomp.
6. bd_score: Es la tabla que contiene los scores seleccionados (5) por componentes principales, el id (entidad federativa) además, contiene las variables auxiliares (media_seg y dif) para el cálculo del subíndice y finalmente, contiene la primera opción del subíndice de seguridad **ind_seg**. A continuación, se detalla cada una de las variables o valores usados en el cálculo.
 - 6.1. media_seg: Media de los scores elegidos, por entidad federativa.
 - 6.2. val_min: Es el valor mínimo de la variable *media_seg*.
 - 6.3. dif: Es la diferencia de la media *media_seg* y el valor mínimo de la media *val_min*.
 - 6.4. val_max: Es el valor máximo de la diferencia *dif*.
 - 6.5. ind_seg: Es el subíndice de seguridad, se obtiene al dividir la diferencia *dif* entre el valor máximo de la diferencia *val_max*, multiplicado por 100.
7. bd_score1: Es la tabla que contiene los scores seleccionados (5) por componentes principales, el id (entidad federativa) además, contiene las variables auxiliares dif1, dif2, dif3, dif4, dif5, ind_seg1, ind_seg2, ind_seg3, ind_seg4 y ind_seg5 para el cálculo del subíndice y finalmente, contiene la segunda opción del subíndice de seguridad **ind_seg_f**. A continuación, se detalla cada una de las variables o valores usados en el cálculo.
 - 7.1. val_{min_i} : Es el valor mínimo de cada uno de los scores, con $i = 1, 2, 3...n$ y n es el número de scores elegidos, en este ejemplo son 5.
 - 7.2. dif_i : Es la diferencia del $score_i$ y el valor mínimo de cada score val_{min_i} .
 - 7.3. val_{max_i} : Es el valor máximo para cada una de las diferencias dif_i .
 - 7.4. ind_{seg_i} : Es el subíndice “i” de cada score. Se obtiene al dividir la diferencia dif_i entre el valor máximo de la diferencia val_{max_i} , multiplicado por 100.
 - 7.5. ind_{seg_f} : Es el subíndice de seguridad (opción 2), se obtiene al promediar los ind_{seg_i} .
8. af.mv5: Es el objeto en donde se realiza el análisis factorial de máxima verosimilitud con la base de datos estandarizada “est_seg”. Se le puso el número 5 para referir que son 5 factores.
9. af.mv6: Es el objeto en donde se realiza el análisis factorial de máxima verosimilitud con la base de datos estandarizada “est_seg”. Se le puso el número 6 para referir que son 6 factores.
10. comu_mv5: Comunalidades del modelo con 5 factores
11. espe_mv5: Varianza específica del modelo con 5 factores
12. res_mv5: Es un data frame con los 5 factores, las comunalidades y la varianza única con un modelo de 5 factores.
13. comu_mv6: Comunalidades del modelo con 6 factores
14. espe_mv6: Varianza específica del modelo con 6 factores
15. res_mv6: Es un data frame con los 6 factores, las comunalidades y la varianza única con un modelo de 6 factores.
16. af.mv5\$PVAL: “Valor p” para el modelo de 5 factores con base en la prueba de hipótesis.
17. af.mv6\$PVAL: “Valor p” para el modelo de 6 factores con base en la prueba de hipótesis.
18. mat_rep_mv5: Matriz reproducida con 5 factores, se puede interpretar como la matriz estimada a partir de los 5 factores.

