
Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes y Big Data

Trabajo Fin de Máster

# Análisis Multivariante de los Indicadores del Desarrollo del Banco Mundial

Autor: Edwin Alexander Betancur Agudelo

Tutor: Jose Luis Vicente Villardón

2022

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

DEPARTAMENTO DE ESTADíSTICA

DOCTOR JOSÉ LUIS VICENTE VILLARDÓN

Profesor Titular del Área de Estadística e Investigación Operativa

de la Universidad de Salamanca

CERTIFICA que **D. Edwin Alexander Betancur Agudelo** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de Máster en Análisis Avanzado de Datos Multivariantes y Big Data presenta con el título, **Análisis Multivariante de los Indicadores del Desarrollo del Banco Mundial**, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a: ___ de septiembre de 2022.

**Doctor José Luis Vicente Villardón**

Los principios teóricos de la economía de mercado y los elementos básicos del liberalismo económico no fueron diseñados, como se creía, por calvinistas y protestantes escoceses, sino por los jesuitas y miembros de la Escuela de Salamanca durante el Siglo de Oro español.

**Friedrich A. Hayek**. Economista austriaco y premio Nobel de economía

Take me back to my San Andrés

To the wave and the coral reefs

Back to be where the sunshine bright

where the sea changes colors day and night

**Maria Cecilia Francis Hall, Miss Chiqui**. Beautiful San Andres (1972)

# Dedicatoria

A Nuris

A Alejandrino

A Isabella y Katrina

A Fernanda

# Agradecimientos

La gratitud es lo que hace posible que reconozcamos el valor, sea de nuestros iguales o de nuestros diferentes, de las cosas grandes o pequeñas, y hasta de nosotros mismos. Aquella virtud que nos permite establecer vínculos sanos y profundos con muchos, con pocos, y además con nosotros mismos.

Es por esto que mis profundos agradecimientos son para mis profesores en este proceso de formación de maestría, creo que son excelentes exponentes de los temas y con una experiencia que enriqueció cada instante. Gracias totales a José Luis Vicente Villardón, Purificación Galindo Villardón, María Purificación Vicente Galindo, Mercedes Sánchez Barba, Carmen Patiño Alonso, Antonio Blazquez Zaballos, Francisco Javier Martín Vallejo, Lorenzo Albas Aso, Mariela Alexandra González, María José Fernández, Carmelo Ávila Zarza, María Cortes Rodríguez, Ana Belén Nieto Librero, Elisa Frutos Bernal, Francisco Javier Delgado Álvarez, Nathalia Diazibeth Tejedor Flores, Mitzi Cubilla Montilla, Carmen Cecilia Rodríguez Martínez, Daniel Caballero Juliá, Miguel Infestas Maderuelo, Fernando Mallo Fernández, Nerea González García, Michael Thon, Rogelio González Sarmiento, y José Antonio Frías.

Doy gracias también y de manera importante a mis tres colegas de la maestría que en diferentes momentos del año académico 2021-2022 fueron clave para el éxito de mi proceso formativo. Gracias a Nansi López Valverde de España, a Nooshina Marafet de Irán, y a Juan Luis Apaza de Perú.

Agradecer a la Universidad de Salamanca por permitirme ser parte de ella, centro importante de pensamiento.

# Resumen

Desde muchas décadas atrás, el Banco Mundial viene realizando esfuerzos importantes para analizar la evolución temporal del desarrollo de un país tomando en cuenta diferentes tópicos como las políticas económicas, la salud, la pobreza, medio ambiente, educación, etc. El objetivo de este trabajo fue proveer una perspectiva multivariante del desarrollo de los países a través de sus indicadores, proporcionar evidencia del impacto del poder discriminante de estos. Además, exploramos los métodos estadísticos de tres vías Sparse STATIS-dual y STATIS-dual, utilizados para la reducción de dimensionalidad. El Análisis de los indicadores tomados de 265 economías, procedentes de la compilación anual que hace el Banco Mundial, pone de manifiesto diferencias entre resultados dependiendo de dos aspectos fundamentales: Primero, el proceso de exclusión de variables explicativas, dada la cantidad de indicadores a seleccionar inicialmente de un total de 1445 que hoy ya tiene el Banco Mundial, y sumado a un problema de datos faltantes; y segundo, el método de análisis, dado que se restringen o no las cargas de los indicadores y esto se refleja en las componentes principales; y sugiere, bajo el principio de parsimonia, seleccionar la mejor hipótesis entre varias igualmente soportadas por los datos, y mejorar la interpretación final. El resultado encontró que los países menos desarrollados están asociados negativamente a la participación de mujeres en un ambiente de negocios, y asociados positivamente a la prevalencia del VIH en mujeres entre 15-24 años; esto sugiere que las mujeres están vinculadas al desarrollo de un país y en ese sentido han de desarrollarse políticas públicas.

**Palabras Clave**: Indicadores del desarrollo Mundial, Políticas públicas, Análisis multivariante, Sparse STATIS dual, Sparse HJ-Biplot, Biplot.

