Taller 2: Análisis Multivariado

Análisis de Componentes Principales (PCA)

Julián Camilo Riaño Moreno

jueves, junio 18, 2020

Contents

Actividad #2

> Realice un análisis de componentes principales (ACP) sobre las variables cuan- titativas (por lo menos 5) de una base de datos relacionada con su área de conocimientos. Muestre el porcentaje de varianza explicado por cada dimensión. Basado en el autovalor medio o en un diagrama de sedimentación, decida cúantos componentes debe retener. Determine la calidad de representación de las vari- ables respecto a las componentes seleccionadas y la contribución que realiza cada varaible a la construcción de cada componente. Defína e interprete los componentes principales obtenidos.

Especificaciones sobre la base de datos utilizada.

Para esta actividad se utilizó la misma base de datos descrita en la actividad 1 del curso de análisis multivariado de la especialización en estadística aplicada, durante el curso del 2020-1. A modo de resumen, se trata de una base de datos corresponde a las frecuencias o porcentaje de resistencia a 4 grupos de antibióticos (Carbapenems, Aminoglucosidos, Fluoroquinolonas y cefalosporinas de 3er generación), en aislados de dos bacterias *Escherichia coli; Klebsiella pneumoniae*; los datos fueron obtenidos en 30 países de las 4 regiones (Sur, norte, este y oeste) de Europa durante los años 2000 al 2018. Se tienen 9 variables: `Bacteria`, `Year`, `Code_C`, `Country`, `Aminoglycosides`, `Carbapenems`, `Fluoroquinolones`, `cephalos_3er_gen`.

Para efectos de esta actividad se extrajeron únicamente las variables de caracter cuantitativo como son las frecuencias de resistencia a cada uno de los grupos de antibióticos.

Debido a las conclusiones previo realizado de correlacción, varianza y distribución de estas variables se consideró que el grupo de los `Carbapenems` presenta baja correlación con las demás variables y para la elaboración de los componentes principales esta variable se asignó como una variable *cuantitativa suplementaria* a través del argumento `quanti.sup`de la función `PCA` del paquete estadístico `factomineR`. Con esto el análisis de APC se realizó únicamente con tres variables centrales: `Aminoglycosides`, `Fluoroquinolones`, `cephalos_3er_gen`.

Solución de la actividad:

Para desarrollar la actividad se realizó un análisis a través de las funciones encontradas en el paquete `factomineR` y `factoextra`.

Table 1: Matriz de *eigenvalues*

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.773	92.45	92.45
Dim.2	0.1901	6.336	98.78
Dim.3	0.0365	1.217	100

Descripción de los componentes principales.

La tabla 1 muestra los resultados de los valores propios o *eigenvalues* de cada una de las componentes principales obtenidas a través del modelo aplicado. Cada componente esta definida como `Dim`. También se puede encontrar la varianza explicada por cada una de las componentes obtenidas. Desde el punto de vista estadístico según este resultado y aplicando el *criterio de Kaiser* se puede definir que la unica componente que debería ser retenida es la primera (`Dim1`), dado que su valor es > 1 (2.773) esto indica que tan solo esta componente explica más la variabilidad de las observaciones que cada una de las variables originales. Además, esta aseveración es soportada en buena medida por el estandar de explicación de varianza de 80% utilizado por buena parte de los estadísticos. Tal como se observa en el la tabla 1 y en la figura 1. la componente 1, explica el 92.45% de la variabilidad de los datos.

