Aplicación del Análisis de Componentes Principales a Variables Económicas y Vinculadas al COVID-19

Análisis Multivariado, Proyecto de Mitad de Curso

Emanuelle Marsella, Maximiliano Saldaña

Junio, 2020

Índice

Resumen ejecutivo

En este proyecto se aplican las técnicas de Análisis de Componentes Principales (ACP) a una base de datos que recoge la información económica y sanitaria de 166 países para construir un Índice de Estímulo Económico, recreando el trabajo hecho en (Elgin, Nasbug y Yalaman, 2020) y a su vez para realizar un análisis exploratorio con el resto de variables de la base. Se logra recrear el Índice con éxito y mediante el análisis posterior se llega a una representación en el plano con alto nivel de variabilidad explicada, a través del cual se puede apreciar que aquellos países con mayor desarrollo económico tendieron a tener una tasa de infección más alta al 31 de marzo de 2020.

Introducción

El objetivo de este trabajo es aplicar las técnicas de Análisis Factorial a la base de datos que surgen del trabajo **Economic Policy Responses to a Pandemic: Developing the COVID-19 Economic Stimulus Index** (Ceyhun Elgin, Gokce Basbug, Abdullah Yalaman; 2020). En dicho artículo los autores bucan explicar cuáles fueron las medidas económicas que tomaron los distintos países del mundo frente a la pandemia de COVID-19. Elaboran un índice de estímulo económico a partir del Análisis de Componentes Principales (ACP) el cual buscan explicar a través de las demás variables de la base, que son de índole sanitario y demográfico. Como primer objetivo de nuestro trabajo se busca reproducir el índice creado por los autores a través de emplear la misma técnica que usaron, y posteriormente utilizar dicho índice en otro ACP que involucre las variables de índole sanitaria, económica y sociodemográfica, estudiando el comportamiento de los individuos y las variables y relación entre estas últimas.

Descripción de los datos

Los datos utilizados fueron elaborados por los autores del artículo de referencia a partir de una recopilación de distintas fuentes. La principal es el FMI (Fondo Monetario Internacional), pero algunos valores fueron reemplazados por otros provenientes de páginas gubernamentales o canales de noticias con el fin de que toda la información estuviese actualizada. La información es de la situación al 31 de marzo de 2020. Por nuestra parte incluimos además la variable densidad poblacional, con datos extraídos del Banco Mundial actualizados a 2018, al considerarse que la variable puede ser de interés en el análisis.

Observaciones y variables

Se cuenta con datos para 166 observaciones (países) y 14 variables. Estas últimas son:

Cuantitativas

- *fiscal*: paquete de políticas fiscales adoptadas, representado como porcentaje del PBI
- *ratecut*: representa el porcentaje que fue recortado de la tasa de interés en relación a la de febrero de 2020
- *macrofin*: tamaño del paquete de medidas macrofinancieras, expresado como porcentaje del PBI
- *bopgdp*: políticas de balanza de pago, expresadas como porcentaje del PBI
- *totalcases*: cantidad de casos totales de COVID-19
- *medage*: edad mediana en el país
- *infectionrate*: tasa de infección, representa la proporción de la población total infectada de COVID-19

- *hospitalbed*: camas de hospital por cada 1000 personas
- *healthexp*: gastos en salud, expresados como porcentaje del PBI
- *stringency*: índice de rigurosidad de respuesta gubernamental (desarrollado por Hale y Webster en 2020)
- *gdppercap*: PBI per cápita del país en dólares a niveles de 2010
- *CESI_INDEX*: índice de estímulo económico (construido por los autores utilizando análisis de componentes principles)
- *denspob*: densidad poblacional del país, medida como cantidad de personas por kilómetro cuadrado de área terrestre.

Cualitativas

- *othermonetary*: variable indicadora que indica si en el país se tomaron otras medidas monetarias
- *otherbop*: variable indicadora que indica si en el país se tomaron otras medidas de políticas de balanza de pago.

Datos faltantes

	Variable	Datos faltantes
1	totalcases	15
2	medage	20
3	gdppercap	23
4	healthexp	20
5	hospitalbed	19
6	stringency	93
7	infectionrate	22

Cuadro 1: Tabla de cantidad de observaciones faltables para las variables.

En el Cuadro 1 se aprecia que varias de las variables cuentan con datos faltantes, en particular las variables *total cases*, *medage*, *gdppercap*, *healthexp*, *hospitalbed*, *stringency* e *infectionrate*. Para poder trabajar con las técnicas de análisis factorial es necesario que contemos con todos los datos, por lo que se tiene que considerar un método para imputar los valores faltantes. Una alternativa posible para el caso de las variables cuantitativas es imputar el valor medio de la variable cuando no contemos con el valor de dicha variable para una observación, opción por la cual optamos por su simplicidad y al no ser el foco del presente trabajo. Para el caso de las variables cualitativas una alternativa de imputación es emplear el modo de la variable cada vez que haya un dato faltante, pero en este caso no es necesario ya que no hay datos faltantes en las variables cualitativas. Debe destacarse que la variable *stringency* cuenta con 93 valores faltantes en las 166 observaciones, por lo que se tomó la decisión de dejar esta variable de lado para no emplear una variable cuya mayoría de valores serían imputados, a pesar de que los autores la emplearon para su análisis.

Análisis de correlaciones

fiscal	ratecut	macrofin	bopgdp	totalcases	medage	gdppercap	healthexp	hospitalbed	infectionrate
1.00	0.16	0.32	-0.14	0.21	0.42	0.60	0.34	0.25	0.26
0.16	1.00	0.19	0.10	0.17	0.21	0.26	0.19	-0.03	-0.01
0.32	0.19	1.00	-0.05	0.23	0.39	0.36	0.20	0.18	0.15
-0.14	0.10	-0.05	1.00	-0.01	0.08	0.05	0.13	0.01	0.05
0.21	0.17	0.23	-0.01	1.00	0.27	0.28	0.41	0.11	0.21
0.42	0.21	0.39	0.08	0.27	1.00	0.60	0.51	0.71	0.26
0.60	0.26	0.36	0.05	0.28	0.60	1.00	0.45	0.34	0.44
0.34	0.19	0.20	0.13	0.41	0.51	0.45	1.00	0.34	0.22
0.25	-0.03	0.18	0.01	0.11	0.71	0.34	0.34	1.00	0.15
0.26	-0.01	0.15	0.05	0.21	0.26	0.44	0.22	0.15	1.00

Cuadro 2: Matriz de correlaciones

En el Cuadro 2 se presentan las correlaciones entre las variables de la base (el orden de las variables en las filas es el mismo que el de las columnas). Se destacan por ser altas las correlaciones entre *gdppercap* y *fiscal* (0.6), *medage* y *gdppercap* (0.6) y entre *medage* y *hospitalbed* (0.71). Al realizar el análisis de componentes principales se buscará obtener un nuevo conjunto de variables incorrelacionadas entre si y posiblemente reducir su número con respecto al conjunto original.