# Abstract

For many decades, the World Bank has been making significant efforts to analyze the temporal evolution of a country's development, taking into account different issues such as economic policies, health, poverty, the environment, education, etc. The aim of this work was to provide a multivariate perspective of the development of countries through their indicators, showing the impact of their discriminating power. In addition, we explore the three-way statistical methods Sparse STATIS-dual and STATIS-dual, used for dimensionality reduction. The analysis of the indicators taken from 265 economies, from the annual compilation carried out by the World Bank, reveals differences between the results based on two fundamental aspects: First, the process of exclusion of explanatory variables, given the number of indicators to select initially from a total of 1445 that the World Bank already has today, and added to a problem of lack of data; and second, the method of analysis, given that the loadings of the indicators be restricted or not and this is reflected in their principal components; and suggests, under the principle of parsimony, selecting the best hypothesis among several equally supported by the data, and improving the final interpretation. The result found that less developed countries are negatively associated with the participation of women in a business environment, and positively associated with the prevalence of HIV in women between 15-24 years; this suggests that women are linked to the development of a country and in this sense, public policies must be developed.

**Keywords**: World Development Indicators, Public Policies, Multivariate Analysis, Sparse STATIS Dual, Sparse HJ-Biplot, Biplot.

# Índice

# Índice de Tablas

# Índice de Figuras

# Introducción

En 1995, la Cumbre Mundial sobre Desarrollo Social de las Naciones Unidas celebrada en Copenhague (https://www.un.org/en/desa), identificó tres ejes temáticos: la erradicación de la pobreza, la generación de empleo y la integración social, en la contribución a la creación de una comunidad internacional que permita la construcción de sociedades seguras, justas, libres y armoniosas que ofrezcan oportunidades y mejores niveles de vida para todos. Hasta hoy se tienen identificado ya no tres, sino cuatro ejes temáticos: Inclusión social, empleo y trabajo decente, desigualdad, y la erradicación de la pobreza. La pobreza implica más que la falta de ingresos y recursos productivos para garantizar medios de vida sostenibles. Sus manifestaciones incluyen el hambre y la desnutrición, el acceso limitado a la educación y otros servicios básicos, la discriminación y exclusión social, así como la falta de participación en la toma de decisiones. Varios grupos sociales soportan una carga desproporcionada de pobreza.

Recientemente, en Julio 2022, en su resolución 72/233, la Asamblea General proclamó el Tercer Decenio de las Naciones Unidas para la Erradicación de la Pobreza (2018-2027). También consideró que el tema del Tercer Decenio, debería ser “Acelerar las acciones globales para un mundo sin pobreza”, en consonancia con la Agenda 2030 para el Desarrollo Sostenible. Se estima que 783 millones de personas vivían con menos de $1,90 al día en 2013, en comparación con 1.867 millones de personas en 1990. El crecimiento económico en los países en desarrollo ha sido notable desde 2000, con crecimiento más rápido del producto interno bruto (PIB) per cápita que los países avanzados. Este crecimiento económico ha impulsado la reducción de la pobreza y la mejora de los niveles de vida. También se

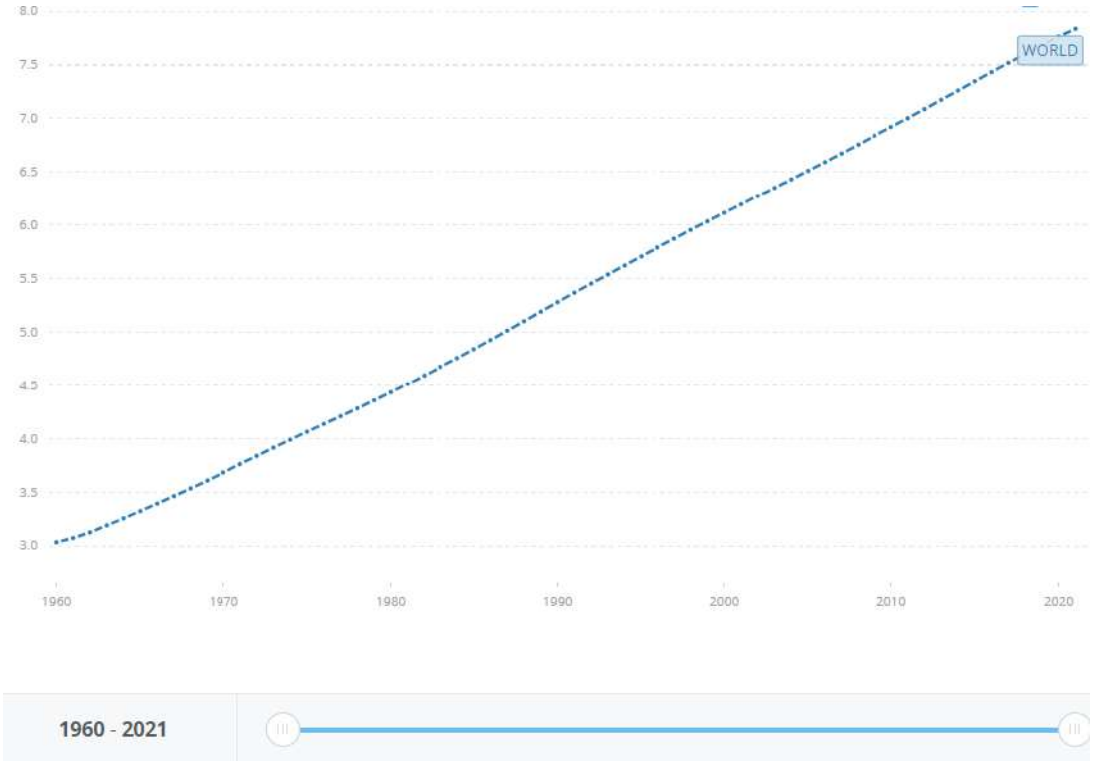
han registrado logros en áreas como la creación de empleo, la igualdad de género, la educación y la atención de la salud, las medidas de protección social, la agricultura y el desarrollo rural, y la adaptación y mitigación del cambio climático.