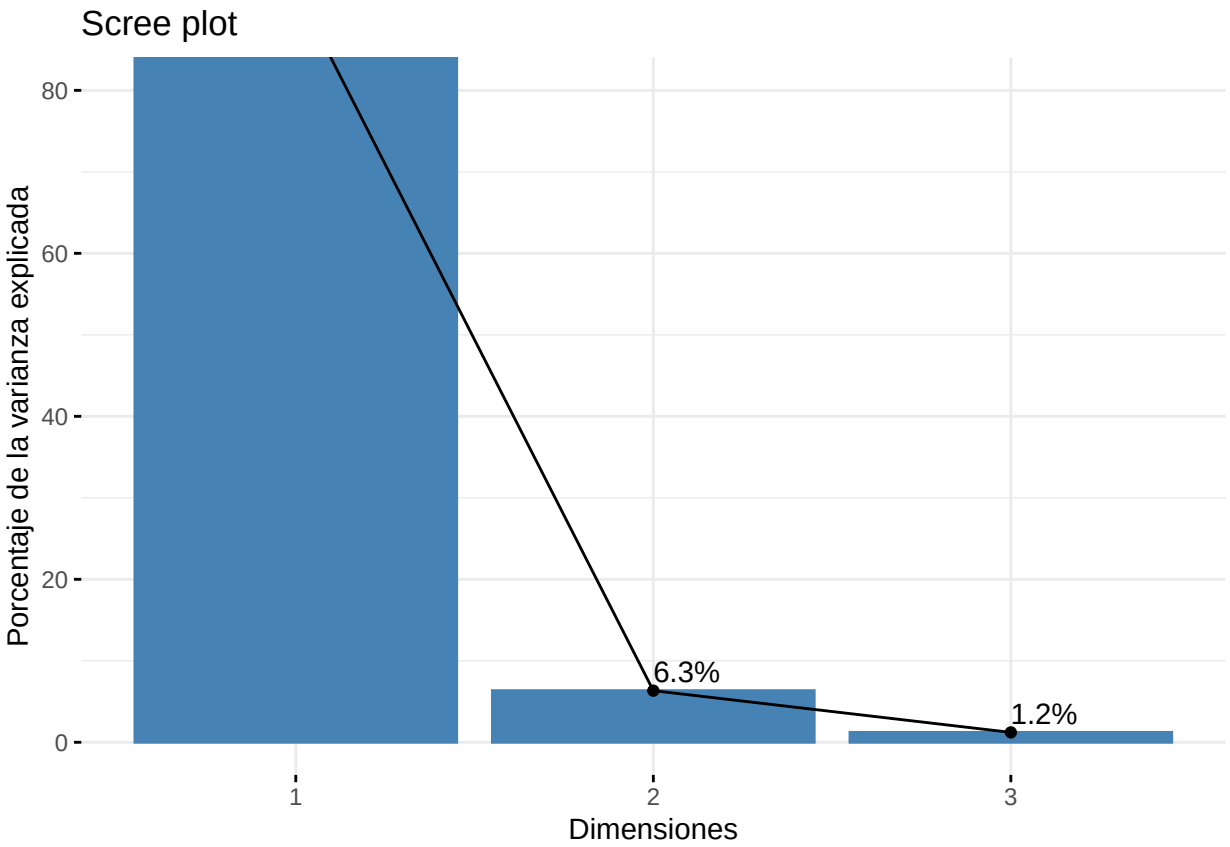

Figure 1: ScreePlot de los valores propios para cada variable

De la misma forma la figura 1. al aplicar el *"criterio del codo"*, que es una valoración cualitativa de la figura 1, dada por la exigencia de retenciaón de las componentes anteriores al punto de máxima flexión entre las curva de dibujada entre el centro de todas las componentes principales. En este caso, se evidencia que está flexión máxima se encuentra en la componente 2, por tanto, se exige la retención de únicamente de la componente 1.

Table 2: Matriz de coordinadas de las variables

	Dim.1	Dim.2	Dim.3
Aminoglycosides	0.9715	-0.1991	0.1283
Fluoroquinolones	0.933	0.3597	0.01422
cephalos_3er_gen	0.9793	-0.1451	-0.1408

Table 3: Vector de cos^2 de las variables

	Dim.1	Dim.2	Dim.3
Aminoglycosides	0.9439	0.03966	0.01647
Fluoroquinolones	0.8704	0.1294	0.0002023
cephalos_3er_gen	0.9591	0.02106	0.01984

Gráficas, coordenadas, cos^2 y contribuciones en los componentes principales.

La tabla 2. muestra la posición en el plano factorial de cada una de las varibles respecto a su componente principal. Esto deja ver que, en las figuras 2 y 3, los vectores de las tres variables se dirigen hacia la porción positiva del eje y en el plano, lo que corresponde al plano de la componnente principal 1. A diferencia de lo ocurrido con la componente principal 2, dónde tanto la variable `Cephalos_3er_gen` y `aminoglucoside`están en el plano negativo de esta componente.

