MAESTRÍA EN ECONOMETRÍA

ANÁLISIS ESTADÍSTICO MULTIVARIADO

Profesora:

- Clemente, Alejandra.

Ayudante:

- Mendoza Greco, Maximiliano.

Alumnos:

- Dignani, Franco;
- Guzzi, David.

Ciclo lectivo: Segundo Trimestre 2024.

UNIVERSIDAD TORCUATO DI TELLA

CONSIDEREACIONES INICIALES

El presente estudio ha utilizado, como **herramienta de cálculo**, el **lenguaje de programación Python**. En este sentido, todas las operaciones y resultados que se expongan fueron llevadas a cabo mediante la utilización de librerías de código abierto del citado lenguaje. Cabe aclarar que, para garantizar la confiabilidad de los resultados, **todas las operaciones han sido replicadas ya sea en lenguaje R o programa Stata y la simetría de resultados fue comprobada**.

No obstante, **ciertos algoritmos no han sido hallados en librerías de código abierto por lo que se ha requerido de una programación propia**, cuyos resultados lograron ser idénticos a los alcanzados por R o Stata. **Ejemplos** de ello aparecen, especialmente, en:

- **Análisis de Componentes Principales:** para la elaboración de gráficos biplot con distintos valores posibles del parámetro c;

- **Análisis de Factores:**

  - para la elaboración de una función para el **Test Multivariado de Mardia** (en librerías de código abierto como SciPy, Scikit-learn, Pingouin, no presentan test de normalidad multivariado, o cuando lo presentan no es el de Mardia);

  - para la programación de una **clase de varias funciones para la realización de Análisis de Factores de acuerdo con el Método de Factores Principales** cuyos resultados repliquen los alcanzados por Stata (para el desarrollo de este fue necesaria la lectura de la documentación de Stata en conjunto con los libros de referencia allí ofrecidos, como Rencher y Christensen, 2012 capítulo 13. Librerías de código abierto como FactorAnalyzer, Statsmodel, Scikit-learn, no brindan la posibilidad de aplicar análisis de factores mediante el método de factores principales –sí, por ejemplo, mediante método de máxima verosimilitud– y, cuando dicen hacerlo, en verdad lo que aplican es el Método de Componentes Princiaples, donde la matriz estimada Psi (Ψ) no es considerada en la estimación de la matriz de varianzas y covarianzas o matriz de correlaciones).

Por último, como complemente del presente estudio, **se brinda**:

- **Archivo de código .ipynb** (Jupyter Notebook) con el respaldo de las operaciones llevadas a cabo;

- **Igual archivo de código .ipynb pero en Google Colab (link)**, para ejecutar en línea el anterior script en caso de ser necesario (para mayor comodidad y evitar posibles problemas de versionado de librerías o dependencias);

- **Repositorio en línea sincronizado en GitHub (link)**, donde podrán encontrarse todos los documentos intervinientes en el presente estudio (presentación, base de datos, archivo .ipynb).

1. ANÁLISIS EXPLORATORIO

INTRODUCCIÓN

En este estudio se han seleccionado los países miembros de la **Organización para la Cooperación y el Desarrollo Económicos (OCDE)** junto con **Argentina** para analizar diversos indicadores económicos y sociales relacionados, de forma general, con la **Agricultura**. La elección de los países de la OCDE se basa en su diversidad económica y geográfica, así como en su relevancia en políticas internacionales y estándares de desarrollo. Por su parte, Argentina fue incluida como un caso de estudio adicional, para indagar posibles contrastes entre nuestra Nación y los países desarrollados. Se considera que este enfoque permite **comparar y contrastar una amplia gama de datos y naciones**, al tiempo de identificar patrones y tendencias clave.

Para el presente análisis, se ha seleccionado el **año 2019** como punto de referencia (dado que se considera representa el último año con condiciones económicas y sociales estables antes del impacto global de la pandemia de COVID-19) en conjunto con las **siguientes variables** que se resumen y definen del siguiente modo:

Variables	Definición
Agricultural land (% of land area)	Mide la proporción de tierras agrícolas, que incluye tierras arables, cultivos permanentes y pastos permanentes.
Arable land (% of land area)	Mide la proporción de tierra que se utiliza para cultivos temporales, pastos temporales y jardines, excluyendo tierras abandonadas por cultivo itinerante.
Agricultural raw materials exports (% of merchandise exports)	Mide proporción de exportaciones de materias primas agrícolas dentro del total de exportaciones de mercancías.
Agricultural raw materials imports (% of merchandise imports)	Mide la proporción de las importaciones de materias primas agrícolas respecto al total de importaciones de mercancías.
Agriculture, forestry, and fishing, value added (% of GDP)	Representa el valor agregado neto de los sectores de agricultura, silvicultura y pesca como porcentaje del PIB. Incluye la producción agrícola, ganadera, caza, pesca y silvicultura, y excluye la depreciación de activos y la degradación de recursos naturales.
Employment in agriculture (% of total employment) (modeled ILO estimate)	Mide la proporción de personas empleadas en el sector agrícola, que incluye actividades de agricultura, caza, silvicultura y pesca, en comparación con el total de empleo.
Rural population (% of total population)	Mide la proporción de personas que viven en áreas rurales como porcentaje de la población.

Nota: Más allá de contar con variables relacionadas de forma general con la Agricultura, la elección de variables responde a la identificación de aquellas variables dentro de la base de datos suministrada (World Bank Group: World Development Indicators) que no presenten datos nulos al tiempo de que las mismas estén expresadas en una misma unidad de medida (variables relativas a una magnitud, como el PBI) para evitar posibles futuros problemas a lo largo del estudio.

ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

Medidas descriptivas:

Indicador / Variable	Agricultural land	Arable land	Agricultural exports	Agricultural imports	Agriculture, forestry, and fishing, value added	Employment in agriculture	Rural population
Cantidad obs.	38	38	38	38	38	38	38
Promedio	0,3823	0,1836	0,0245	0,0138	0,0233	0,0501	0,2160
Desvío estándar	0,1752	0,1381	0,0287	0,0064	0,0157	0,0444	0,1119
Mínimo	0,0270	0,0120	0,0014	0,0055	0,0022	0,0068	0,0196
Percentil 25%	0,2973	0,0667	0,0069	0,0089	0,0128	0,0211	0,1351
Mediana	0,4374	0,1634	0,0134	0,0115	0,0188	0,0372	0,1936
Percentil 75%	0,4868	0,2740	0,0274	0,0178	0,0326	0,0574	0,2904
Máximo	0,7242	0,5990	0,1168	0,0291	0,0641	0,1811	0,4627
Coef. de Variación	0,4582	0,7521	1,1742	0,4652	0,6723	0,8860	0,5179
Kurtosis	-0,4182	0,7749	4,1573	0,0087	0,9200	1,9896	-0,3474
Asimetría	-0,4175	0,8868	2,1202	0,9792	1,1671	1,6272	0,5370
ARG	0,4259	0,1491	0,0088	0,0105	0,0532	0,0733	0,0801

La **matriz de datos** presenta **38 observaciones** pertenecientes a países de la **OCDE** y, sumando a **Argentina**, da **39 en total**. Si bien el **análisis considerará el conjunto de 39 observaciones**, aquí se dividen ambos grupos para **enriquecer la descripción** de cada una de las variables, que se pasa a comentar de forma general brevemente.

- En la gran mayoría de las variables se observa una **disparidad** importarte entre los **valores mínimos y máximos**. Esto nos da cuenta de la **heterogeneidad de economías** presentes en matriz de datos;

- A pesar de ello, considerando la matriz de **forma agrupada** puede observarse que esta heterogeneidad parecería desaparecer dado que los guarismos para los **promedios** y las **medianas** no son muy disímiles, lo que indicaría que los **casos extremos** dentro de la matriz no serían significativos. Por otro lado, no obstante, algunas variables presentan altos **desvíos estándar**, como el caso de *Agricultural land, Arable land, y Rural population*.