De otro lado, la cooperación entre el Grupo del Banco Mundial (GBM) (https://www.worldbank.org/en/home) y las Naciones Unidas (ONU) existe desde la fundación de las dos organizaciones (1944 y 1945, respectivamente) y se centra en áreas económicas y sociales de interés mutuo, y de ellos está el reducir la pobreza, promover el desarrollo sostenible e invertir en las personas. La Misión del Grupo Banco Mundial tiene dos objetivos ambiciosos: Acabar con la pobreza extrema y promover la prosperidad compartida. El Banco Mundial cada año publica los Indicadores del desarrollo mundial, WDI por sus siglas en ingles World Development Indicators, que proporciona una compilación de estadísticas relevantes, de alta calidad y comparables internacionalmente sobre el desarrollo global y la lucha contra la pobreza. Su objetivo es ayudar a usuarios de todo tipo (formuladores de políticas, estudiantes, analistas, profesores, administradores de programas y ciudadanos) a encontrar y utilizar datos relacionados con todos los aspectos del desarrollo, incluidos aquellos que ayudan a monitorear y comprender el progreso hacia los dos objetivos. Se utilizan seis secciones temáticas para organizar los indicadores: visión del mundo, personas, medio ambiente, economía, estados y mercados, y vínculos globales.

Los WDI de hecho son la colección principal de indicadores de desarrollo del Banco Mundial, compilados a partir de fuentes internacionales reconocidas oficialmente, y disponibles a través de la Iniciativa de Datos Abiertos del Banco Mundial, y presenta los datos de desarrollo global más actualizados y precisos disponibles, e incluye datos nacionales, regionales y globales. Se puede tener

acceso a cada indicador vía URL conociendo el identificador del indicador, por ejemplo SP.POP.TOTL para el caso del indicador de población mundial (http://data.worldbank.org/indicator/SP.POP.TOTL).

Figura 1 *Gráfico lineal del indicador de Total Population.*

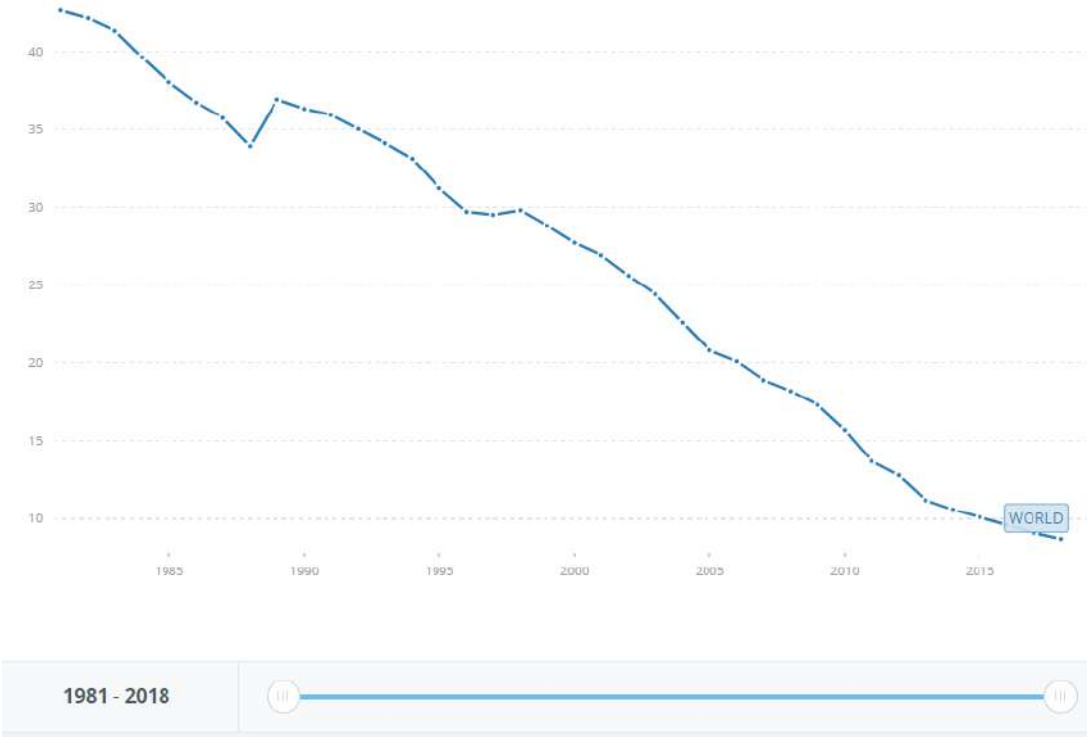



*Nota.* Tomado de *Population, Total* [Line Chart], por World Bank, Recuperado 20 Mayo de 2022. (http://data.worldbank.org/indicator/SP.POP.TOTL).

También se puede acceder a cualquier otro indicador (https://data.worldbank.org/indicator/), e inclusive agregar en la búsqueda la economía de interés (país o agrupaciones de estos) y se obtiene la grafica correspondiente haciendo descargable un CSV, Excel o XML del indicador para esa economía para todos los años. U otra manera es ir a las tablas por tema (http://wdi.worldbank.org/tables). Por otro lado, se puede lograr información ya sea por país, o mejor aun, de las economías agrupadas ya sea por región, por ingresos,

etc. (https://data.worldbank.org/country), ver que países lo conforman, los valores de algunos indicadores y gráficos de línea del mismo, y se puede descargar, en formato CSV, Excel o XML, tanto un indicador para todos los años, o todos los indicadores para esa economía.