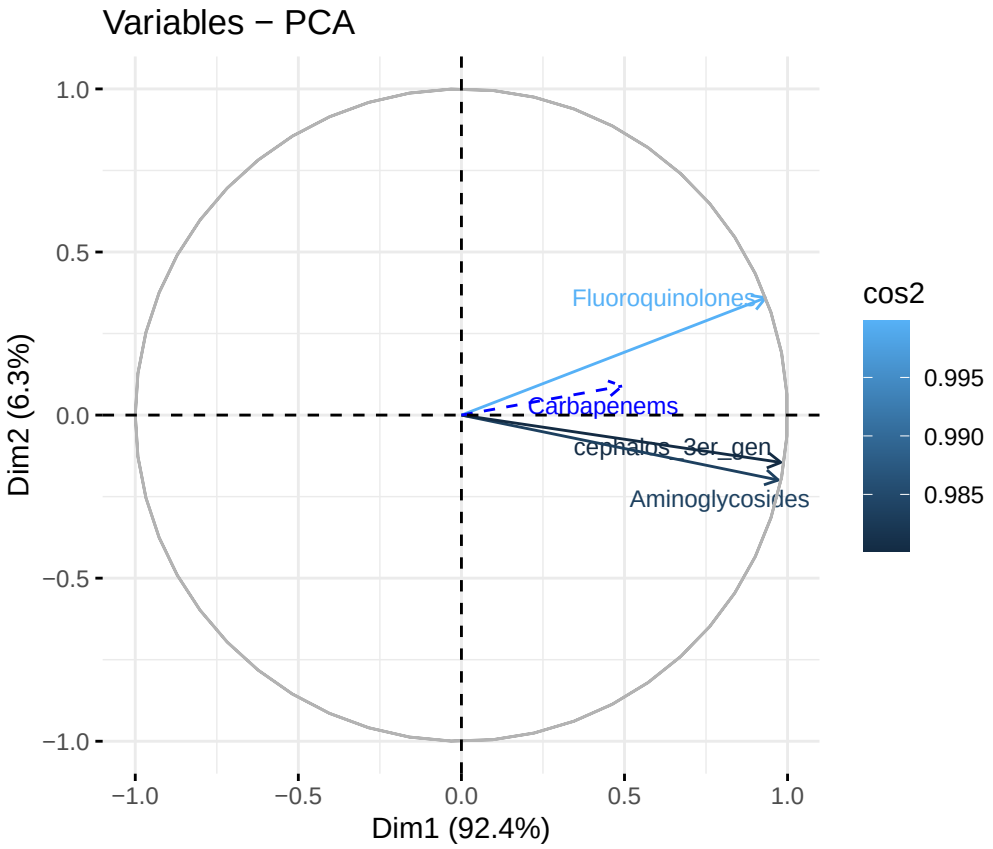

Figure 2: Gráfica de calidad de la representación por cos^2 de las variables

Por otra parte, la figura 2 y 3 muestran que los angulos de las líneas obtenidas para cada una de la variables originales a través del análisis de compnentes prinicipales y que corresponden a `Cephalos_3er_gen` y `aminoglucoside` son muy cercanos de manera que confirma que existe una alta correlación entre estas dos variables; en el caso de la variable `Fluoroquinolone`cuyo angulo es mayor con respecto a las otras dos, lo que indica menor correlación entre estas.

Table 4: Matriz de contribuciones por variables

	Dim.1	Dim.2	Dim.3
Aminoglycosides	34.03	20.86	45.1
Fluoroquinolones	31.38	68.06	0.5542
cephalos_3er_gen	34.58	11.08	54.34

La tabla 3. muestra los valores de cos^2 de las componentes prinicipales, lo que corresponde a la importancia de un componente principal para cada una de las variables originales. Esto quiere decir que la dimensión 1 tiene la mayor importancia con la variable `Aminoglicoside`(0.9439) que lo que tienen las componentes 2 (0.03966) y 3 (0.01647); lo mismo sea dicho para las otras dos variables.

Esto se puede evidencia en la figura 2. a través de un sistema de color (azul), siendo, el más oscuro la calidad de la representación de la componente en la variable y más claro la menos calidad.