- Con relación a la **distribución de las variables**, la gran mayoría presentan distribuciones positivas sesgadas hacia la derecha al tiempo de evidenciar colas de distribución alargadas (dados los valores extremos presentes).

- Comparativamente, **Argentina y OCDE** (considerando su mediana) presentan **similitudes** en variables como *Agricultural land, Arable land, y Agricultural imports*. Sin embargo, *Agricultural exports y Rural population* se encuentran por encima en naciones OCDE, y variables como *Agricultural value added y Agricultural employmernt* son superiores relativamente en Argentina.

MATRICES: COVARIANZAS Y CORRELACIONES

Matriz de Varianzas y Covarianzas:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	0,0299						
2	0,0145	0,0186					
3	-0,0015	-0,0009	0,0008				
4	0,0000	0,0003	0,0001	0,0000			
5	-0,0002	-0,0006	0,0002	0,0000	0,0003		
6	0,0005	-0,0012	0,0001	0,0000	0,0006	0,0019	
7	0,0029	0,0022	-0,0001	0,0002	-0,0001	0,0007	0,0127

Medidas de variabilidad global:

Medida	Valor
Varianza total	0,0642
Varianza media	0,0092
Varianza generalizada	0,0000
Varianza efectiva	0,0013

Comentarios:

- Se observa **variancia máxima** en *Agricultural land* y **varianza mínima** en *Agricultural imports*.
- Predominan **relaciones positivas** entre las variables, siendo la más significativa entre *Agricultural land* y *Arable land*. Por su parte, la **relación negativa** más significativa se produce entre *Agricultural land* y *Agricultural exports* (contraintuitivo, en principio).
- Dado el rango que pueden tomar los valores (0 a 1, dado que son porcentajes), todas las medidas de variabilidad global indicarían un **bajo nivel de varianza global,** incluso en aquellas medidas que sí consideran las relaciones o covarianzas entre las distintas variables (como la varianza generalizada y la varianza efectiva).
- El análisis de la **matriz de correlación**, que presenta guarismos no influidos por las unidades de medida de las distintas variables, permitirá confirmar los anteriores comentarios.

Matriz de Correlaciones:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	1						
2	0,6069	1					
3	-0,3079	-0,2349	1				
4	-0,0218	0,3737	0,3101	1			
5	-0,0596	-0,2523	0,3555	0,0419	1		
6	0,0674	-0,1942	0,0926	0,0634	0,8028	1	
7	0,1476	0,1410	-0,0322	0,2899	-0,0485	0,1450	1

Medidas de dependencia global:

Medida	Valor
Dependencia conjunta	0,0642
Dependencia efectiva	0,0092

Comentarios:

- Las variables, al ser estandarizadas, confirmarían los **grados y direcciones de relación** comentados anteriormente. Incluso, los guarismos de relación serían más altos en términos absolutos.
- Las medidas globales indicarían un **bajo nivel de dependencia conjunta**.

2. ANÁLISIS DE COMPONENTES PRINCIPALES

ANÁLISIS DE COMPONENTES PRINCIPALES

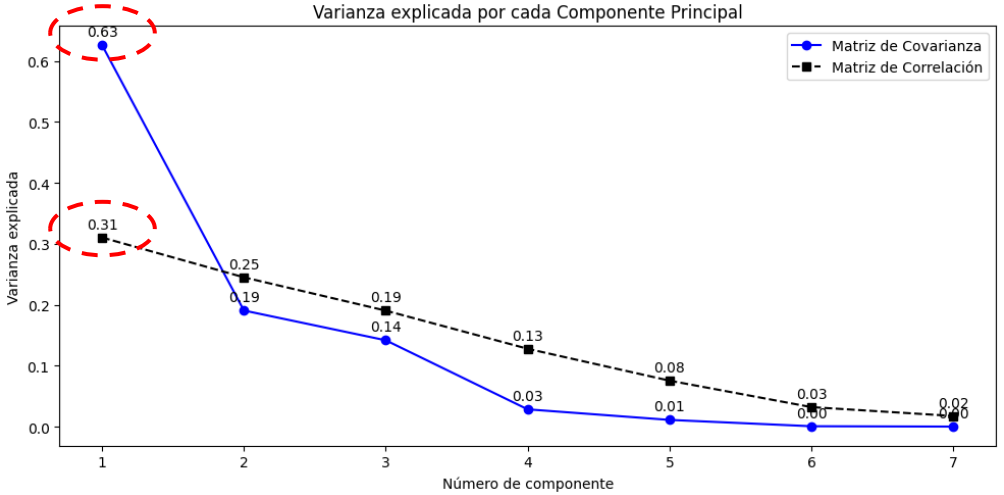
Se presentan a continuación los resultados del **Análisis de Componentes Principales:**

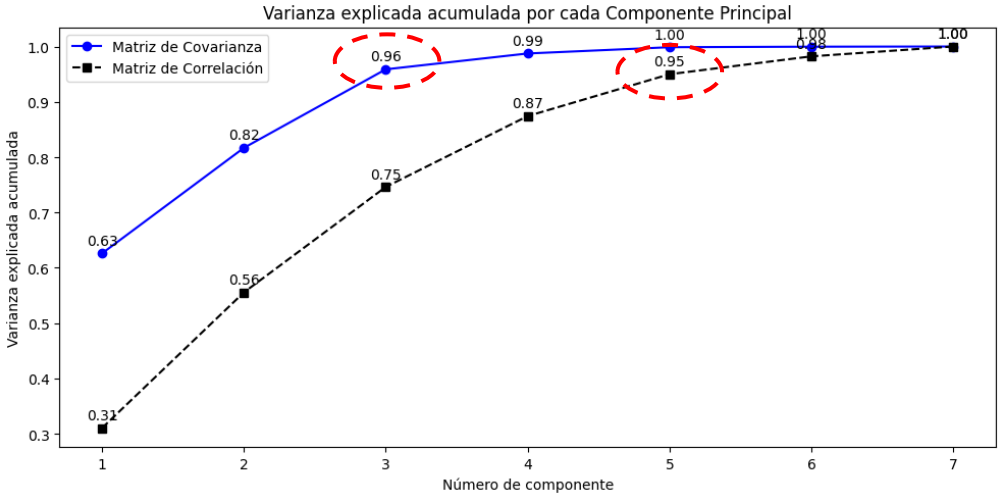
Para una **matriz de datos no estandarizados** (**Matriz de Covarianzas**):

N° de Componente	Varianza	Varianza Explicada	Varianza Exp. Acum.
Componente 1	0,04	0,63	0,63
Componente 2	0,01	0,19	0,82
Componente 3	0,01	0,14	0,96
Componente 4	0,00	0,03	0,99
Componente 5	0,00	0,01	1,00
Componente 6	0,00	0,00	1,00
Componente 7	0,00	0,00	1

Para una **matriz de datos estandarizados** (**Matriz de Correlaciones**):

N° de Componente	Varianza	Varianza Explicada	Varianza Exp. Acum.
Componente 1	2,23	0,31	0,31
Componente 2	1,76	0,25	0,56
Componente 3	1,37	0,19	0,75
Componente 4	0,92	0,13	0,87
Componente 5	0,54	0,08	0,95
Componente 6	0,23	0,03	0,98
Componente 7	0,13	0,02	1





Comentarios:

- Dado el objetivo de **reducción de cantidad de variables con la menor pérdida de varianza posible** para la matriz de datos analizada, se ha realizado un Análisis de Componentes Principales (PCA) de **dos formas distintas**, cuyos **resultados difieren sensiblemente**.
- Dada la matriz de datos no estandarizada, PCA nos indica que el **primer componente** explica en un 63% la **varianza total** de los datos. Sin embargo, dada la matriz de datos estandarizada, PCA nos indica que este mismo guarismo es del 31%.
- Al presentar la información de varianzas explicadas de **forma gráfica y conjunta** para ambos análisis de PCA, puede concluirse que las **diferencias entre ambos resultados se mantienen**: mientras que el primer análisis indicaría que para contar con un nivel de varianza explicada del 95% se requerirían 3 componentes, el segundo análisis concluiría que mismo guarismo podría conseguirse con tan solo 5 componentes.
- A pesar de que PCA con datos no estandarizados "conservaría" mayor varianza en las primeras componentes, **se preferirá trabajar con los resultados de PCA para datos estandarizados** (si bien sabemos que las variables están en una misma unidad de medida, preferimos trabajar con datos estandarizados por si pequeñas relaciones de variabilidad entre variables no han sido visualmente percibidas por nosotros, los analistas).

ANÁLISIS DE COMPONENTES PRINCIPALES

Se presenta a continuación la **interpretación de las dos primeras Componentes Principales**, dado PCA con datos estandarizados (Matriz de Correlaciones):

Coeficiente 1 y Coeficiente 2 para cada **variable**:

Variable / Coeficiente	Coeficiente 1	Coeficiente 2
Agricultural land	-0,3549	0,4082
Arable land	-0,4678	0,4219
Agricultural exports	0,3852	0,0762
Agricultural imports	-0,0173	0,4622
Agriculture, forestry, and fishing, value added	0,5456	0,3133
Employment in agriculture	0,4495	0,4094
Rural population	-0,0827	0,4125

Correlación entre **Componente 1 y Componente 2** con **variables**:

Variable / Componente	Componente 1	Componente 2
Agricultural land	-0,5233	0,5347
Arable land	-0,6897	0,5526
Agricultural raw materials exports	0,5678	0,0998
Agricultural raw materials imports	-0,0255	0,6056
Agriculture, forestry, and fishing, value added	0,8045	0,4104
Employment in agriculture	0,6627	0,5364
Rural population	-0,1219	0,5404

Matriz de datos estandariza (solo 20 observaciones) junto a **primeros dos componentes**:

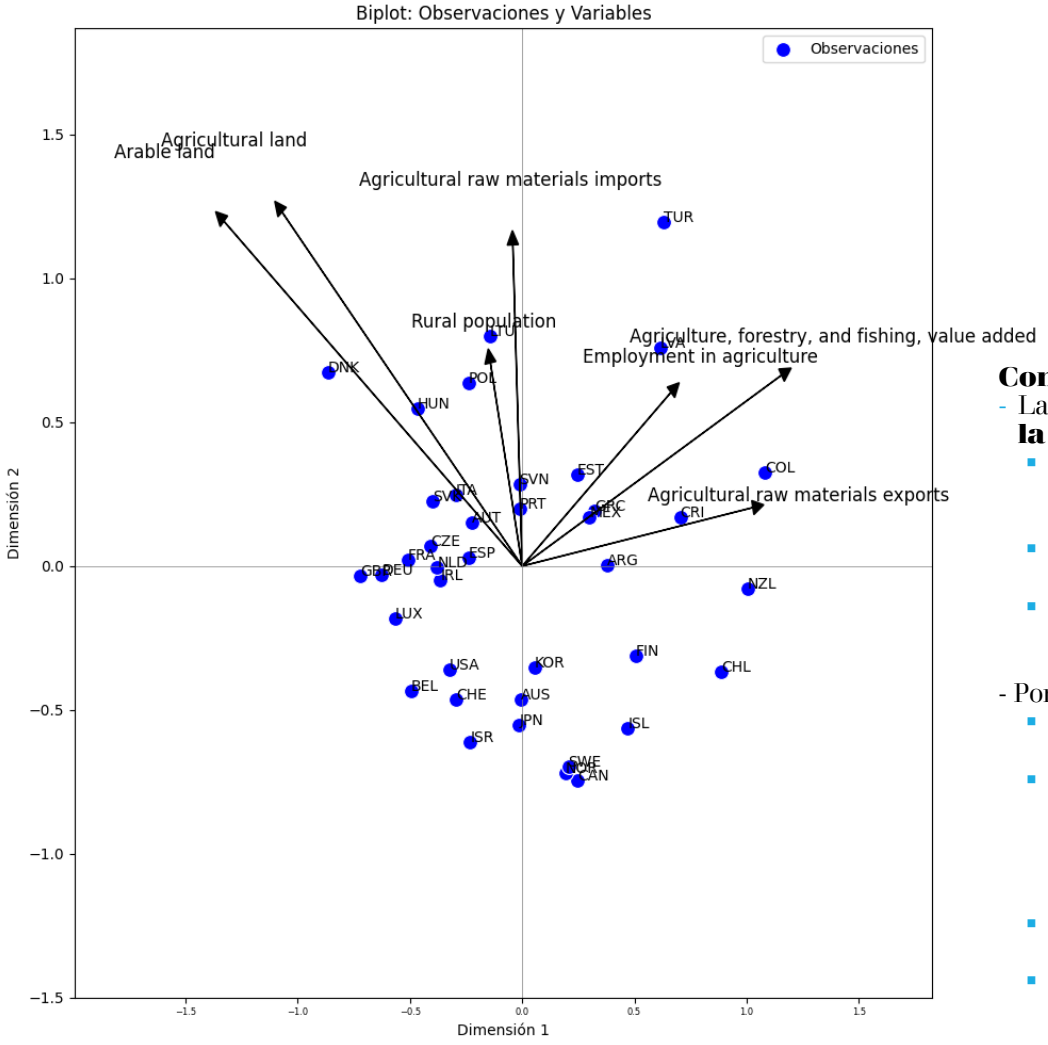
País	Agri. land	Arable land	Agri. exports	Agri. imports	Value added	Emplo yment	Rural population	Comp. 1	Comp. 2
Argentina	0,25	-0,25	-0,54	-0,51	1,80	0,51	-1,18	1,14	0,01
Australia	0,51	-1,05	-0,11	-1,11	-0,18	-0,57	-0,66	-0,01	-1,32
Austria	-0,36	-0,16	-0,31	0,70	-0,84	-0,32	1,80	-0,68	0,43
Belgium	0,37	0,74	-0,41	-0,51	-1,07	-0,94	-1,71	-1,50	-1,23
Canada	-1,84	-1,02	0,45	-0,87	-0,44	-0,81	-0,24	0,74	-2,12
Switzerland	-0,01	-0,60	-0,80	-1,29	-1,09	-0,58	0,44	-0,89	-1,31
Chile	-1,34	-1,21	1,59	-1,01	0,97	0,91	-0,79	2,68	-1,05
Colombia	0,09	-1,20	0,58	-0,79	2,47	2,51	-0,21	3,26	0,92
Costa Rica	-0,21	-0,99	-0,21	-0,54	1,11	2,33	-0,12	2,12	0,48
Czechia	0,42	1,02	-0,39	-0,48	-0,34	-0,55	0,43	-1,23	0,20
Germany	0,54	1,12	-0,59	-0,25	-1,00	-0,88	0,12	-1,89	-0,09
Denmark	1,58	3,05	0,05	1,75	-0,71	-0,65	-0,82	-2,61	1,92
Spain	0,82	0,39	-0,48	-0,49	0,04	-0,24	-0,16	-0,72	0,08
Estonia	-0,88	-0,16	1,81	2,23	-0,06	-0,43	0,86	0,74	0,90
Finland	-1,78	-0,80	1,67	0,93	-0,04	-0,29	-0,60	1,53	-0,88
France	0,80	1,08	-0,53	-0,45	-0,55	-0,58	-0,17	-1,53	0,06
UK	1,97	0,51	-0,69	-0,53	-1,11	-0,92	-0,44	-2,18	-0,09
Greece	0,41	-0,13	-0,13	-0,81	0,86	1,38	-0,06	0,98	0,55
Hungary	1,13	2,13	-0,62	-0,35	0,57	-0,08	0,63	-1,40	1,56
Ireland	1,58	-0,87	-0,74	-1,25	-0,93	-0,15	1,36	-1,10	-0,14