19. mat_res_mv5: Matriz residual, es la matriz de diferencias entre la matriz de correlación (integrada con las variables originales) y la matriz reproducida (con la agrupación de las variables en 5 factores propuestos)
20. mat_rep_mv6: Matriz reproducida con 6 factores, se puede interpretar como la matriz estimada a partir de los 6 factores.
21. mat_res_mv6: Matriz residual, es la matriz de diferencias entre la matriz de correlación (integrada con las variables originales) y la matriz reproducida (con la agrupación de las variables en 6 factores propuestos)
22. scores_min: Cálculo de factor scores mediante enfoque de mínimos cuadrados, estos son a nivel observación, en este caso son 32 por las entidades federativas.
23. scores_reg: Cálculo de factor scores mediante enfoque de regresión, estos son a nivel observación, en este caso son 32 por las entidades federativas.
24. diferencia: Es matriz en donde se hace la diferencia entre el cálculo de los factor scores con el método de mínimos cuadrados vs el de regresión. De esta manera si la diferencia es cercana a cero, las interpretaciones que se den sobre las entidades serán aplicables a ambos enfoques.
25. bd_score2: Es la tabla que contiene los factor scores seleccionados (5) por análisis de factores, el id (entidad federativa) además, contiene las variables auxiliares (media_seg y dif) para el cálculo del subíndice y finalmente, contiene la primera opción del subíndice de seguridad **ind_seg**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 6.1 al 6.5
26. bd_score3: Es la tabla que contiene los factor scores seleccionados (5) por análisis de factores, el id (entidad federativa) además, contiene las variables auxiliares dif1, dif2, dif3, dif4, dif5, ind_seg1, ind_seg2, ind_seg3, ind_seg4 y ind_seg5 para el cálculo del subíndice y finalmente, contiene la segunda opción del subíndice de seguridad **ind_seg_f**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 7.1 al 7.5
27. seg1: Es la base en donde se eliminan las variables que tienen poca correlación.
28. est_seg1: Es la base de datos “seg1” estandarizada.
29. cor_seg1: Es la matriz de correlación de las variables de la base “seg1”.
30. det_seg1: Es el valor del determinante de la matriz de correlaciones “cor_seg1”
31. seg_pc1: Es el objeto en donde se realiza el análisis de componentes principales usando la base “est_seg1”, devolviendo los resultados como un objeto de clase princomp.
32. bd_score4: Es la tabla que contiene los scores seleccionados (3) por componentes principales, el id (entidad federativa) además, contiene las variables auxiliares (media_seg y dif) para el cálculo del subíndice y finalmente, contiene la primera opción del subíndice de seguridad **ind_seg**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 6.1 al 6.5
33. bd_score5: Es la tabla que contiene los scores seleccionados (3) por componentes principales, el id (entidad federativa) además, contiene las variables auxiliares dif1, dif2, dif3, ind_seg1, ind_seg2, ind_seg3, para el cálculo del subíndice y finalmente, contiene la segunda opción del subíndice de seguridad **ind_seg_f**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 7.1 al 7.5.
34. af.mv3: Es el objeto en donde se realiza el análisis factorial de máxima verosimilitud con la base de datos estandarizada “est_seg1”. Se le puso el número 3 para referir que son 3 factores.
35. comu_mv3: Comunalidades del modelo con 3 factores
36. espe_mv3: Varianza específica del modelo con 3 factores
37. res_mv3: Es un data frame con los 3 factores, las comunalidades y la varianza única con un modelo de 3 factores.
38. af.mv3\$PVAL: “Valor p” para el modelo de 3 factores con base en la prueba de hipótesis.
39. mat_rep_mv3: Matriz reproducida con 3 factores, se puede interpretar como la matriz estimada a partir de los 3 factores, usando “est_seg1”
40. mat_res_mv3: Matriz residual, es la matriz de diferencias entre la matriz de correlación (integrada con las variables originales), usando “est_seg1” y la matriz reproducida (con la agrupación de las variables en 3 factores propuestos)
41. scores_min_v2: Cálculo de factor scores mediante enfoque de mínimos cuadrados, estos son a nivel