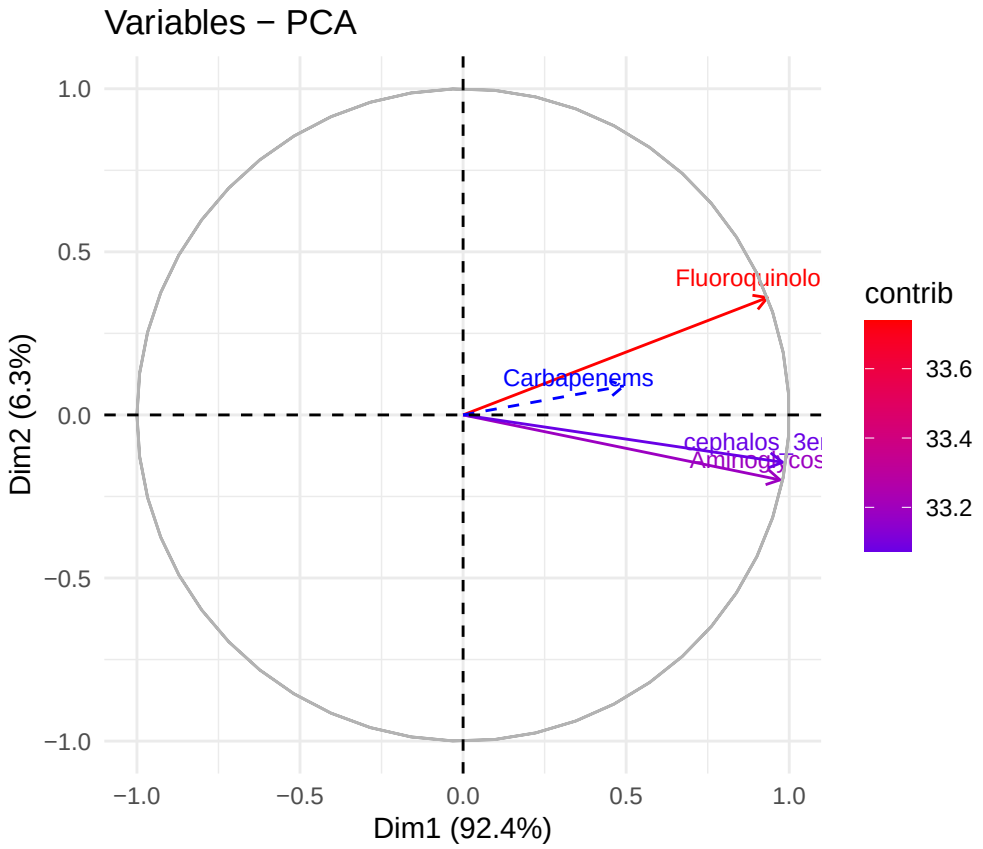

Figure 3: Gráfica de contribución de las variables en las componentes

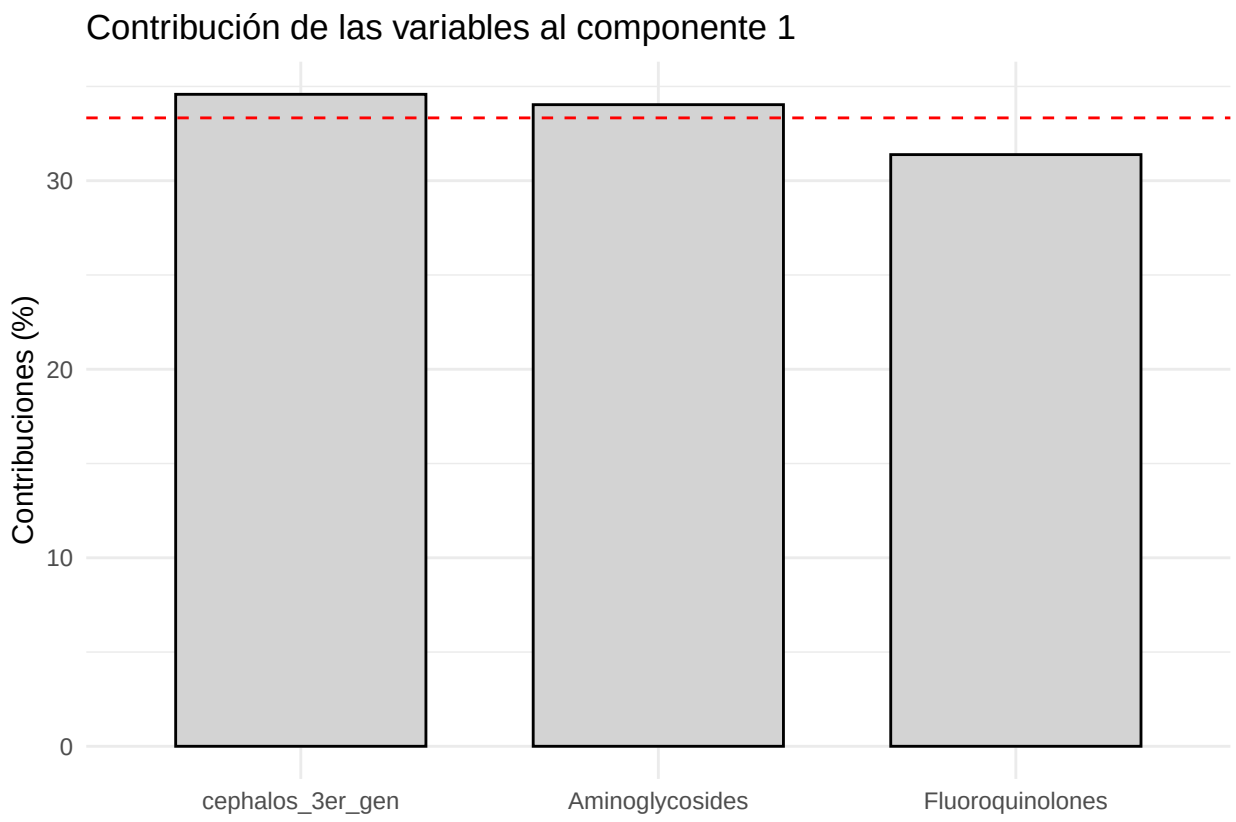

Figure 4: Contribuciones de las variables en cada la componente principa 1

Seguidamente, se realiza una gráfica de contribuciones de variables a las compomentes principales obtenidas. Esta información se encuentra en la tabla 4. La gráfica de contribuciones, figura 2. Como se puede observar, se realiza una discriminación por color dónde el color rojo indica la máxima contribución de la variable al componente relacionado y el colo azul es la menor. Si bien el vector de la variable `Fluoroquinolones` se ve de color rojo por su alta contribución no se hace al componente prinicipal 1, sino al 2. Aproximadamente el 68% del compomente 2 es consecuencia de la variable `Fluoroquinolones`. Mientras que su contribución en el compomente 1 es de 31.38%.