Comentarios:

- La **interpretación de los Componente Principales** puede realizarse o bien analizando la **matriz de coeficientes** (autovectores) provenientes de PCA, o bien al observar la **correlación** existente entre los componentes y las variables originales, dado que los resultados alcanzados por ambos ofrecen iguales conclusiones.
- Dado que los componentes principales surgen (bajo este modelo PCA) del producto matricial entre la matriz de datos estandarizados y los coeficientes, puede interpretarse lo siguiente:
  - **Componente Principal 1:** Representaría, en forma conjunta y de forma positiva, un producto ponderado del valor agregado, el empleo y las exportaciones en la agricultura. A su vez, implicaría, de forma negativa, un producto ponderado de la tierra arable y agrícola. En este sentido, es de esperar que países con altos guarismos en el primer grupo y bajos guarismos en el segundo grupo, presenten un alto Componente Principal 1. Ejemplo: Colombia.
  - **Componente Principal 2:** Representaría, en forma conjunta y de forma positiva, un producto ponderado de todas las variables de la matriz de datos, siendo las importaciones agrícolas la variable que presenta mayor peso en la ponderación, seguidas de tierras arables y agrícolas. Ejemplo: Dinamarca.

ANÁLISIS DE COMPONENTES PRINCIPALES

Se presenta un **análisis de estadística descriptiva en conjunto con una representación *biplot* de las dos primeras Componentes Principales**:



Indicador / Variable	Comp. 1	Comp. 2
Cantidad obs.	39	39
Promedio	0,0000	0,0000
Desvío estándar	1,4743	1,3100
Mínimo	-2,6090	-2,1212
Percentil 25%	-1,1256	-1,0336
Mediana	-0,0392	0,0073
Percentil 75%	0,9386	0,6697
Máximo	3,2579	3,4038
Kurtosis	-0,4980	-0,0901
Asimetría	0,4785	0,3901

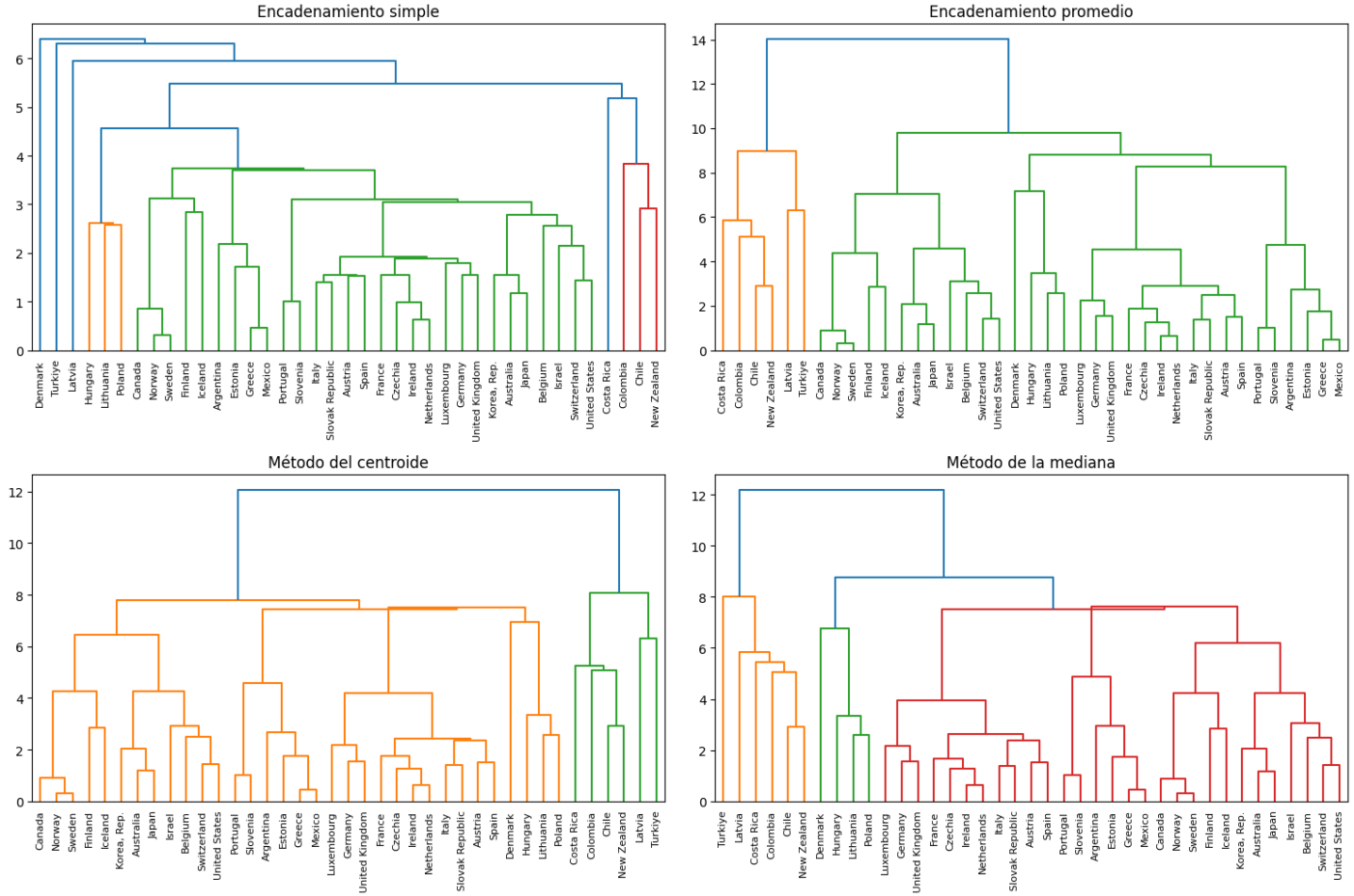
Comentarios:

- Las **estadísticas descriptivas** de las dos primeras Componentes Principales reflejan que **las características de la matriz de datos original han sido preservadas,** dado que:
  - Se mantiene la **disparidad entre valores máximos y mínimos**, dando cuenta de la heterogeneidad entre países. Sin embargo, y dado que la **distancia entre la mediana y el promedio es no significativa**, las diferencias anteriores (máx. vs. mín.) vendrían dados por unos pocos casos extremos;
  - Ante la **estandarización de datos**, los componentes poseen ahora **promedio** prácticamente cero y **desvío estándar** cercano a uno;
  - Por último, dado lo anterior y en base a coeficientes de asimetría y kurtosis, podría afirmarse que las variables poseen una **distribución** ligeramente simétrica hacia la derecha con colas menos alargadas en comparación con las variables originales (es decir, se aproximan a una distribución normal).
- Por otro lado, de la **visualización del gráfico biplot**, puede destacarse:
  - Nota: para una **representación conjunta de observaciones y variables**, se optó por un valor de **c = 0,5** (consiguiendo un **biplot simétricamente escalado**);
  - En relación con las **variables** (autovectores), se **confirma visualmente la interpretación de los Componentes Principales,** siendo relevantes las variables valor agregado, el empleo, y exportaciones en la agricultura (de forma positiva), y tierra arable y agrícola (de forma negativa) para el **Componente 1**, y siendo relevantes todas las variables de forma positiva para el **Componente 2**, con una mayor participación en variables como tierra arable y agrícolas, e importaciones;
  - En relación con las **observaciones**, puede interpretarse de forma preliminar la distribución de cada país en términos de las dos primeras Componentes Principales. Se reservan comentarios para la sección de Cluster Jerárquico;
  - Por último, y si bien para la representación conjunta (c=0,5) la magnitud de las variables (autovectores) fue alterada, el **gráfico confirma de igual manera las relaciones de dependencia entre variables previamente examinadas en la matriz de correlación** (representación de orden dos). Por ejemplo, elevada correlación entre *Agricultural land y Arable land*, baja correlación entre estas dos y *Agricultura exports*, etc.