Observación preliminar de valores atípicos

Se realiza un sondeo de los valores atípicos de las variables, porque pueden afectar el ACP al confundirse las relaciones entre las variables. Esto se llama efecto tamaño, donde los valores atípicos causan que haya una estructura común al conjunto de variables que termina por ser expresada en la primer componente principal (Blanco, 2006).

Para buscar las posibles observaciones que se puedan considerar atípicas, se realiza un análisis exploratorio a partir de medidas de resumen aplicadas sobre las variables como el mínimo, el máximo, la media, la mediana y el rango intercuartílico. Además, se utilizan gráficos (*boxplots* e histogramas) y tablas.

País	Medidas de BOP (% PBI)
Algeria	6.00
Croatia	2.94
Switzerland	2.90
Brazil	1.69
Peru	0.90
Georgia	0.63
Colombia	0.43
Iran	0.33
Iceland	0.30
Argentina	0.22
Afghanistan	0.00

Cuadro 3: Tabla de medidas de balanza de pago como porcentaje de PBI, ordenada de forma decreciente.

Las medidas de balanza de pago adoptadas por los países se dividen en dos tipos, un primer tipo expresadas en la variable *bopgdp* como porcentaje del PBI, y de segundo tipo expresadas en la variable indicadora *otherbop* que toma el valor 1 si el país tomó medidas de ese tipo y 0 si no lo hizo. En el Cuadro 3 podemos ver que solo hay 10 países que tomaron medidas de balanza de pago de primer tipo, siendo Argelia el país

que efectuó mayor cantidad de estas medidas. Por otro lado, solo 32 países tomaron alguna otra medida de balanza de pagos, mientras que los restantes 134 no lo hicieron.

Argelia, que era la observación que tenía el mayor valor de la variable *bopgdp*, es también la observación que tiene el menor valor de la variable *fiscal*, con un porcentaje de medidas fiscales de -7.2. Existe además un conjunto de países que toman un valor de la variable por encima del límite máximo superior definido por el boxplot como $Q_3 + 1{,}5 \cdot RI$, donde RI es el rango intercuartílico. (Figura 4, Anexo 1)

Para la variable *ratecut* tanto el primer cuartil como la mediana coinciden con 0, lo cual indica una acumulación de países que toman ese valor, derivándose esto del hecho de que 95 países no experimentaron cambios en su tasa de interés respecto a febrero de 2020. Mientras tanto, 5 tuvieron un aumento de la misma y el resto una disminución. Hay un conjunto de 21 países que superan el límite extremo superior (calculado como un 40 % de disminución de la tasa de interés) y podrían ser considerados como atípicos. Entre estas observaciones se encuentran Reino Unido, Noruega, Croacia, Nueva Zelanda y Estados Unidos, siendo este último el único país que tuvo una reducción del 100 de su tasa de interés. (Figura 5, Anexo 1).

Para la variable *macrofin* también hay una gran cantidad de observaciones atípicas, destacándose en particular dos de ellas que se alejan incluso más que el resto, tomando valores próximos a 25, siendo que las siguientes observaciones más altas están en torno a 15. Estas dos observaciones con valores altos corresponden a los países Baréin y Oman. En cuanto al resto de los países, vemos una acumulación de observaciones con valores en torno al 0, valor que toman 102 países, más del 60 % del total. (Figura 6, Anexo 1).

Considerando la variable *bopgdp*, la mayoría de las observaciones tienen valor 0. Luego está Argelia que tiene el valor más alto próximo a 6, y las dos siguientes (con valores próximos a 3) son Croacia y Suiza.

En la variable *gdppercap* hay bastante dispersión como es de esperarse y un 75 % de los países cuentan con un país menor o igual a 16656.4 También se observa que hay un conjunto de países con valores atípicos de la variable, siendo la observación que toma el mayor valor es Luxemburgo, con un PBI per cápita de 110742 dólares que podemos atribuir al pequeño número de habitantes pero gran desarrollo económico de este país. (Figura 7, Anexo 1).

La variable medage toma un rango de valores entre 15 y 48, y muestra una gran simetría siendo que tanto la media como la mediana toman el valor de 31.52, lo cual atribuímos en parte a la imputación en esta variable utilizando la media. El país con menor edad mediana es Niger con un valor de 15 años, y el país con mayor edad mediana es Japan con 48 años. (Figura 8, Anexo 1).

En la variable *healthexp*, se destaca una observación atípica que corresponde a Estados Unidos con un gasto en salud del 17.07 % del PBI, casi el doble que el país que le sigue, Reino Unido con un valor de 9.76. Al igual que con *medage* nuevamente los datos son simétricos, ya que la media coincide con la mediana, y lo atribumos al mismo motivo. (Figura 9, Anexo 1).

Considerando la variable *hospitalbed*, observamos nuevamente cierta simetría ya que la media toma el valor de 3.01 y la mediana toma el valor de 2.8. Se aprecia además la presencia de algunas observaciones atípicas, siendo Japón la que toma el mayor valor con un número de 13.4 camas de hospital por cada 1000 habitantes. (Figura 10, Anexo 1).

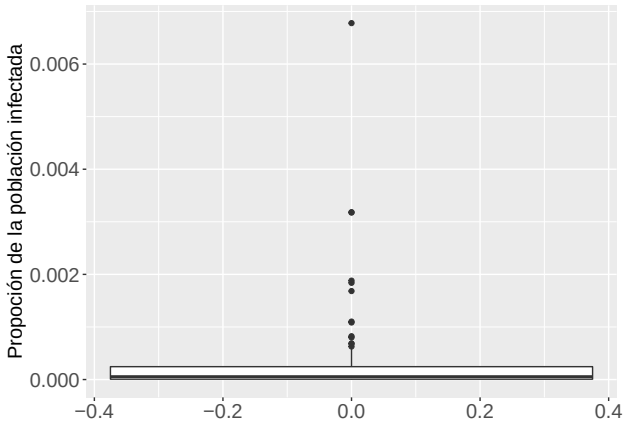

Figura 1: Gráfico de caja de la tasa de infección de los países.