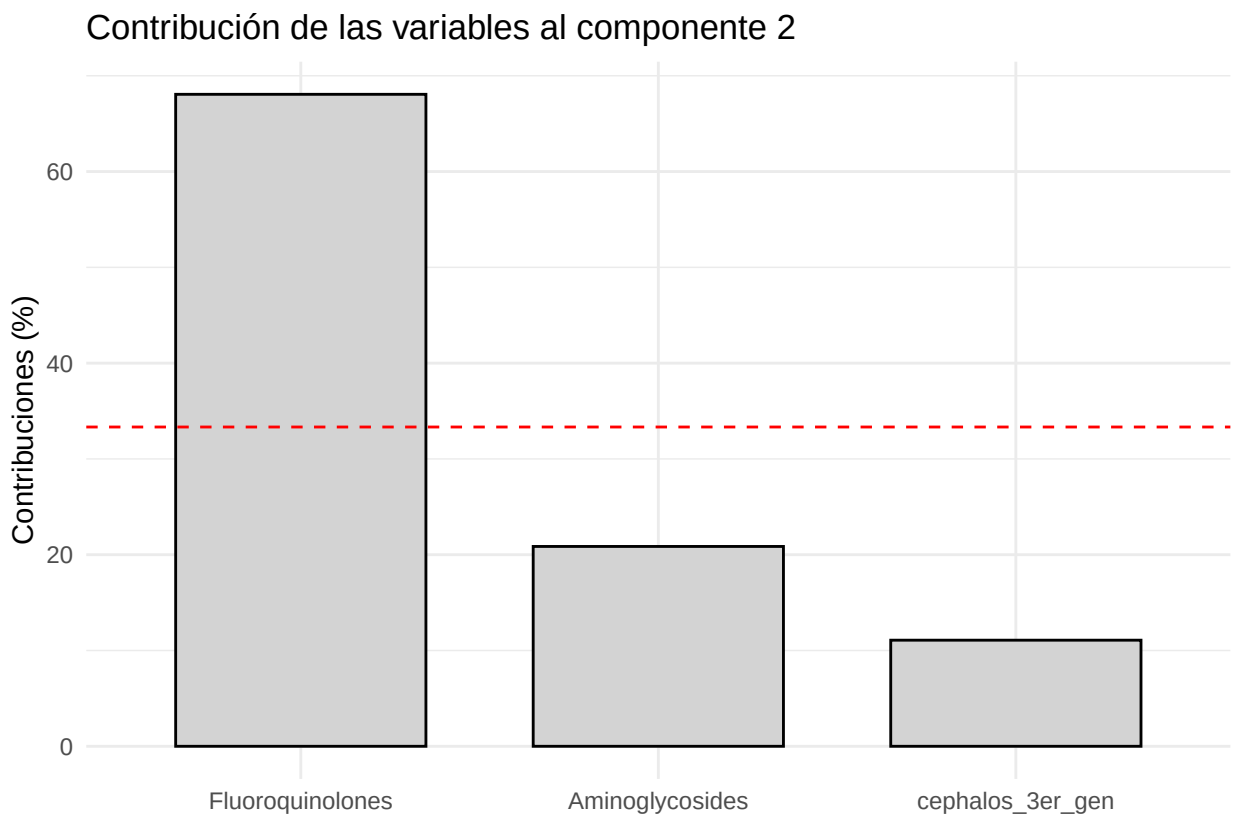

Figure 5: Contribuciones de las variables en cada la componente principa 2

Cabe resaltar también el vector de la variable suplementaria `Carbapenems` que como se predijo a través de los resultados preliminares, contribuye muy poco en las componentes principales resultantes. El peso de su contribución se da en color azul y la distancia del vector desde el centro del plano es corta lo que sugiere que contribuye poco a todas las compomentes, y su mayor contribución es al componente 1.

Las figuras 4, 5 y 6. corresponde a gráficas de barras que comparan el % de contribucion de cada una de las variables para cada una de las componentes principales. En cada una de estas gráficas se trazó una línea roja que corresponde al contribución promedio esperada. Esto indica que la variable que supere este valor umbrar contribuye más que de lo esperado por las variables originales.