3. ANÁLISIS DE CLUSTERS JERÁRQUICO

ANÁLISIS DE CLUSTERS JERÁRQUICO

Antes del Análisis de Clusters, se comentan **breves razones por las cuales no se han considerado ciertos métodos de encadenamiento:**



Método de encadenamiento simple

- Formación de clústeres alargados y dispersos (fenómeno de "chaining“) dado que este método tiende a ser *space-contracting*;
- Formación de clústeres poco equilibrado dado que en este caso se conforman tres grupos de tres países y uno de treinta;
- Dificultad de interpretación debido a grupos dispersos.

Método de encadenamiento promedio

- No forma clústeres tan compactos ni diferenciados dado que al utilizar el promedio los límites entre los grupos pueden volverse menos claros, lo que dificulta la interpretación visual del dendrograma;
- Dado que buscamos identificar grupos bien diferenciados este método dificulta la identificación de patrones significativos en el análisis.

Método del centroide

- Al igual que el método de la mediana, por la falta de monotonicidad la representación de los grupos se vuelve poco clara y difícil de interpretar;
- Al presentar “Reversals” (representados en el dendrograma con los crossovers) esto hace que las observaciones tienden a fusionarse rápidamente, lo que puede resultar en clústeres menos diferenciados y más densamente compactados, afectando la claridad del análisis.

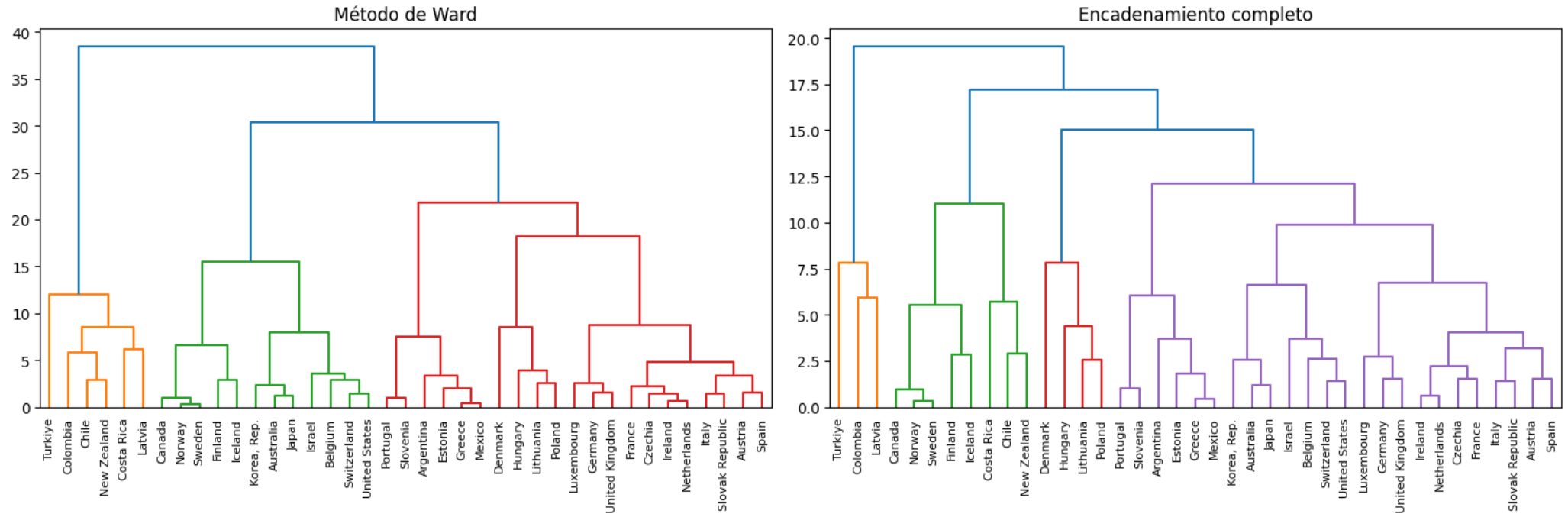
Método de la mediana

- Método de la mediana carece de monotonicidad generando "crossovers" en los dendrogramas;
- Esta falta de coherencia en las fusiones de los clústeres dificulta significativamente la interpretación visual del dendrograma, volviéndolo una opción menos adecuada para este análisis.

Nota: para todos los dendogramas se utilizó la distancia euclidiana como métrica de proximidad entre los países.

ANÁLISIS DE CLUSTERS JERÁRQUICO

Se presenta a continuación una breve **descripción** de los **métodos de encadenamientos seleccionados:**



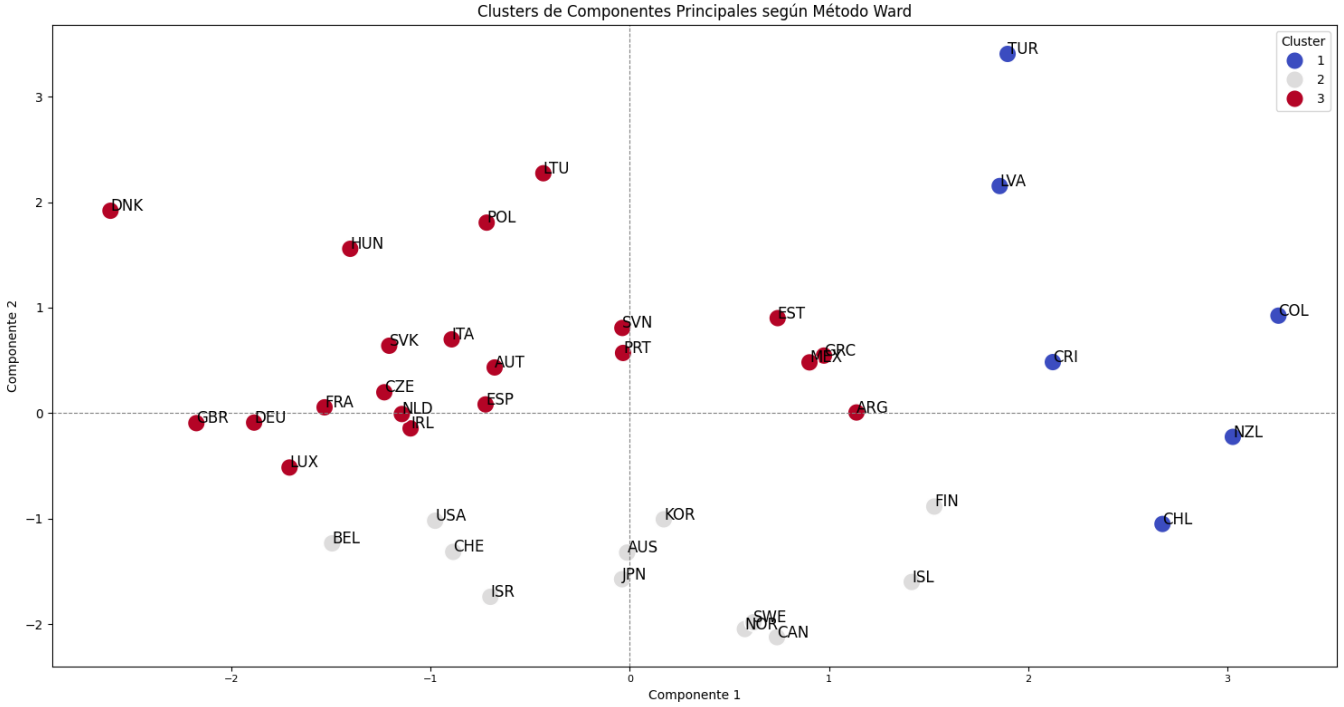
Comentarios:

- De forma general, la **selección del Método de Ward** responde a:
  - **Minimiza la varianza interna dentro de los clústeres**, lo que garantiza que los países dentro de un mismo grupo sean lo más homogéneos posible;
  - Dado lo anterior, tiene una **sensibilidad moderada a la presencia de *outliers***. Esto permite que los grupos homogéneos se formen de manera coherente, incluso cuando hay observaciones muy disímiles. Por ello, **Turquía y Letonia** se podrían comportar como *outliers* respecto al resto. Estos países se agrupan más tarde en el proceso, lo que sugiere que sus características agrícolas y rurales son significativamente diferentes a las de los otros países.

- Por su parte, la selección del **encadenamiento completo** se debe a:
  - Generación de **clústeres separados entre sí,** simplificando la identificación de países y manteniendo distancias significativas con otros clústeres (a diferencia de Ward que busca minimizar la varianza interna, lo permite ver dos perspectivas diferentes en la agrupación de los países);
  - **Mayor sensibilidad ante la presencia de *outliers*** que el método de Ward. Por ejemplo, países como Colombia, Turquía y Dinamarca, que tienen comportamientos atípicos, son agrupados en etapas más tardías del proceso, lo que resalta aún más sus diferencias en comparación con los demás clústeres.

ANÁLISIS DE CLUSTERS JERÁRQUICO

Se presenta a continuación una **interpretación de los Clusters dado el Método de Ward en el espacio de las Componentes Principales:**



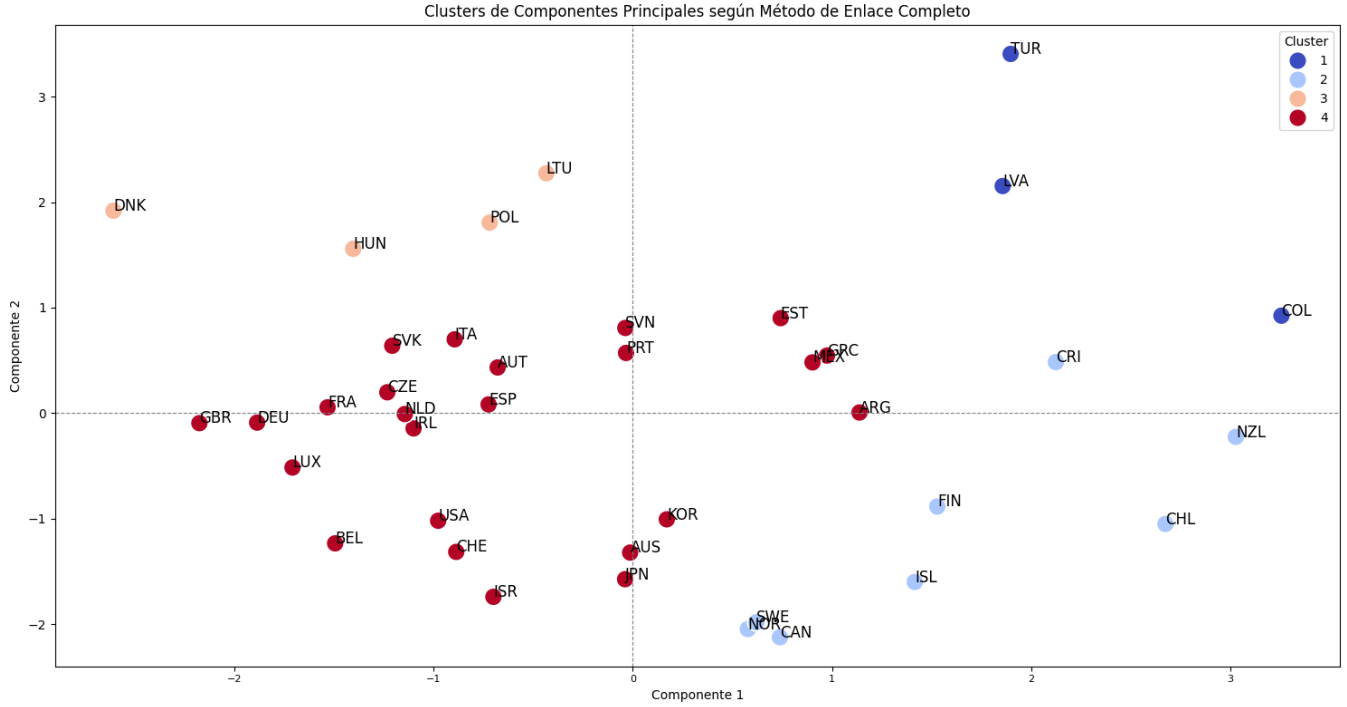
Comentarios:

- Recordar que, las componentes principales poseen la siguiente interpretación:
  - **Componente Principal 1:** Representaría, en forma conjunta y de forma positiva, un producto ponderado del valor agregado, el empleo y las exportaciones en la agricultura. A su vez, implicaría, de forma negativa, un producto ponderado de la tierra arable y agrícola. En este sentido, es de esperar que países con altos guarismos en el primer grupo y bajos guarismos en el segundo grupo, presenten un alto Componente Principal 1.
  - **Componente Principal 2:** Representaría, en forma conjunta y de forma positiva, un producto ponderado de todas las variables de la matriz de datos, siendo las importaciones agrícolas la variable que presenta mayor peso en la ponderación, seguidas de tierras arables y agrícolas.

- Dado esta conceptualización, y en función al reconocimiento de **tres clústers** según Método de Ward, los mismos **pueden ser definidos de la siguiente forma**:
  - **Clúster 1:** Lo que los caracteriza como un grupo homogéneo y diferente del resto es que poseen **elevados guarismos en componente 1:** naciones con alto valor agregado, empleo y exportaciones agrícolas, al tiempo de poseen bajo territorio para estos fines. Se aprecian países como Turquía y Letonia con alto componente 2, a su vez;
  - **Clúster 2:** De forma similar, logran la característica de clúster por medio de presentar **guarismos negativos de componente 2**: naciones con bajos guarismos en importaciones agrícolas y territorio para dichos fines. A su vez (generalmente), el grupo se divide entre aquellos que presentan guarismos negativos y positivos de componente 1, siendo la diferencia en si presentan o no territorio para dichos fines, respectivamente.()
  - **Clúster 3:** Presentan, mayoritariamente, **guarismos positivos en componente 2** (en general, producto ponderado positivo de todas las variables).