En cuanto a la tasa de infección respecto a la población (*infectionrate*), se observa en la Figura 1 que hay una gran acumulación de tasas de infección en torno a valores próximos a 0, y un 75 % de las tasas son menores a 0.00025 (que representa un caso por cada 4000 habitantes aproximadamente). Se distinguen especialmente del resto de los países San Marino, Islandia y Luxemburgo. Aquí puede estar entrando en juego que dichos países cuentan con poblaciones muy reducidas, acumuladas en ciudades y en el caso de San Marino y Luxemburgo, limítrofes con países que también contaban con tasas de infección altas (el primero con Italia y el segundo con Bélgica y Alemania). En particular, San Marino tiene una tasa de infección de 0.0068 que representa un caso por cada 150 habitantes aproximadamente.

Aplicación del Análisis de Componentes Principales

Partimos de una matriz de datos centrados X, una matriz diagonal:

$$D = \begin{pmatrix} 1/I & 0 & 0 & \dots & 0 \\ 0 & 1/I & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1/I \end{pmatrix}$$

y una matriz diagonal igual a los inversos de las variables:

$$M = \begin{pmatrix} 1/s_1^2 & 0 & 0 & \dots & 0 \\ 0 & 1/s_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1/s_J^2 \end{pmatrix}$$

.

Al realizar el Análisis de Componentes Principales desde la perspectiva de la nube de variables, la matriz D representa la métrica utilizada en el espacio vectorial euclidiano, y M representa los pesos asociados a las variables. Utilizar esta métrica tiene la interpretación de que se están estandarizando las variables al dividirlas entre sus varianzas.

Se proyecta el conjunto de variables en el subespacio W de dimensión $k \leq I$, de forma tal de minimizar la inercia del subespacio, ya que esta representa la variabilidad de la nube de puntos original que no queda explicada por el subespacio. Esto se logra representando esta nueva nube de puntos en un conjunto de ejes $(\Delta_1, ..., \Delta_I)$ que se construye como combinación lineal de los vectores de una base ortonormal del subespacio, W, formada por los vectores propios asociados a la matriz $X^T DXM$. Las componentes principales son los

conjuntos de coeficientes que permiten expresar a las variables como combinación lineal de los vectores propios ortonormales de esa base. La inercia del subespacio se minimiza eligiendo una base ortononormal conformada por los k vectores propios asociados a los k valores propios de mayor magnitud.

De esto, se desprende claramente que cada coeficiente representa la correlación entre la variable asociada y el eje sobre el que se está proyectando, dado que si el valor de la variable aumenta en una unidad, el valor en el eje aumenta en una cantidad proporcional al coeficiente.

El análisis desde la perspectiva de los individuos es análogo al anterior, solo que la matriz M representa la métrica y la matriz D representa los pesos asociados a los individuos, y los factores son los vectores propios asociados a la matriz XMX^TD.

Construcción del Índice de Estímulo Económico

Inicialmente, intentaremos replicar el índice de estímulo económico elaborado por los autores del artículo. Para esto, realizamos un análisis de componentes principales únicamente con las variables de índole económico *fiscal*, *ratecut*, *macrofin*, *bopgdp*, *othermonetary* y *otherbop*. Si bien *othermonetary* y *otherbop* son variables cualitativas al ser indicadoras resulta adecuado incluirlas en el análisis al tener su media una interpretación clara; es la proporción de países que que toman el valor 1 en estas variables (tomaron medidas de este tipo).

Se pudo comprobar que el Índice de Estímulo Económico que se obtuvo realizando el ACP y definiéndolo como los valores de la primera componente principal es equivalente al obtenido por los autores. Esto nos lleva a pensar que el procedimiento de imputación de valores faltantes fue el mismo, la imputación de la media de las variables. Sin embargo, observamos que las coordenadas de las variables no coinciden con las nuestras, y pudimos observar que esto se debe a que el índice que presentan los autores es el resultado de dividir el que hallamos nosotros por la raíz cuadrada del vector propio asociado, lo cual se puede deber a que utilizaron una metodología de estandarización o software distinto al nuestro.

Al realizar el ACP observamos que la primera componente principal logra explicar un 27.13 % de la variabilidad total de la nube de puntos, la segunda un 19.76 % y la tercera un 18.04 %. En nuestra opinión, el porcentaje de inercia explicado por el primer eje, que es el que los autores toman para definir el índice de estímulo económico (*CESI_INDEX*) es bajo; y consideramos más adecuado caracterizar la nube de puntos con dos o tres dimensiones, logrando así explicar el 46.89 % o el 64.93 % la inercia, respectivamente. Sin embargo, se debe tener en cuenta que el objetivo de los autores del artículo al elaborar este índice es modelizarlo con un modelo lineal utilizando a las demás variables de la base como variables explicativas, lo cual justifica que optaran por mantener una sola dimensión. En nuestro caso, como nuestro objetivo es utilizar dicho índice como variable en un análisis de componentes principales y realizar un estudio similar al de los autores del artículo, pero a través de un segundo ACP en lugar de un modelo lineal, optaremos también por conservar solo la primera componente principal en el análisis y evaluar su desempeño.

Variable	Correlación
fiscal	0.74
ratecut	0.50
macrofin	0.71
bopgdp	-0.21
othermonetary	0.50
otherbop	-0.21

Cuadro 4: Correlación de cada variable con el índice de estímulo económico.

Podemos observar en el Cuadro 4 la correlación de cada una de las variables utilizadas en el análisis de componentes principales con el índice de estímulo económico definido. Las correlaciones, que nos permiten medir la calidad de representación de cada variable en el primer eje, resultan ser altas y con signo positivo para las variables *fiscal* y *macrofin*, lo cual significa que las medidas de tipo fiscal y macrofinanciero tienen un impacto positivo en el índice de estímulo económico construido.

Las variables *ratecut* y *othermonetary* tienen correlaciones menores a las variables previamente mencionadas, y de signo positivo, lo cual parece indicar que las otras medidas de tipo monetaria, así como también las medidas de reducción de tasa de interés, también tienen un impacto positivo en el índice de estímulo económico, si bien la magnitud de las correlaciones no son lo suficientemente altas como para poder afirmarlo.

Por su parte, las variables asociadas a medidas de políticas de balanza de pago *bopgdp* y *otherbop* tienen correlaciones bajas, lo cual significa que no influyen demasiado en la construcción del índice de estímulo económico y por lo tanto no están bien representadas por el mismo.