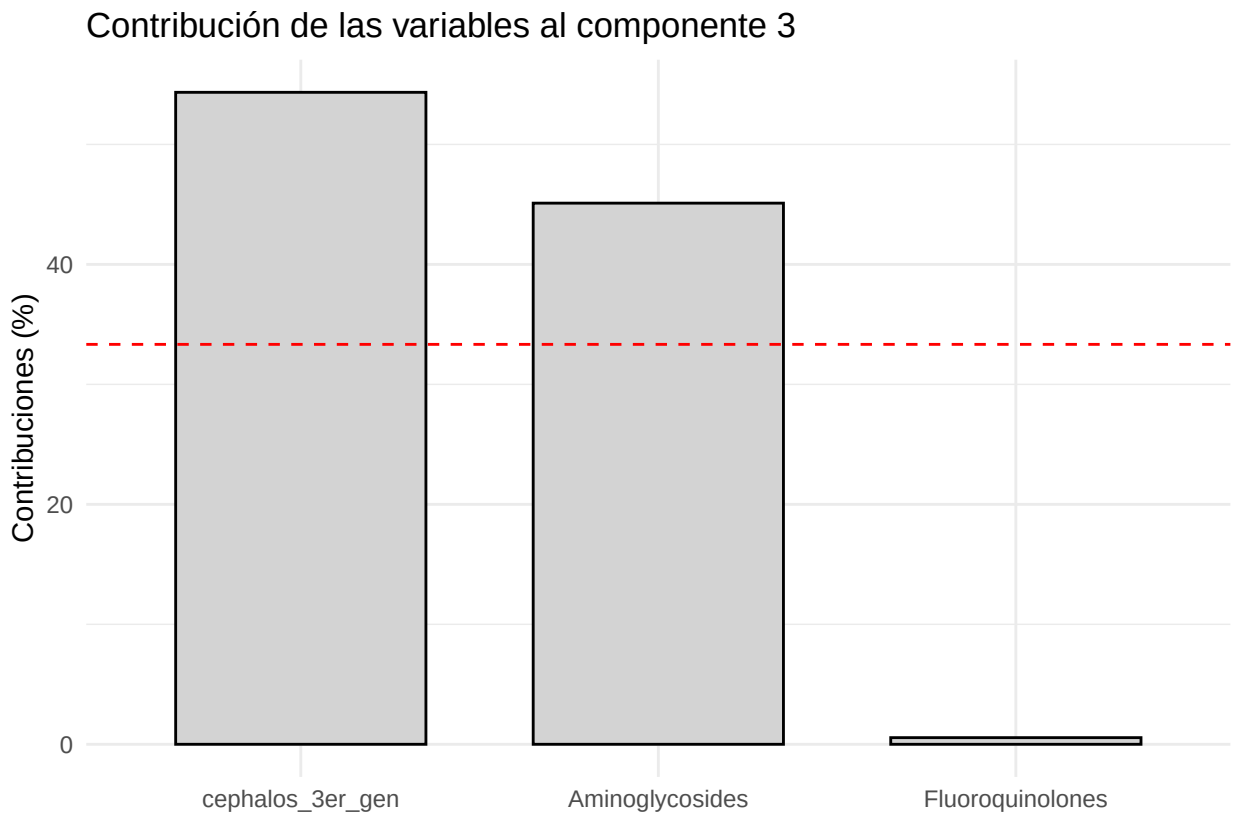

Figure 6: Contribuciones de las variables en cada la componente principa 3

Por su parte, la figura 4. muestra que las variables `cephalos_3er_gen` y `Aminoglycosides` son las que contribuyen en mayor medida a la componente principa 1 en una medida mayor que el la contribución esperada.

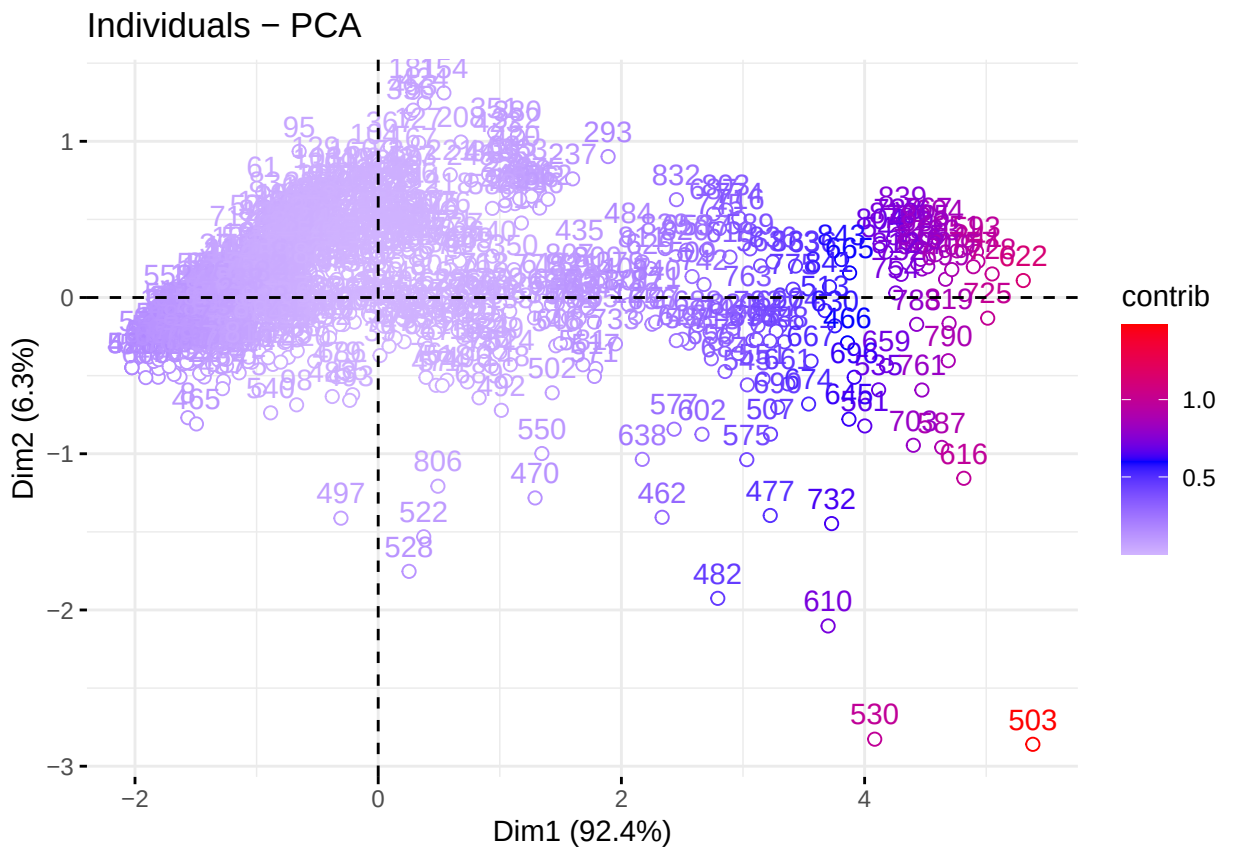

Figure 7: Gráfica de la distribución y contribuciones de las observaciones en las componentes principales

La figura 5. muestra que las variables `Fluoroquinolones` es la que contribuye en mayor medida a la componente principa 2 en una medida mayor que el la contribución esperada.