Figura 2 *Gráfico lineal del indicador de Poverty headcount ratio.*

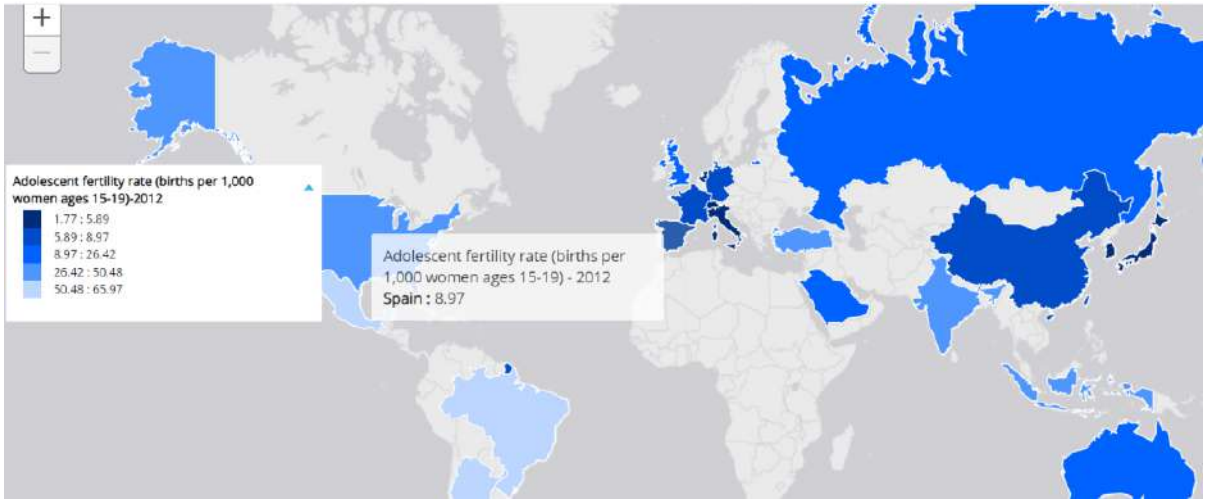



*Nota.* Tomado de *Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)* [Line Chart], por World Bank, Recuperado 20 Mayo de 2022. (https://data.worldbank.org/indicator/SI.POV.DDAY?view=chart).

Actualmente se puede descargar versiones masivas de archivos Excel y CSV de la base de datos de WDI directamente, incluidos los metadatos (https://datatopics.worldbank.org/world-development-indicators/), en el apartado de 'Access Data', opciones 'Excel download' y 'CSV download'. Hay opción de consulta en línea, 'Query database' en el apartado 'Open Data & DataBank', para obtener visualizaciones en mapas, por países o grupos de ellos, etc, y hacer una vista previa

de la base de datos según filtros (https://databank.worldbank.org/source/world-development-indicators#). Es importante anotar que al seleccionar un filtro para todos los países, todas las series y todos los años, indica para el Mapa en línea que : *Your selection (23781464 cells) exceeds the maximum report limit (2500000 cells). Please use "Download Options" or remove some variables to continue.* Del ejemplo automático en la web, para un grupo de 20 países, ocasiones 10 (años), y series (variables, indicadores) 55, se tiene este mapa por indicador cada vez (por ejemplo tasa de fertilidad adolescente).

Figura 3 *Mapa mundial con Indicador Adolescent fertility rate (births per 1,000 women ages 15-19).*

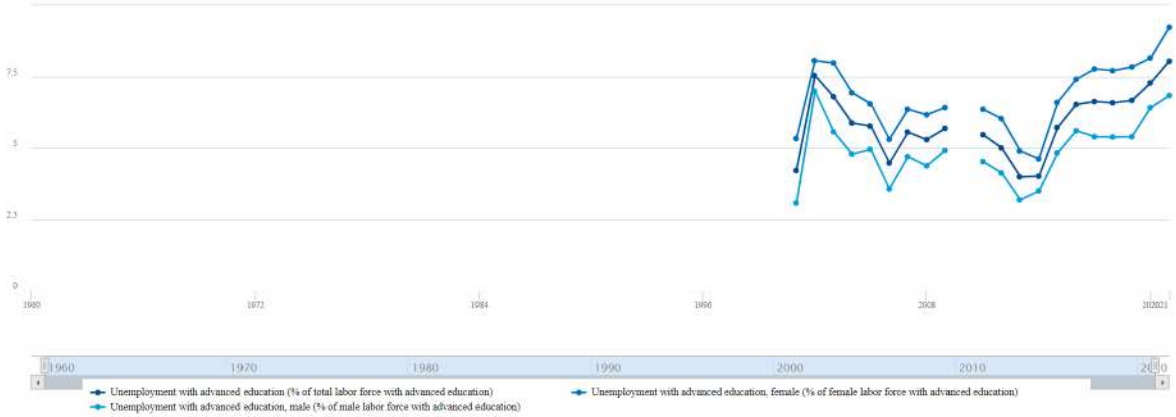



*Nota.* Tomado de *Adolescent fertility rate (births per 1,000 women ages 15-19)* [Map Chart], por World Bank, Recuperado 20 Mayo de 2022. (https://databank.worldbank.org/source/world-development-indicators).