Análisis de Componentes Principales considerando el Índice de Estímulo Económico

A continuación, se procede a emplear esta nueva variable en un análisis de componentes principales. Como primer paso, creamos a partir de *infectionrate* la variable *noinfectionrate*, para indicar la proporción de habitantes del país no infectados en lugar de los infectados. Esto lo hacemos con el fin de facilitar la interpretación de la variable, dado que un valor elevado de esta variable representa en realidad una situación negativa de la situación del país frente a la pandemia en lugar de positiva. Por lo tanto, trabajamos con la tasa de no infección que sí representa una situación positiva, lo que consideramos que puede facilitar la lectura de los gráficos y la interpretación de lo que representa la dirección de esta variable en el subespacio de dimensión reducida respecto a la de las demás variables, así como también la interpretación de las coordenadas de la variable en los distintos ejes.

El conjunto de variables empleado en este análisis esta conformado por *medage*, *gdppercap*, *healthexp*, *hospitalbed*, *noinfectionrate*. La variable *CESI_INDEX*, el índice de estímulo económico obtenido a partir del ACP es utilizado como variable suplementaria en el análisis, y tampoco se incluyen las variables empleadas para construirlo al ser la intención utilizar el índice como su representante. Inicialmente había sido tenida en cuenta para la construcción de las componentes, pero al ser la variable peor representada en el plano principal y comportarse de forma similar a otras variables explicadas mayoritariamente por el segundo eje del análisis (en especial como *gdppercap*) se tomó como alternativa estudiar el comportamiento de su proyección en el espacio construido a partir de las demás variables para ver su interacción con ellas.

Además, también es excluida la variable que indica la cantidad de casos totales de COVID-19 (*totalcases*), ya que la información ya se expresa en la variable *noinfectionrate* y al estar medida en términos absolutos dificulta la comparación entre observaciones. Adicionalmente se hizo la prueba de incluir una variable que expresaba la densidad poblacional de los países, pero se concluyó que la misma no aportaba al análisis. Estos resultados se encuentran en el Anexo 2.

	Valor propio	% de varianza	% acumulado de varianza
comp 1	2.67	53.32	53.32
comp 2	0.97	19.45	72.77
comp 3	0.67	13.49	86.26
comp 4	0.48	9.56	95.81
comp 5	0.21	4.19	100.00

Cuadro 5: Valores propios asociados a las componentes principales y porcentajes de inercia de cada uno.

En el Cuadro 5 podemos ver que la primer componente principal logra explicar un 53.32 % de la variabilidad de la nube de puntos, y la segunda componente principal un 19.45 %. Esto significa que con el primer plano principal logramos explicar prácticamente un 72.77 % de la variabilidad total, una cantidad importante a nuestro juicio dado que reducimos la cantidad de dimensiones a dos, siendo 5 el número original de variables en la matriz de datos. Por este motivo y por el hecho de que permite una representación e interpretación gráfica clara se opta por trabajar con estos dos ejes.

Perspectiva de la Nube de Individuos

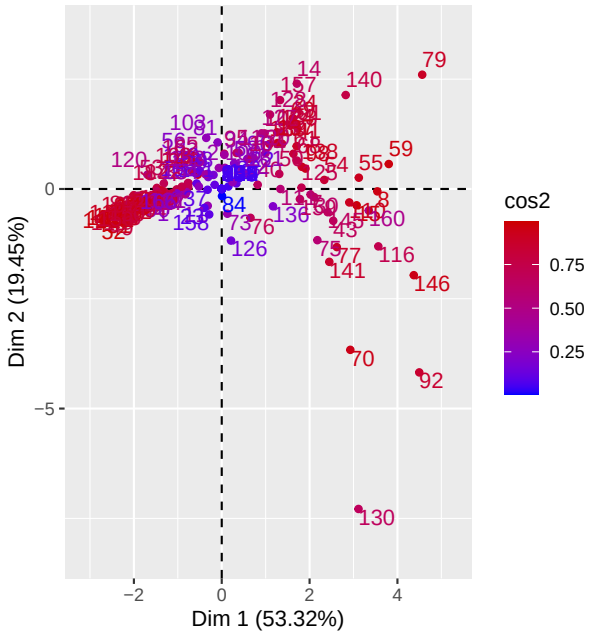

Figura 2: Proyecciones de los individuos en el primer plano principal.

En la Figura 2, podemos observar las proyecciones de los individuos en el primer plano principal, los cuales se encuentran coloreados en función del $cos^2(\theta)$, es decir, el coseno al cuadrado del ángulo que forman los vectores originales de los individuos con este plano. Un $cos^2(\theta)$ próximo a 0 significa que el ángulo θ está próximo a un múltiplo de $\pi/2$, es decir, es casi ortogonal al plano, y en consecuencia el individuo no se encuentra bien explicado por el mismo. Por otro lado, un $cos^2(\theta)$ próximo a 1 signfica que el vector original es casi colineal con el plano y que por lo tanto se encuentra bien explicado, y la proyección es buena. Es deseable que ninguno de los individuos sea especialmente influyente a la hora de generar los componentes principales. Un punto resulta influyente cuando es lejano al baricentro y además el $cos^2(\theta)$ es próximo a 0, lo que puede tener como consecuencia una disminución en la inercia explicada por el subespacio considerado. A primera vista la Figura 2 indicaría que esta situación no se da, puesto que hay un conjunto de puntos con $cos^2(\theta)$ próximos a 0 pero se encuentran cerca del baricentro; mientras que también hay un conjunto de puntos que se encuentran más apartados del baricentro, pero con $cos^2(\theta)$ relativamente altos.

Gráficamente se puede notar que el primer eje tiende a dar menores valores a países de menor desarrollo económico y mayores a países de menor desarrollo económico (que se corresponden con valores mayores en el conjunto de variables que se usaron para caracterizarlos). Mientras tanto el segundo eje tiende a dar valores menores a países con mayor tasa de infección y mayores a países con valores menores de dicha variable.