Finalmente la figura 6. muestra que las variables `cephalos_3er_gen` y `Aminoglycosides` son las que contribuyen en mayor medida a la componente principa 3 en una medida mayor que el la contribución esperada; sin embargo, esta componente no se puede obervar en las figuras 2 y 3, y tampoco es mandatorio retenerla en este análisis.

Se Adicionó una gráfica de contribución de las observaciones en las componentes principales (figura 7) en esta podemos obervar que la mayor parte de las observaciones se encuentra cercanas al centro del plano factorial. Sin embargo la dispersión de las observaciones tiende hacia la derecha, es decir hacia el plano positivo del eje y, lo que corresponde a la componente principal 1. Se realizó una discriminación por color de la contribución: las observaciones de color rojo son las de mayor contribución en las componentes y las de color más grisaceo son las de menos contribución con una contribución intermedia las de color azul. Como se puede ver las observaciones 530 y 503 (outsiders) son las que más aportan en la componente 1 de manera individual.

Análisis multivariado de APC

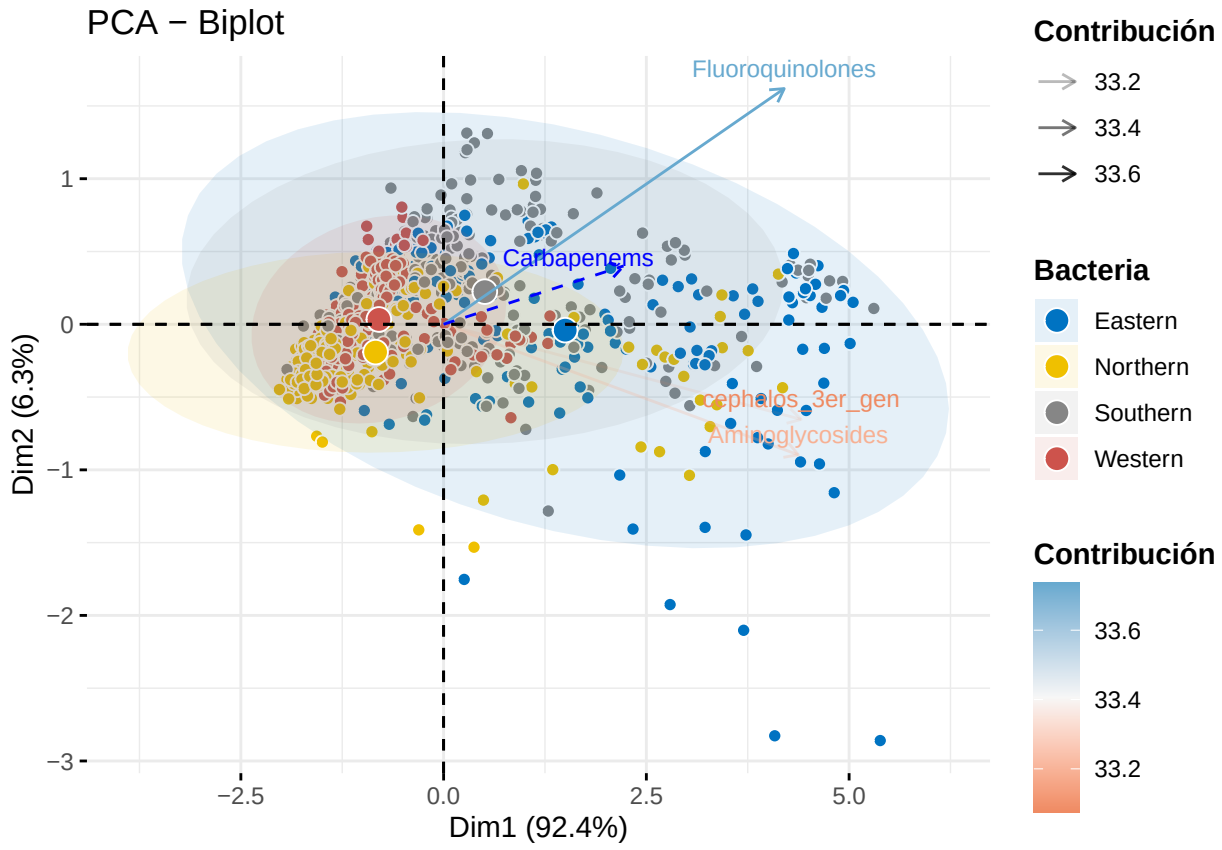

Figure 8: Gráfica de las componentes principales y la distribución de observaciones por region

Se realizaron dos análisis más que pueden encontrarse en las figuras 8 y 9. Estos corresponde a un gráfico tipo *Biplot* que permite hacer discriminaciones de las observaciones en le plano factoria introduciendo variables categóricas. Este análisis es particularmente interesante ya que se realizó una discriminación de las observaciones en el plano a través de especie de `Bacteria` (figura 9) y por `Region` (figura 8).