ANÁLISIS DE CLUSTERS JERÁRQUICO

Se continúa con la **interpretación de los Clusters dado el encadenamiento completo en el espacio de las Componentes Principales:**



Comentarios:

- En función al reconocimiento de **cuatro clústers** según el encadenamiento completo, los mismos **pueden ser definidos de la siguiente forma**:

  - **Clúster 1:** Representa un **subgrupo del grupo 1 del método de Ward**, por lo que su caracterización se mantiene. Se infiere que el **elevado componente 2** de Turquía y Letonia, con al alto **componente 1 de Colombia** (el mayor entre las observaciones) pesaron para esta división.
  - **Clúster 2:** Representa la unión parcial entre los grupos 1 y 2 (solo aquellas observaciones positivas en componente 1) del anterior método. Se caracterizan por poseer **positivos guarismos en componente 1 y negativos en componente 2:** no poseen, en general, tierra agrícola;
  - **Clúster 3: Partición del grupo 3** del método de Ward. Se excluye a Dinamarca, Hungría, Polonia y Lituania del anterior grupo 3 **mejorando la heterogeneidad entre los grupos**, ya que estos poseen guarismos elevados en componente 2.
  - **Clúster 4: Se incorporan aquellos países pertenecientes al grupo 2 del análisis anterior que presentaban guarismos negativos de componente 2**. A grandes rasgos, dada la actual configuración, el grupo puede interpretarse como un **promedio ponderado de ambos componentes**, dada su concentración en torno al origen de coordenadas.

- Dado el **manejo de posibles *outliers*** (como Turquía y Letonia, por un lado, y de Dinamarca, Hungría, Polonia y Lituania, por otro) en conjunto a la **apertura del grupo 2** según método de Ward (aquí grupos 3 y 4) en naciones con o sin (en general) territorio para fines agrícolas, se considera que **el método de encadenamiento completo ofrece mejores resultados** en términos de lograr grupos homogéneos dentro de sí y heterogéneos entre sí.

4. ANÁLISIS DE FACTORES

ANÁLISIS DE FACTORES

Se presenta a continuación un **Análisis de Factores considerando un modelo de un único factor:**

Factor	Varianza (autovalor)	Prop. explicada	Prop. exp. acum.
Factor 1	1,8282	0,4982	0,4982
Factor 2	1,2632	0,3442	0,8424
Factor 3	0,8155	0,2222	1,0646
Factor 4	0,2328	0,0634	1,1281
Factor 5	-0,0323	-0,0088	1,1193
Factor 6	-0,1703	-0,0464	1,0729
Factor 7	-0,2674	-0,0729	1

Variable	Coef. 1	Uniqueness	Communality
Agricultural land	-0,3620	0,8689	0,1311
Arable land	-0,5573	0,6894	0,3106
Agricultural exports	-0,4277	0,8170	0,1830
Agricultural imports	-0,0015	1,0000	0,0000
Agriculture value added	0,8288	0,3132	0,6868
Employment in agriculture	0,7171	0,4858	0,5142
Rural population	-0,0502	0,9975	0,0025

Comentarios:

- Nota: **Antes de efectuar el Análisis de Factores**, fue realizado un **Test de Hipótesis de Normalidad Multivariada de Mardia**. De acuerdo con los resultados de este (que pueden ser encontrados en archivo .ipynb), si bien se rechazaría la hipótesis nula según Simetría de Mardia (p-value igual a cero), dicha conclusión no se mantendría según Kurtosis de Mardia. Sin embargo, y a pesar de no existir una definición exhaustiva por parte de la hipótesis de no normalidad de datos multivariados, **se efectuará análisis según Método de Factores Principales**.

- Del **Análisis de Factores**, de acuerdo con el método de factores principales, **para el primer factor,** puede destacarse:
  - El mismo **conservaría** aproximadamente el **50% de la varianza total** de la matriz de datos;
  - Analizando el **primer vector de coeficientes** (*loadings*), puede resaltarse **guarismos positivos para valor agregado y empleo en Agricultura**, siendo el resto guarismo negativos;
  - En relación con la **bondad de ajuste del modelo** de análisis de factores (cuán bien reconstruyen los factores a las variables), puede argumentarse que el mismo sería **deficiente** dados los **altos niveles de variabilidad específica o *uniqueness*** (no asociados a los factores) presentes en la mayoría de las variables.

ANÁLISIS DE FACTORES

Se presenta a continuación un **Análisis de Factores considerando un modelo de un único factor:**

Matriz de Correlación estimada:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	1						
2	0,2017	1					
3	-0,1548	-0,2384	1				
4	0,0006	0,0009	-0,0007	1			
5	-0,3	-0,4618	0,3545	-0,0013	1		
6	-0,2596	-0,3996	0,3067	-0,0011	0,5943	1	
7	0,0182	0,028	-0,0215	0,0001	-0,0416	-0,036	1

Matriz de diferencias entre Matriz correlación estimada vs. Matriz correlación real:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	1						
2	-0,4052	1					
3	0,1530	-0,0035	1				
4	0,0223	-0,3728	-0,3107	1			
5	-0,2405	-0,2095	-0,0010	-0,0431	1		
6	-0,3270	-0,2054	0,2141	-0,0645	-0,2085	1	
7	-0,1294	-0,1130	0,0107	-0,2898	0,0068	-0,1810	1

Comentarios:

- Nota: Para el presenta análisis, dado que el Método de Factores Principales trabaja por defecto con datos estandarizados, no se considerará la Matriz de Varianzas y Covarianzas, por lo que **se efectúa la estimación de la Matriz de Correlaciones**.
- **Otra manera de analizar la bondad de ajuste** del modelo de análisis factorial es examinar cuan bien replica las varianzas y covarianzas (correlaciones, por ser datos estandarizados) de la matriz de datos original. De este modo, se presenta la estimación de la matriz, por un lado, y la matriz de diferencias presentes en la estimación, por otro.
- De acuerdo con la segunda matriz, se observa diferencias significativas (por exceso o defecto) en correlación entre *Agricultural land y Arable land*, *Arable land y Agricultural imports*, *Agricultural land y Employment*, etc. Esto daría cuenta de un **deficiente ajuste del modelo**.

ANÁLISIS DE FACTORES

Se presenta a continuación un **Análisis de Factores considerando un modelo de dos factores:**

Factor	Varianza (autovalor)	Prop. explicada	Prop. exp. acum.
Factor 1	1,8282	0,4982	0,4982
Factor 2	1,2632	0,3442	0,8424
Factor 3	0,8155	0,2222	1,0646
Factor 4	0,2328	0,0634	1,1281
Factor 5	-0,0323	-0,0088	1,1193
Factor 6	-0,1703	-0,0464	1,0729
Factor 7	-0,2674	-0,0729	1

Variable	Coef. 1	Coef. 2	Uniqueness	Communality
Agricultural land	-0,3620	-0,5729	0,5407	0,4593
Arable land	-0,5573	-0,6024	0,3265	0,6735
Agricultural exports	-0,4277	0,0578	0,8137	0,1863
Agricultural imports	-0,0015	-0,4064	0,8349	0,1651
Agriculture value added	0,8288	-0,3121	0,2158	0,7842
Employment in agriculture	0,7171	-0,4543	0,2794	0,7206
Rural population	-0,0502	-0,3159	0,8977	0,1023

Comentarios:

- Nota: Por idénticas razones, **se efectúa análisis según Método de Factores Principales**.

- **De la incorporación de un segundo factor al Análisis de Factores,** puede destacarse:
  - Se produce un incremento de varianza explicada del 34%. Se conservaría, de este modo, aproximadamente el **84% de la varianza total** de la matriz de datos;
  - Analizando el **segundo vector de coeficientes** (*loadings*), pueden resaltarse **guarismos negativos** en prácticamente todas las variables. Debe notarse aquí que Stata arriba a iguales guarismos, pero de signo contrario. Esto puede explicarse (consideramos) por una diferencia en el método SVD (singular value descomposition) empleado. Para **garantizar resultados**, los interpretaremos con su **signo inverso** (es el caso de slide 19);
  - En relación con la **bondad de ajuste del modelo,** los niveles de **variabilidad específica o *uniqueness*** (no asociados a los factores) disminuyeron en relación con el modelo de un solo factor contemplado anteriormente, lo que daría cuenta de un incremento de performance del modelo.