	País	cos2_plano	norma
130	San Marino	0.67	9.69
92	Luxembourg	0.85	6.64
79	Japan	0.90	5.55
160	United States	0.47	4.94
146	Switzerland	0.95	4.92
70	Iceland	0.94	4.83
116	Norway	0.73	4.46
140	South Korea	0.77	4.05
59	Germany	0.97	3.91
14	Belarus	0.64	3.69

Cuadro 6: Coseno al cuadrado con el plano principal y norma de las observaciones (10 países con mayores valores de la norma)

En el Cuadro 6 se presentan los valores de la suma de los cosenos al cuadrado de los ángulos que forman las observaciones con los dos primeros ejes y la distancia respecto al baricentro, ordenando las observaciones decrecientemente según esta última. Se aprecia que la observación correspondiente a San Marino (que ya se había identificado como atípica en la descripción de los datos) tiene la mayor norma, con un valor de 9.69 y una suma de cosenos al cuadrado de 0.67. Se debe tener en cuenta que podría estar afectando el análisis teniendo en cuenta que su calidad de representación en el primer plano no es muy alta y su norma es particularmente alta. Efectuar el ACP considerando la observación influyente como un dato suplementario representa una ganancia en inercia explicada en el plano del 4 % aproximadamente (queda en 76.63 %), por lo que se toma la decisión de considerarla como una observación suplementaria en el resto del análisis para tener una mejor descripción de las variables/individuos en las dimensiones elegidas. El resto de observaciones con norma alta cuentan con valores del coseno al cuadrado en general mayores a 0.6 por lo que no se pensarían influyentes. Estados Unidos tiene un valor de este indicador de la calidad de representación de 0.47, pero incluirla como complementaria no supone un aumento de la inercia explicada considerable, por lo que no se toma como atípica influyente.

	País	cos2_plano	norma
84	Kuwait	0.01	1.64
10	Bahamas	0.06	0.65
21	Botswana	0.06	0.00
26	Burundi	0.06	0.00
27	Cabo Verde	0.06	0.00
36	Congo, R	0.06	0.00
38	Cote Ivory	0.06	0.00
45	Ecuador	0.06	0.00
51	Eswatini	0.06	0.00
64	Guinea Bissau	0.06	0.00

Cuadro 7: Coseno al cuadrado del ángulo con el plano principal y norma de las observaciones (10 países con mayores valores del coseno cuadrado)

En el Cuadro 7 anterior se presentan las 10 observaciones con la menor calidad de representación en el plano. Se destaca que practicamente todos son países en vías de desarrollo. No obstante, todos estos países se encuentran muy cercanos al baricentro, por lo que su influencia es limitada, simplemente es dificil distinguirlos del individuo medio. En un principio se podría pensar que es un problema de representación de este tipo de países en general, pero al identificar los valores del $cos^2(\theta)$ del resto de las observaciones en el plano podemos ver que a su vez hay países en desarrollo bien representados y países desarrollados con mala representación.

Perspectiva de la Nube de Variables

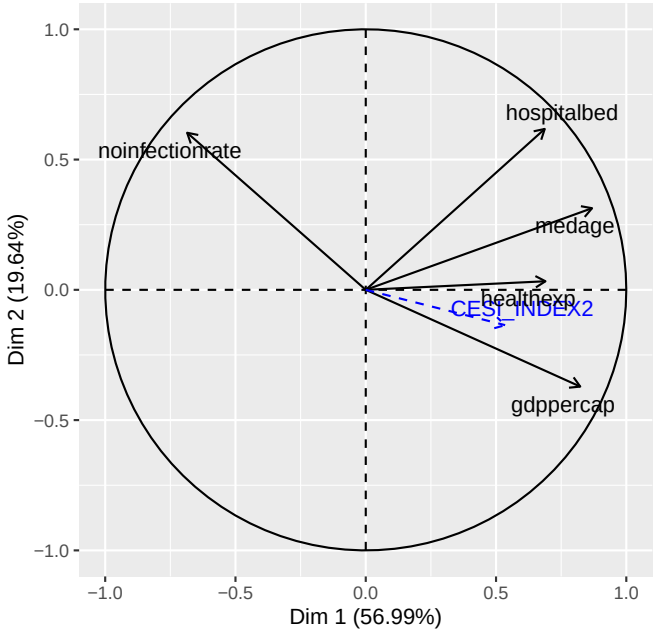

Figura 3: Proyecciones de las variables utilizadas para el ACP en el primer eje principal, y el índice de estímulo económico construido como variable suplementaria.

La Figura 3 permite ver representados los vectores asociados a las variables proyectadas en el primer plano principal, teniendo en cuenta que el largo del vector en relación al círculo unitario indica qué tan bien está representada esa variable en el primer plano principal y el valor sobre los ejes su correlación con la dimensión que representa cada uno. Podemos ver que en general los vectores tienen una longitud relativamente cercana al círculo, destacándose las variables *medage*, *hospitalbed*, *noinfectionrate* y *gdppercap*. A excepción de *healthexp*, notamos que todas las variables necesitan de ambos ejes para poder estar bien representadas.

Por otro lado, vemos que todas las variables apuntan en la misma dirección del primer componente principal a excepción de *noinfectionrate* que apunta en sentido contrario. Esto quiere decir que un valor mayor en *noinfectionrate* se expresa con una disminución del valor de la primera componente, mientras que un valor mayor en cualquiera de las demás variables se expresa con un aumento en dicha componente. Esto implica que aquellos países con menores tasas de infección van a tender ser a su vez los que tengan menores valores de las demás variables. La segunda componente explica principalmente a *noinfectionrate* y *hospitalbed* y observando las direcciones de las proyecciones de las variables un aumento en cualquiera de las dos variables se expresa en un aumento en el valor de esta componente.

	Dim.1	Dim.2
medage	0.87	0.31
gdppercap	0.82	-0.37
healthexp	0.69	0.03
hospitalbed	0.69	0.62
noinfectionrate	-0.69	0.60

Cuadro 8: Coordenadas de las variables originales en el primer plano principal

En el Cuadro 8 se puede ver que las variables *medage* y *gdppercap* tienen una correlación alta (entre 0.8 y 0.9) con la primera componente principal, lo cual ya se identificaba en el gráfico anterior dado que los

vectores tienen una longitud considerable y apuntan en la dirección de esta componente. Sucede lo contrario con *noinfectionrate*, dado que tiene una correlación negativa con la primer componente principal lo cual se identificaba en la dirección del vector.

Las variables *medage* y *gdppercap* tienen correlaciones similares con la componente 2, aunque una con signo positivo y la otra con signo negativo. Esto es fácil de ver en el gráfico, dado que los dos vectores tienen longitud similares y ángulos opuestos respecto a este eje. Las variables *hospitalbed* y *noinfectionrate* tienen correlaciones similares en magnitud con cada uno de los ejes, a excepción del signo de la correlación con el primer eje (en el caso de *noinfectionrate* es negativa).

Las correlaciones de las demás variables con la segunda componente son más bajas que con la primera, lo cual significa que no representa esas variables tan bien como sí lo hace el otro eje, que se podría asociar como representante de características del país (económicas y respecto al sistema de salud).