- observación, en este caso son 32 entidades federativas, considerando 3 factores y “est_seg1”
42. scores_reg_v2: Cálculo de factor scores mediante enfoque de regresión, estos son a nivel observación, en este caso son 32 entidades federativas, considerando 3 factores y “est_seg1”
 43. diferencia_v2: Es matriz en donde se hace la diferencia entre el cálculo de los factor scores con el método de mínimos cuadrados vs el de regresión, considerando 3 factores y “est_seg1”
 44. bd_score6: Es la tabla que contiene los factor scores seleccionados (3) por análisis de factores, el id (entidad federativa) además, contiene las variables auxiliares (media_seg y dif) para el cálculo del subíndice y finalmente, contiene la primera opción del subíndice de seguridad **ind_seg**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 6.1 al 6.5
 45. bd_score7: Es la tabla que contiene los factor scores seleccionados (3) por análisis de factores, el id (entidad federativa) además, contiene las variables auxiliares dif1, dif2, dif3, ind_seg1, ind_seg2 y ind_seg3 para el cálculo del subíndice y finalmente, contiene la segunda opción del subíndice de seguridad **ind_seg_f**. Las variables y valores usados para el cálculo del subíndice de seguridad son las mismas que las descritas en los puntos: 7.1 al 7.5
 46. tr: Base con todas las variables de las tablas bd_score al bd_score7, la unión se realizó por el ID de la entidad federativa, la base incluye los 8 subíndices.
 47. tabla: Es la base con los 8 subíndices calculados por la entidad federativa. Se realizó un cambio del nombre de los subíndices asociando el método usado, el número de componentes o factores elegidos y la opción de cálculo del subíndice. Además, se incluye la variable “ent”, que refiere a los nombres de las entidades federativas. A continuación, se detalla:
 - 47.1. cp5_opc1: Opción 1 del subíndice, se extrajo de “bd_score”.
 - 47.2. cp5_opc2: Opción 2 del subíndice, se extrajo de “bd_score1”.
 - 47.3. afc5_opc1: Opción 3 del subíndice, se extrajo de “bd_score2”.
 - 47.4. afc5_opc2: Opción 4 del subíndice, se extrajo de “bd_score3”.
 - 47.5. cp3_opc1: Opción 5 del subíndice, se extrajo de “bd_score4”.
 - 47.6. cp3_opc2: Opción 6 del subíndice, se extrajo de “bd_score5”.
 - 47.7. afc3_opc1: Opción 7 del subíndice, se extrajo de “bd_score6”.
 - 47.8. afc3_opc2: Opción 8 del subíndice, se extrajo de “bd_score7”.
 48. tabla_final: Es la tabla con el formato final, integrada por la variable “ent” y los 8 subíndices descritos en el punto anterior.

ANEXO

3 factores y 4 factores.

Para 3 factores

```
# Cargas estimadas rotadas para 3 factores por máxima verosimilitud
af.mv3<-factanal(est_seg, factors = 3, rotation = "varimax", n.obs = nrow(est_seg))
# Se almacenan las cargas
cargas_mv3<-af.mv3$loadings
vx3<-colSums(cargas_mv3^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx3, `Proporción de la varianza` = vx3/nrow(cargas_mv3),
      `Varianza acumulada` = cumsum(vx3/nrow(cargas_mv3)))

##                                     Factor1   Factor2   Factor3
## SS loadings                 3.3144280 3.2911470 2.2320395
## Proporción de la varianza  0.1949664 0.1935969 0.1312964
## Varianza acumulada        0.1949664 0.3885632 0.5198597
```

Prueba de hipótesis

```
# Valor p para el modelo de 3 factores
af.mv3$PVAL
```

```
## objective
## 0.01429138
```

Como el valor p es menor a 0.05 ($0.01429138 < 0.05$) se rechaza H_0 , es decir, 3 factores en el modelo no son suficientes para capturar la dimensionalidad completa del conjunto de datos.

Para 4 factores

```
# Cargas estimadas rotadas para 5 factores por máxima verosimilitud
af.mv4<-factanal(est_seg, factors = 4, rotation = "varimax", n.obs = nrow(est_seg))
# Se almacenan las cargas
cargas_mv4<-af.mv4$loadings
vx4<-colSums(cargas_mv4^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx4, `Proporción de la varianza` = vx4/nrow(cargas_mv4),
      `Varianza acumulada` = cumsum(vx4/nrow(cargas_mv4)))

##                                     Factor1   Factor2   Factor3   Factor4
## SS loadings                 3.11763 2.9797372 2.3846514 1.7790141
## Proporción de la varianza  0.18339 0.1752787 0.1402736 0.1046479
## Varianza acumulada        0.18339 0.3586687 0.4989423 0.6035902
```

```
# Valor p para el modelo de 2 factores
af.mv4$PVAL
```

```
## objective
## 0.06523034
```

Como el valor p es mayor a 0.05 ($0.06523034 > 0.05$) no se rechaza H_0 , es decir, 4 factores en el modelo son suficientes para capturar la dimensionalidad completa del conjunto de datos. Aunque, no se rechaza H_0 , en valor de p es muy cercano a 0.05.