En esta web, se permite aplicar filtros específicos, otra ilustración será seleccionar WDI como base de datos, en países a Latinamerica & The Caribbean, en Series (indicadores) seleccionar 3: Unemployment with advanced education (% of total labor force with advanced education), Unemployment with advanced education, female (% of female labor force with advanced education), y

Unemployment with advanced education, male (% of male labor force with advanced education), y en años (time) seleccionar todos (desde 1960 hasta 2021), y luego descargar la imagen en format SVG vector image (que es la mejor calidad que entrega, superando a PNG, JPEG, y PDF) para este grafico de linea.

Figura 4 *Gráfico lineal de 3 Indicadores de Unemployment with advanced education (female, male, total) para Latinamerica & The Caribbean entre 1960-2021.*




*Nota*: Tomado de *Unemployment with advanced education (%total, %male, %female labour forcé with advanced education)* [Line Chart], por World Bank, Recuperado 20 Mayo de 2022. (https://databank.worldbank.org/source/world-development-indicators).

También a partir de las API (de sus siglas en ingles, Application Programming Interface) que se han desarrollado, se accede a consultas (Queries) de la base de datos desde otras plataformas de software, como es el caso del software estadístico R de código abierto. Desde el link inicial (https://datatopics.worldbank.org/world-development-indicators/), en el apartado 'API documentation' se accede al apartado 'Developer Information', (https://datahelpdesk.worldbank.org/knowledgebase/topics/125589-developer-information) donde se tiene acceso a todos los desarrollos API que tiene el Banco Mundial para los Queries en WDI, tales como: New Features and Enhancements in

the V2 API, API Basic Call Structures, Country API Queries, Aggregate API Queries, Indicator API Queries, Topic API Queries, Advanced Data API Queries, Metadata API Queries, y SDMX API Queries. Como ejemplo acceder a la información de un indicador, NY.GDP.MKTP.CD, en formato JSON (JavaScript Object Notation), el API sera de la forma: http://api.worldbank.org/v2/indicators/NY.GDP.MKTP.CD?format=json, o como ejemplo para el caso de todos los indicadores de un país, Brasil 'br': http://api.worldbank.org/v2/country/br?format=json, o para el caso de todos los tópicos: http://api.worldbank.org/v2/topic?format=json, o todos los indicadores de un tópico: http://api.worldbank.org/v2/topic/5/indicator o http://api.worldbank.org/v2/indicator?topic=5.

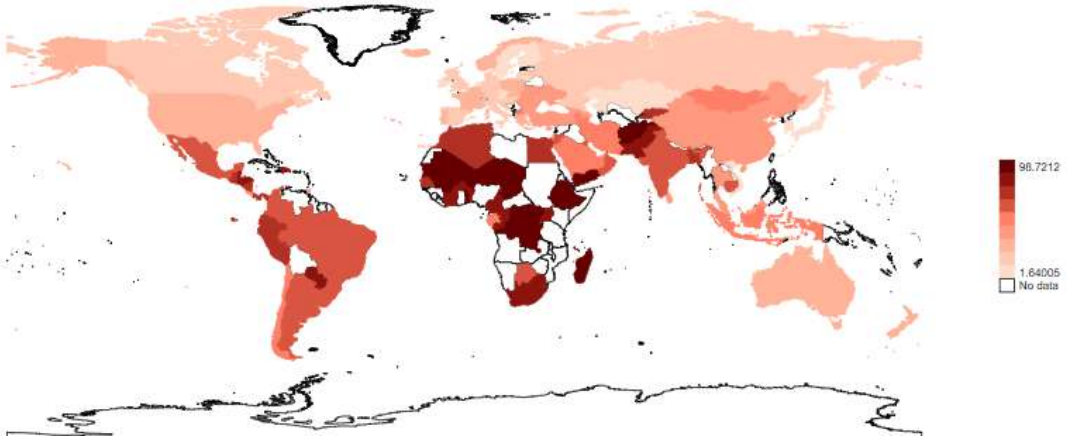
Estos desarrollos API soportan 2 formas básicas para los Queries que hacen: basado en estructura URL, como ejemplo para un Income clasificado como LIC que países lo tienen (http://api.worldbank.org/V2/incomeLevel/LIC/country) y otro basado en estructura argumento, ejemplo para los países filtrar los que tienen clasifiaicon de Income como LIC (http://api.worldbank.org/V2/country?incomeLevel=LIC). Ver las anotaciones de aclaración para otros ejemplos y formatos de descarga en el Data Help Desk del Banco Mundial (https://datahelpdesk.worldbank.org/knowledgebase/articles/898581-api-basic-call-structures). Esto es clave porque pueden ser llamados desde el software estadístico R de código abierto (a través de los librerías httr y jsonlite, y comandos GET y fromJSON, aplicable a archivos en formato JSON: JavaScript Object Notation) para acceder y recopilar datos. Estos API son desarrollados dentro del proyecto Drupal, que es una plataforma que se usa para añadir, editar o eliminar contenido de una página web, es un sistema de gestión de contenidos, CMS por sus siglas en ingles,

y de cogido abierto. Se puede encontrar más información de Drupal y la API para Banco Mundial (https://www.drupal.org/project/wbapi), y de donde es descargable dicho paquete. Uno de estos API es el 'Data Catalog API' que proporciona información sobre los miles de conjuntos de datos relevantes para el desarrollo mundial (https://datacatalog.worldbank.org/home), siendo el de interés en este caso WDI (https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators), donde se puede hacer descarga directa desde ese catalogo las versiones en Excel o en CSV.