Considerando ahora la variable *CESI_INDEX* incluida como suplementaria, las correlaciones del índice de estímulo económico con los dos ejes principales, son 0.53 y -0.14, con el primer y segundo eje respectivamente. Esto significa que dicha variable se ve más relacionada con el eje que a su vez representa mejor a las variables *medage*, *gdppercap* y *healthexp*. Aún así debe ser tenido en cuenta que la correlación no resulta muy alta con dicho eje, lo que compromete la interpretación de la variable en el subespacio considerado.

Variable	Correlación
medage	0.49
gdppercap	0.58
healthexp	0.31
hospitalbed	0.22
noinfectionrate	-0.22

Cuadro 9: Correlación del índice de estímulo económico construido con las variables originales de la base.

Las correlaciones de *CESI_INDEX* más altas con las variables originales son de 0.58 con *gdppercap* y 0.49 con *medage* (Cuadro 9). Esto puede ser interpretado como que los países que tienen un mayor desarrollo tuvieron la capacidad de realizar un mayor gasto de estímulo económico, si bien esta apreciación debe ser tomada con cautela debido a que las correlaciones no son demasiado altas.

Como resultado a destacar de este análisis, se llegó a que el segundo eje es el que caracteriza principalmente a las variable *noinfectionrate* y *hospitalbed*; mientras que el primer eje caracteriza sobre todo a las demás variables, y también a las anteriormente mencionadas. Además, observando cómo se comporta la variable *CESI_INDEX* al proyectarla sobre el primer plano principal, vemos que se correlaciona principalmente con el eje que describe a las variables de características del país.

Conclusiones

Se logró recrear con éxito el Índice de Éstimulo Económico de los países generado en el artículo de referencia *(Elgin, Basbug y Yalaman, 2020)*, encontrándose únicamente una diferencia de estandarización en los factores de las variables. Al proceder con el objetivo de realizar un ACP considerando el índice y un conjunto de las variables de los datos se consideró pertinente íncluir el índice como variable suplementaria en lugar de considerarla dentro de las variables activas, debido a su pobre representación en los primeros ejes. Al analizar la nube de individuos proyectada se excluyó la observación correspondiente a San Marino al ser atípica y resultar influyente, lográndose así una ganancia del 4 % de inercia explicada en el primer plano aproximadamente. Con el plano principal se logró inicialmente una inercia explicada del 72.77 % por lo que se optó por trabajar en el análisis con las primeras dos dimensiones, y posteriormente al excluir a San Marino de las observaciones activas se llega a que los primeros dos ejes explican un 76.63 % de la inercia total, por lo que se conserva esta configuración del ACP. Si bien todas las variables necesitan de ambas componentes consideradas para ser explicadas, la primera explica más el conjunto de las variables que

expresan características económicas, del sistema de salud y demográficas; la segunda la tasa de no infección y la cantidad de camas de hospital cada 1000 habitantes. La proyección del índice de estímulo se comporta como el primer grupo de variables, aunque su correlación con la primer componente no resulta muy elevada en comparación con las variables que sí formaron parte del ACP. En la representación de las variables en el primer plano principal se puede observar que la dirección de la tasa de no infección es opuesta a la de las variables socioeconómicas, incluyendo el índice de estímulo económico. Es decir, los países con una mayor población infectada fueron aquellos que cuentan con un mayor PBI per cápita, una mayor edad mediana, un mayor número de camas de hospital, y que tomaron un mayor número de medidas económicas. La conclusión de esto es, a grandes rasgos, que los países con un mayor nivel de desarrollo socioeconómico fueron aquellos que presentaron altas tasas de infección. Este resultado, que puede resultar poco intuitivo, en realidad se debe a que los datos utilizados son al 31 de marzo de 2020, cuando muchos de los países desarrollados se enfrentaban a una ola de infecciones. En consecuencia, vemos la dirección similar del índice de estímulo económico con las variables sociodemográfica: países más desarrollados, que fueron de los primeros en infectarse, tendieron a ser los que se vieron forzados a tomar medidas de estímulo económico para intentar enfrentarse a los estragos de la pandemia.

Bibliografía

- Jorge Blanco. (2006). *Introducción al Análisis Multivariado*. Montevideo: IESTA.
- C. Elgin, G. Basbug y A. Salaman. (2020). *Economic Policy Responses to a Pandemic: Developing the COVID-19 Economic Stimulus Index*. CEPR Press.
- Marco Scavino. (2020). *Análisis Multivariado I - Análisis Factorial de una nube de puntos en un espacio vectorial euclidiano*. Montevideo: IESTA.

Anexos

Anexo 1: Gráficos para el análisis de observaciones atípicas

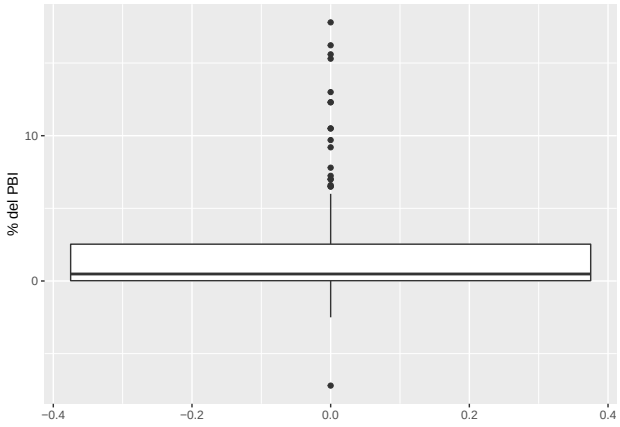

Figura 4: Gráfico de caja de medidas fiscales.

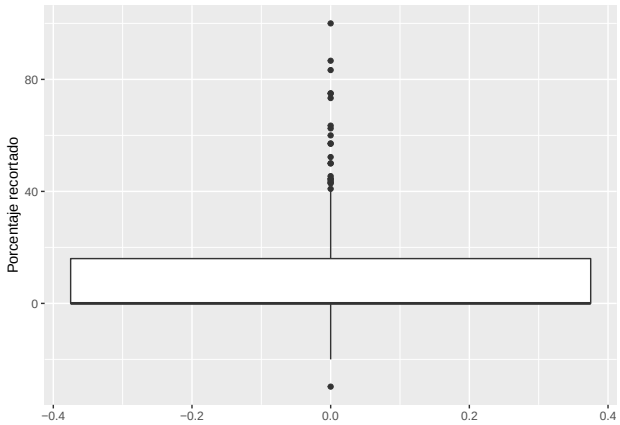

Figura 5: Gráfico de caja del recorte de la tasa de interés respecto a la de febrero de 2020.