VERSIÓN 2

Para 4 factores

```
# Cargas estimadas rotadas para 4 factores por máxima verosimilitud
af.mv4_v2<-factanal(est_seg1, factors = 4, rotation = "varimax", n.obs = nrow(est_seg1))
# Se almacenan las cargas
cargas_mv4_v2<-af.mv4_v2$loadings
vx4_v2<-colSums(cargas_mv4_v2^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx4_v2, `Proporción de la varianza` = vx4_v2/nrow(cargas_mv4_v2),
      `Varianza acumulada` = cumsum(vx4_v2/nrow(cargas_mv4_v2)))

##                               Factor1  Factor2  Factor3  Factor4
## SS loadings            2.746943 2.5208080 2.4153203 1.3192531
## Proporción de la varianza 0.228912 0.2100673 0.2012767 0.1099378
## Varianza acumulada      0.228912 0.4389793 0.6402560 0.7501937
```

Prueba de hipótesis

```
# Valor p para el modelo de 4 factores
af.mv4_v2$PVAL
```

```
## objective
## 0.01137698
```

Como el valor p es menor a 0.05 ($0.01137698 < 0.05$) se rechaza H_0 , es decir, 4 factores en el modelo no son suficientes para capturar la dimensionalidad completa del conjunto de datos.

Para 5 factores

```
# Cargas estimadas rotadas para 5 factores por máxima verosimilitud
af.mv5_v2<-factanal(est_seg1, factors = 5, rotation = "varimax", n.obs = nrow(est_seg1))
# Se almacenan las cargas
cargas_mv5_v2<-af.mv5_v2$loadings
vx5_v2<-colSums(cargas_mv5_v2^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx5_v2, `Proporción de la varianza` = vx5_v2/nrow(cargas_mv5_v2),
      `Varianza acumulada` = cumsum(vx5_v2/nrow(cargas_mv5_v2)))
```

```
##                               Factor1  Factor2  Factor3  Factor4  Factor5
## SS loadings            2.5068080 2.3558911 2.3263072 1.2397626 1.2074055
## Proporción de la varianza 0.2089007 0.1963243 0.1938589 0.1033136 0.1006171
## Varianza acumulada      0.2089007 0.4052249 0.5990839 0.7023974 0.8030145
```

Prueba de hipótesis

```
# Valor p para el modelo de 5 factores
af.mv5_v2$PVAL
```

```
## objective
## 0.06135688
```

Como el valor p es mayor a 0.05 ($0.06135688 > 0.05$) no se rechaza H_0 , es decir, 5 factores en el modelo son suficientes para capturar la dimensionalidad completa del conjunto de datos.

Para 6 factores

```
# Cargas estimadas rotadas para 6 factores por máxima verosimilitud
af.mv6_v2<-factanal(est_seg1, factors = 6, rotation = "varimax", n.obs = nrow(est_seg1))
# Se almacenan las cargas
cargas_mv6_v2<-af.mv6_v2$loadings
vx6_v2<-colSums(cargas_mv6_v2^2)
# Se imprime la proporción de varianza explicada
rbind(`SS loadings` = vx6_v2, `Proporción de la varianza` = vx6_v2/nrow(cargas_mv6_v2),
      `Varianza acumulada` = cumsum(vx6_v2/nrow(cargas_mv6_v2)))
```

	Factor1	Factor2	Factor3	Factor4	Factor5
## SS loadings	2.4662702	2.2659822	2.2439585	1.3417039	1.16390354
## Proporción de la varianza	0.2055225	0.1888318	0.1869965	0.1118087	0.09699196
## Varianza acumulada	0.2055225	0.3943544	0.5813509	0.6931596	0.79015152
##		Factor6			
## SS loadings		0.90034366			
## Proporción de la varianza		0.07502864			
## Varianza acumulada		0.86518016			

Prueba de hipótesis

```
# Valor p para el modelo de 6 factores
af.mv6_v2$PVAL
```

```
## objective
## 0.1153948
```

Como el valor p es mayor a 0.05 ($0.1153948 > 0.05$) no se rechaza H_0 , es decir, 6 factores en el modelo son suficientes para capturar la dimensionalidad completa del conjunto de datos.