ANÁLISIS DE FACTORES

Se presenta a continuación un **Análisis de Factores considerando un modelo de dos factores:**

Matriz de Correlación estimada:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	1						
2	0,5469	1					
3	-0,1880	-0,2732	1				
4	0,2334	0,2457	-0,0242	1			
5	-0,1212	-0,2738	0,3364	0,1255	1		
6	0,0007	-0,1259	0,2805	0,1835	0,7361	1	
7	0,1992	0,2183	0,0398	0,1285	0,0570	0,1075	1

Matriz de diferencias entre Matriz correlación estimada vs. Matriz correlación real:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1	1						
2	-0,0600	1					
3	0,1199	-0,0384	1				
4	0,2551	-0,1280	-0,3343	1			
5	0,0617	-0,0215	-0,0190	0,0837	1		
6	-0,0668	0,0683	0,1879	0,1201	-0,0667	1	
7	0,0516	0,0774	-0,0076	-0,1615	0,1054	-0,0375	1

Comentarios:

- Nota: Por idénticas razones, **se efectúa la estimación de la Matriz de Correlaciones**.
- De acuerdo con la segunda matriz, se observa diferencias significativas solo en dos correlaciones (por exceso o defecto): en correlación entre *Agricultural exports y Agricultural imports*, y entre *Agricultural land y Agricultural imports*. La disminución de diferencias significativas daría cuenta de una **mejora en ajuste del modelo**.

ANÁLISIS DE FACTORES

Se presenta a continuación una comparativa del **Análisis de Factores entre un modelo de un único factor y un modelo de dos factores:**

Único factor:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1							
2	0,4052						
3	0,153	0,0035					
4	0,0223	0,3728	0,3107				
5	0,2405	0,2095	0,001	0,0431			
6	0,327	0,2054	0,2141	0,0645	0,2085		
7	0,1294	0,113	0,0107	0,2898	0,0068	0,181	

Dos factores:

Variables	Agricultural land (1)	Arable land (2)	Agricultural exports (3)	Agricultural imports (4)	Agriculture, forestry, and fishing, value added (5)	Employment in agriculture (6)	Rural population (7)
1							
2	0,06						
3	0,1199	0,0384					
4	0,2551	0,128	0,3343				
5	0,0617	0,0215	0,019	0,0837			
6	0,0668	0,0683	0,1879	0,1201	0,0667		
7	0,0516	0,0774	0,0076	0,1615	0,1054	0,0375	

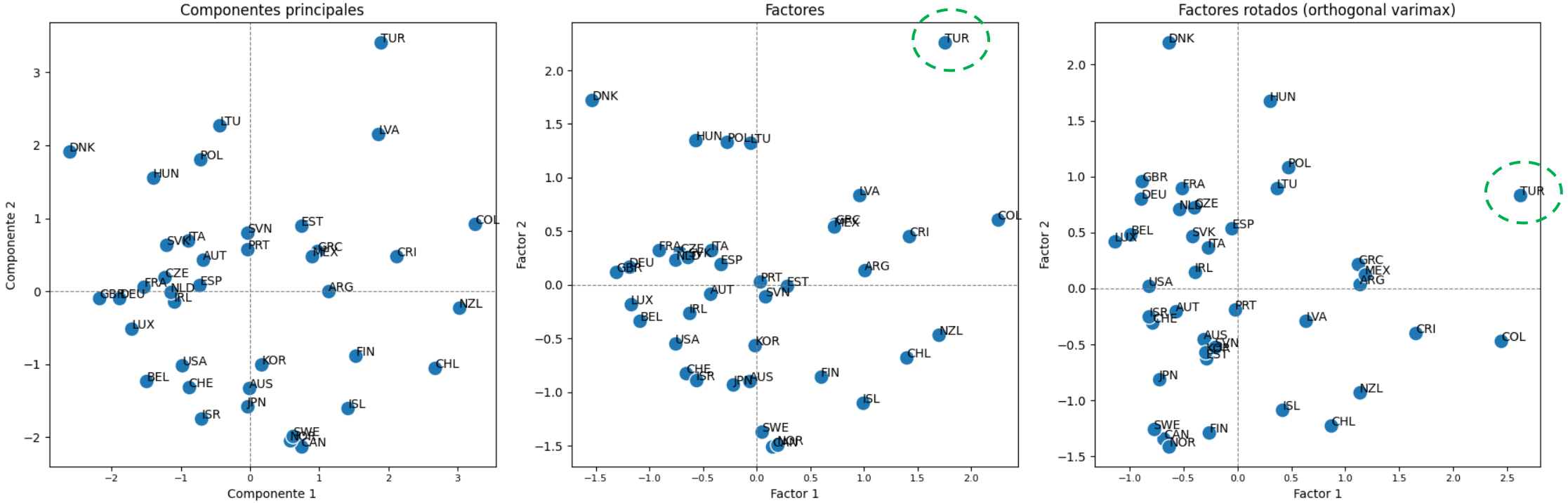
Comentarios:

- Se presenta, de **forma comparativa**, los **valores en términos absolutos** (dado que no interesa si la estimación es por exceso o defecto) **de las matrices de diferencias** de slides anteriores para ambos modelos (único factor vs. dos factores), donde los **valores son mapeados con diferentes escales de colores: verdes** (**buena estimación**) y **rojo** (**mala estimación**).

- Visualizando las mismas, puede argumentarse que la incorporación de un factor adicional al modelo mejora sensiblemente la estimación de la matriz de correlaciones y, ergo, de la matriz de datos original. En este sentido, existe respaldo para concluir que el **modelo de dos factores resultaría ser el más adecuado para representar la estructura de asociación entre las variables.**

PCA Y ANÁLISIS DE FACTORES

Se presentan a continuación **breves comentarios finales en relación con los métodos PCA y Análisis de Factores:**



Comentarios:

- De forma general, se entiende por **PCA y Análisis de Factores** a dos posibles **técnicas de reducción de dimensionalidad de una matriz de datos**, al tiempo intentar **preservar la mayor cantidad posible de las características intrínsecas propias de dicha matriz**. PCA, por su parte, intentará realocar la varianza (entendida esta como las características intrínsecas de la matriz de datos) en unos pocos componentes principales; Análisis Factorial, por su parte, intentará la reconstrucción de las variables a partir de una combinación lineal de variables artificiales, permitiendo en dicho proceso explicar no solo la varianza de los datos sino también sus covariaciones o correlaciones. En este sentido, es de esperar que **ambos métodos no sean contrapuestos sino complementarios**, y es lo que se observa en los gráficos presentes en esta sección:
  - La **visualización** de las primeas dos Componentes Principales y los primeros dos Factores (sin rotar) dan cuenta de una **simetría de resultados** (en diferentes escalas);
  - De la rotación de los Factores podemos observar una mayor intensidad ya sea o bien en Factor 1 o bien en Factor 2, lo que nos brinda una mejora en la **interpretación de los resultados**. Por ejemplo, Turquía: dada la rotación, ¿el componente 1 es el que prima por sobre el componente 2?;
  - Preguntas como la anterior confirmar la **complementariedad** existente entre ambos métodos: la reducción de datos que permite PCA acompañado de la mejora en interpretabilidad de los componentes mediante la rotación de factores.

MUCHAS GRACIAS