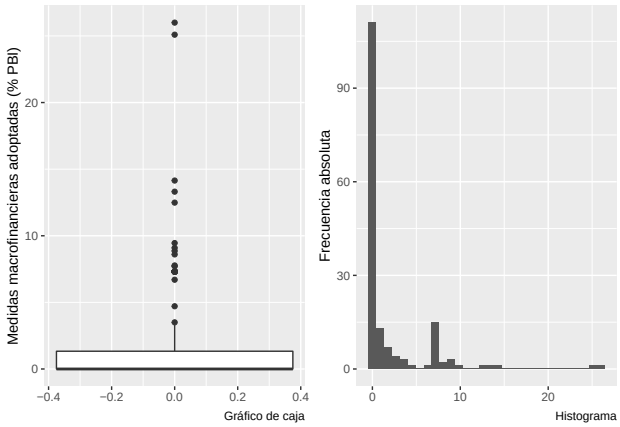

Figura 6: Gráficos para las medidas macrofinancieras adoptadas.

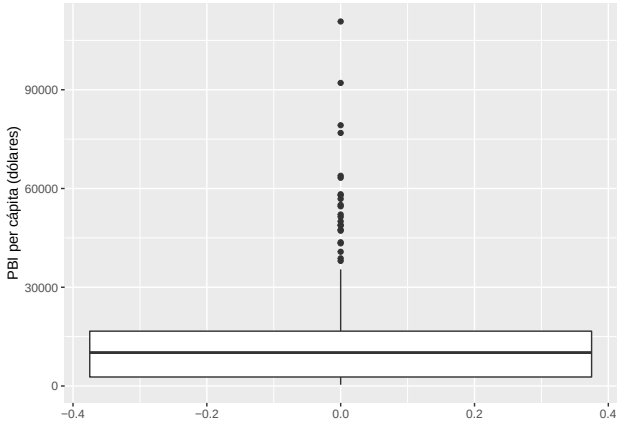

Figura 7: Gráfico de caja del PBI per cápita de los países.

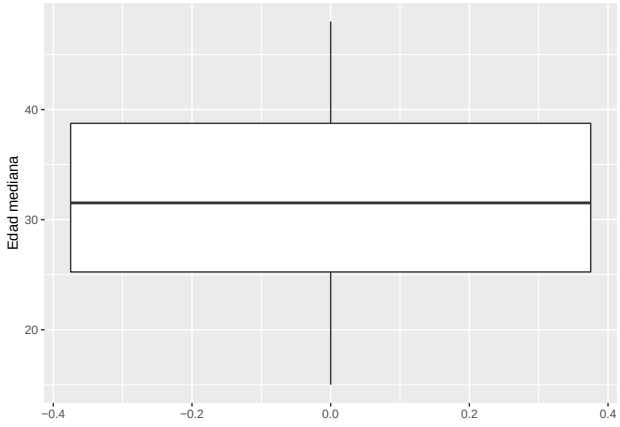

Figura 8: Gráfico de caja para la edad mediana de los países.

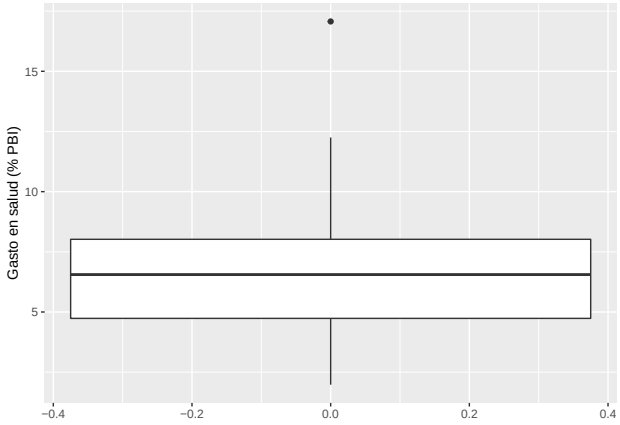

Figura 9: Gráfico de caja para el gasto en salud de los países.

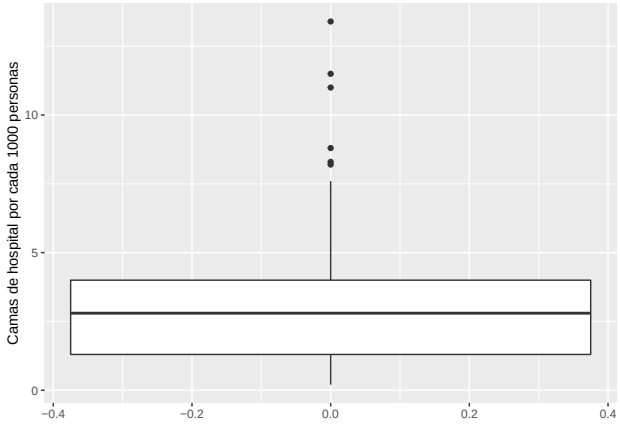

Figura 10: Gráfico de caja de la cantidad de camas de hospital en los países.

Anexo 2: Inclusión de la variable densidad poblacional

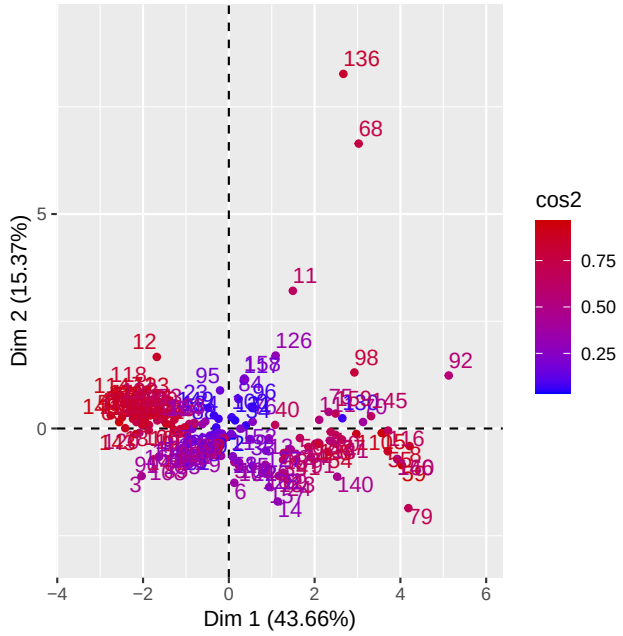

Figura 11: Proyecciones de los individuos en el primer plano principal utilizando denspob.