Por otro lado, el Banco Mundial anima a desarrollar aplicaciones de software que utilicen y agreguen valor a los conjuntos de datos. Ya se han desarrollado muchas aplicaciones de terceros, incluidos módulos de integración tanto para Drupal, R, Python y STATA. Uno de los más conocidos y usados es el WBOPENDATA (https://github.com/jpazvd/wbopendata), módulo de STATA (paquete de computación estadística ampliamente utilizado en el mundo empresarial y académico) el cual se basa en las principales colecciones de datos del Banco Mundial, y entre otras permite acceder a WDI. El módulo se conecta a la API de datos abiertos del Banco Mundial y brinda acceso directo a la última versión de los datos del Banco a través de la interfaz de STATA; no es necesario descargar ni administrar los datos innecesariamente. En dicho modulo, se admiten Actualmente se admiten tres posibles opciones de descarga:

- País - todos los indicadores para todos los años para un solo país.
- Tema - todos los indicadores dentro de un tema específico, para todos los años y todos los países.
- Indicador - todos los años para todos los países para un solo indicador.

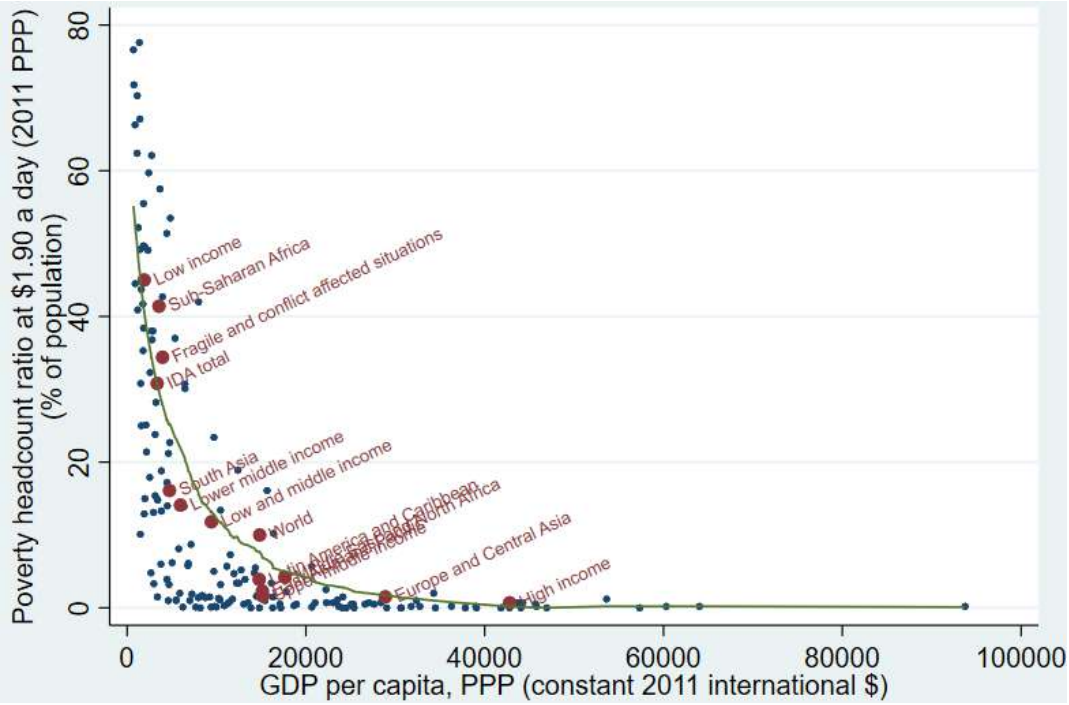
Figura 5 *Mapa mundial con el indicador: Learning poverty: Share of Children at the End-of-Primary age below minimum reading proficiency adjusted by Out-of-School Children (%).*




*Nota*: Tomado *de Learning poverty: Share of Children at the End-of-Primary age below minimum reading proficiency adjusted by Out-of-School Children (%).* [Map Chart], por Joao Pedro Azevedo, 2011. (https://github.com/jpazvd/wbopendata/blob/master/doc/wbopendata.md).

Una de las ventajas importantes de WBOPENDATA es que facilita la reproducibilidad de cualquier análisis utilizando WDI en STATA, así como el seguimiento de conjuntos de datos antiguos, y permite también hacer mapas.