La figura 8, preve algunos resultados que se verán en el análisis de Clusters que se realizará más adelante. En este se encuentra que los paises del este europeo `Eastern`son los que más contribuyen en la componente principal 1. Aquí tambien se encuentran algunos países de la región sur de europa y el norte, sugieriendo que los países del oeste `Western` son los que menos contribuyen a la resistencia antibiótica. En especial los paises del este contribuye a la resistencia atibiótica de grupos como la Aminoglucosidos y las cefalosporinas de tercera generación, que son antibióticos última línea para el tratamiento de las infecciones producidas por estas dos bacterias.

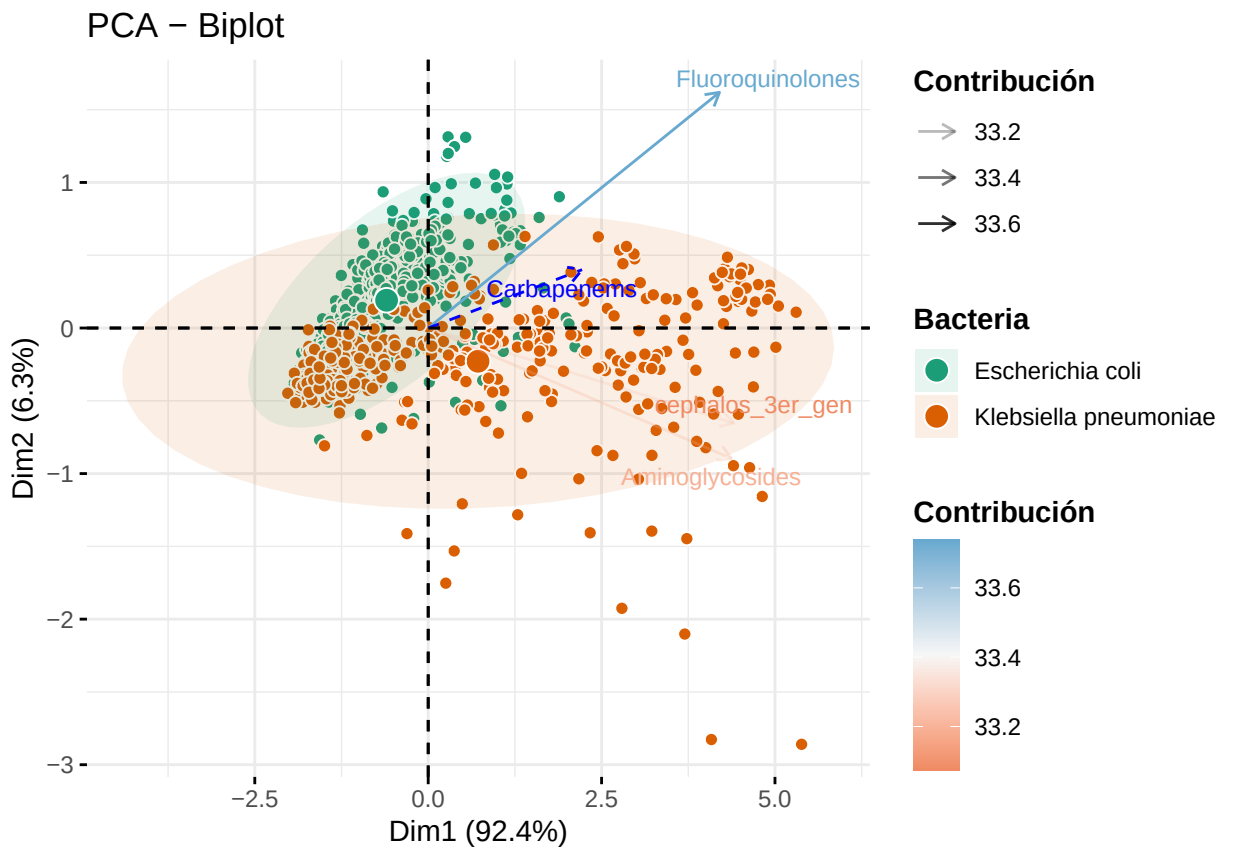

Figure 9: Gráfica de las componentes principales y la distribución de observaciones por especie de bacteria

Por último, la figura 9, muestra un la dispersión de las observaciones en le plano factoria y agrupadas por tipo de bacteria. Como se veía en el primer análisis realizado, la bacteria Klebsiella Pneumoniae es la que más contribuye a la resistencia antibiótica. Esto queda soportado, dado que se evidencia una amplia diferencia en la distribución de las observaciones que corresponden a esta en la componente principal 1, particularmente interesante es que contribuyen a la a resistencia atibiótica de grupos como la Aminoglucosidos y las cefalosporinas de tercera generación y que aunado con los resultados de la figura 8, se puede sugerir que el impacto mayor de esta resistencia es el los países del este de europa.