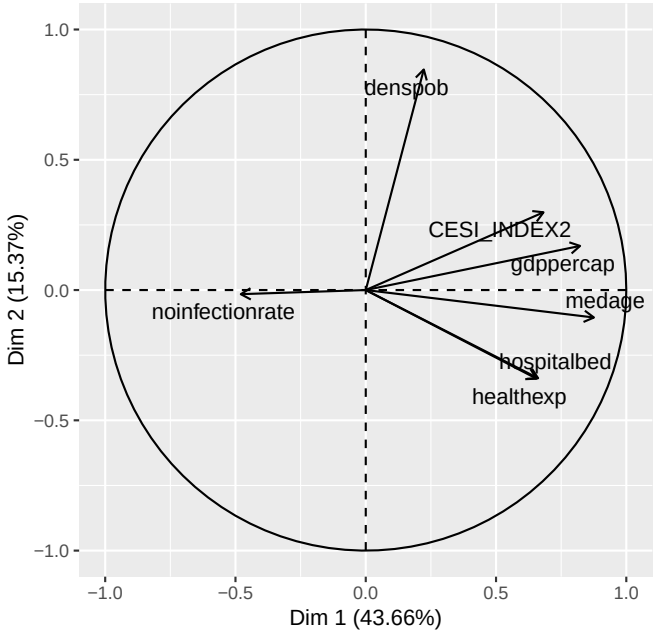

Figura 12: Proyecciones de las variables en el primer plano principal utilizando denspob.

	Dim.1	Dim.2
medage	0.88	-0.11
gdppercap	0.82	0.17
healthexp	0.66	-0.34
hospitalbed	0.66	-0.34
noinfectionrate	-0.48	-0.02
CESI_INDEX2	0.68	0.30
denspob	0.22	0.85

Cuadro 10: Coordenadas de las variables en los ejes del primer plano principal.

Podemos observar en las Figuras 11 y 12 que al realizar un Análisis de Componentes Principales incluyendo la variable densidad poblacional, solo logramos explicar con el primer plano principal un 50.03 % de la inercia total de la nube de puntos, bastante inferior que el 72.77 % de inercia explicada por el plano principal del ACP que finalmente optamos por mantener.

En la Figura 12 Vemos que el segundo eje explica principalmente la variable *denspob*, teniendo con la misma una correlación de 0.847, mientras que el resto de las variables tienen una correlación baja con ese eje, siendo -0.340 la segunda mayor en términos absolutos. Por otro lado, el primer eje es en el plano principal el que explica en mayor medida a todas las demás variables, incluyendo a *noinfectionrate*, con una correlación baja en magnitud de tan solo -0.479.

La conclusión que obtenemos de esto es que el segundo eje se incluye principalmente para explicar a la variable *denspob*, que fue incluída por nuestra parte porque creímos que podía resultar valiosa en el análisis al presuntamente relacionarse con la tasa de infecciónn, lo cual no fue así. Más aún, su inclusión perjudica el resultado del ACP dado que la inercia explicada por el primer plano principal resultante de explicar a esta variable es pobre, y no permite un segundo eje que explique en mayor medida otras variables como *noinfectionrate*, resultado al que llegamos en el cuerpo del texto. Por lo tanto, optamos por finalmente no incluir esta variable en la base.

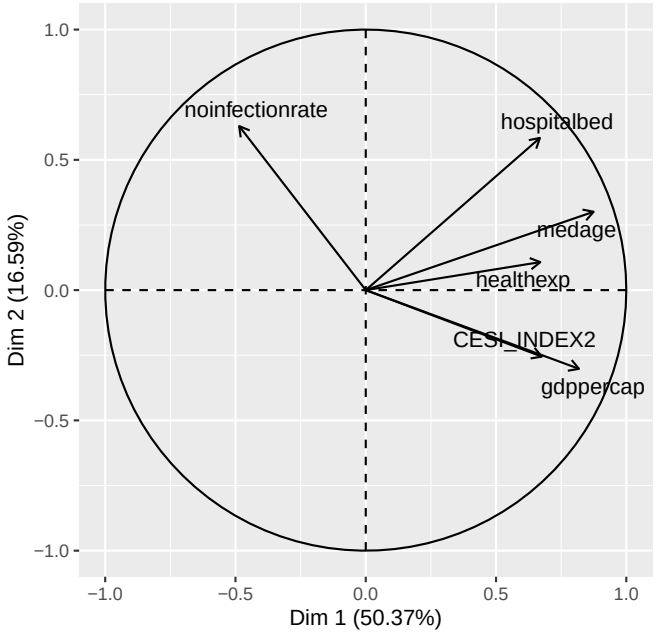

Figura 13: Proyecciones de las variables en el primer plano principal utilizando Índice de Estímulo Económico como variable activa.

Como se mencionó en el cuerpo del texto, en primera instancia se incluyó a *CESI_INDEX* como variable

activa para construir los ejes, opción que fue descartada debido a su pobre representación en los mismo y a su comportamiento similar al resto de las variables representadas principalmente por la primera componente (Figura 13). En lugar de esto se la incluyó como variable suplementaria.

Anexo 3: Gráfico de las proyecciones de los individuos en el plano, con San Marino como observación complementaria

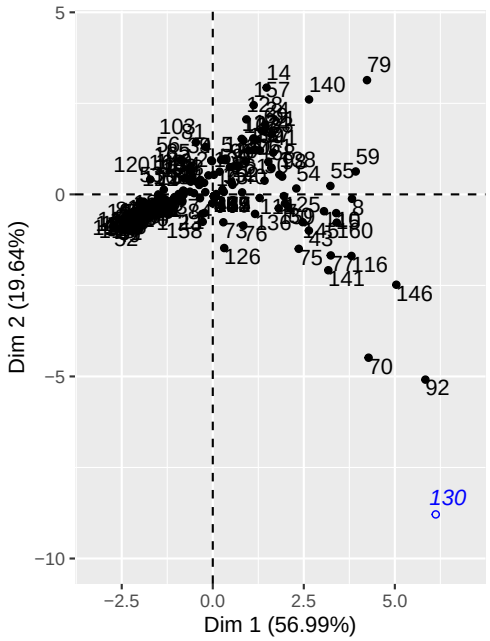

Figura 14: Proyecciones de los individuos en el primer plano principal, con la observación San Marino como suplementaria.