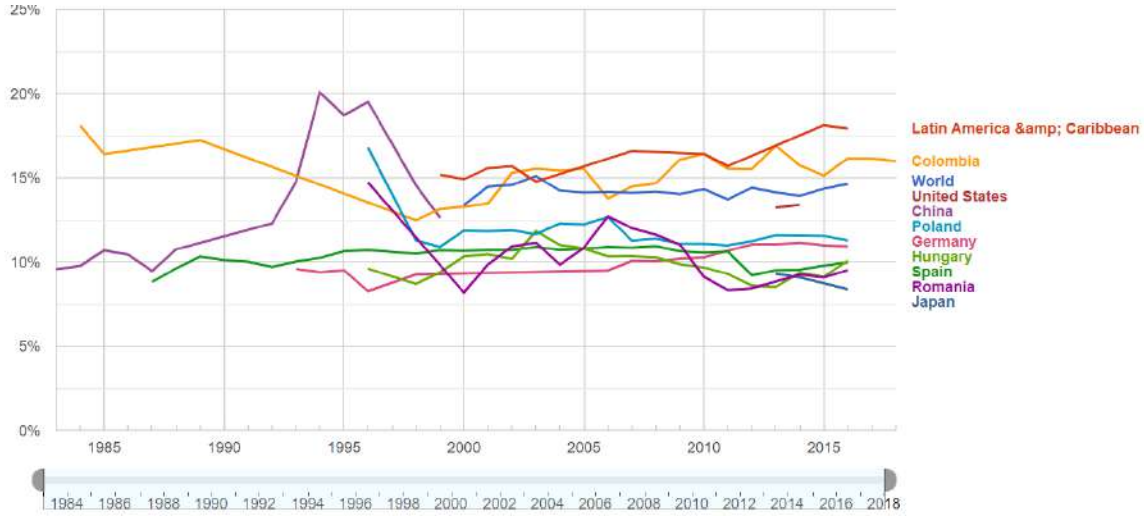
Figura 6 *Comparacion indicadores: Poverty headcount ratio at $1.90 a day Vs. GDP per capita based on purchasing power parity (PPP).*




*Nota*: Tomado de *Example 11, Bencharmk latest poverty levels by percapital income, highlighting regional* [Line Chart], por Joao Pedro Azevedo, 2011. (https://github.com/jpazvd/wbopendata/blob/master/doc/wbopendata.md).

Dentro de las aplicaciones de terceros, y en especifico para WDI, contamos con Explorador de datos públicos de Google (en Ingles: Google's Public Data Explorer) que utiliza las herramientas de Google, incluidos los gráficos de movimiento, para visualizar y explorar datos del Banco Mundial y además muchos otros proveedores. Para el caso de WDI se accede desde google (https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_), donde igualmente se accede a un tópico, se selecciona una serie (variable), y los países de interés, y muestra para todos las ocasiones (años).

Figura 7 *Grafico lineal de Indicador: Government expenditure on education (% of government expenditure), para las economias: Latinamerica & The Caribbean, Colombia, World, United States, China, Poland, Germany, Hungary, Spain, Romania, y Japan, entre 1980-2018.*

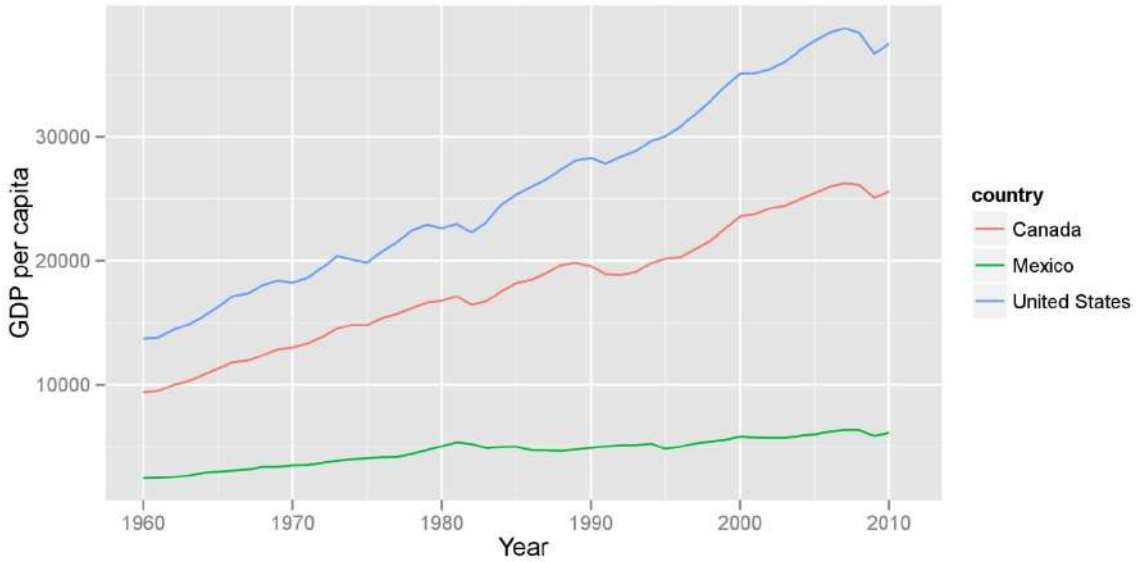



*Nota*: Tomado de *Government expenditure on education (% of government expenditure)* [Line Chart], por Google's Public Data Explorer, Recuperado 15 de Julio de 2022. (https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_ ).

El Google's Public Data Explorer también permite ver año a año, el indicador por cada país tanto en un diagrama descendente de barras (bar chart) como una animación, así como en un mapa (map chart) igualmente animado que utiliza círculos (bubble chart) grandes o pequeños para ilustrar la evolución en el tiempo, y permite escalas logarítmicas. La salida grafica de esta página con el resultado no toma mucho tiempo, solo permite un indicador a la vez.

Si se usa el software estadístico R de código abierto, se tiene par módulos similares disponibles: el modulo WDI y el wbstats, ambos módulos ofrecen opciones para leer datos del Banco Mundial directamente, y ambos paquetes se integran con ggplot2 para gráficos (https://github.com/vincentarelbundock/WDI).

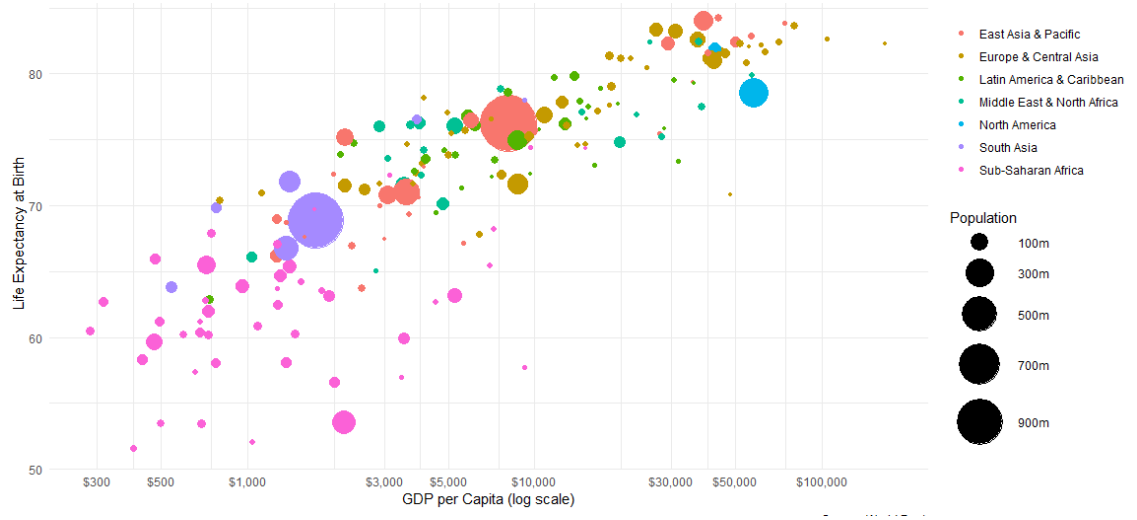
Figura 8 *Grafico lineal de Indicador: GDP per capita, para las economias: Canada, Mexico, United States, entre 1960-2010.*




*Nota*: Tomado de *Download and use the data* [Line Chart], por vincentarelbundock en Repositorio GitHub, 2018. (https://github.com/vincentarelbundock/WDI)

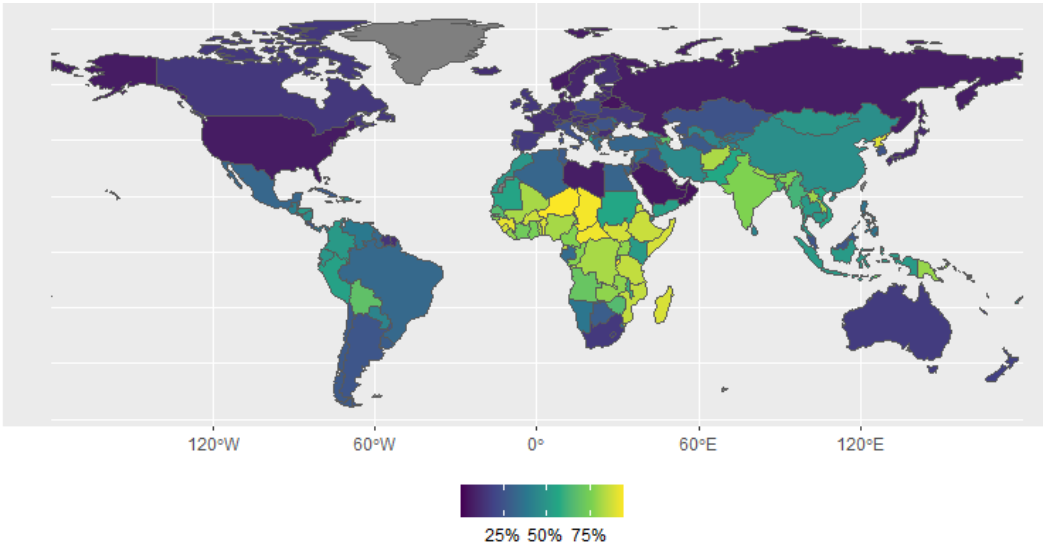
En el caso de wbstats, es un paquete para búsqueda y descarga de datos desde la API Banco Mundial (https://github.com/gshs-ornl/wbstats).

Figura 9 *Grafico de Hans Rosling's Gapminder que compara dos Indicadores: Life Expectancy at Birth vs GDP per capita, para varias economías regionales, inciando en 2016.*




*Nota*: Tomado de *An Hans Rosling's Gapminder using wbstats* [Gapminder chart], por Jesse Piburn en Repositorio GitHub, 2016. (https://github.com/gshs-ornl/wbstats)

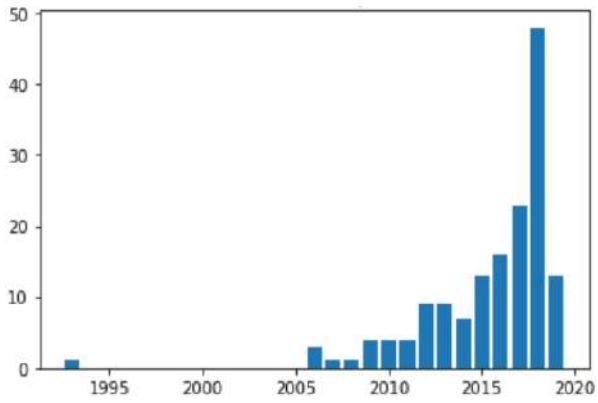
Figura 10 *Mapa mundial con el indicador Self-employed, total (% of total employment) (modeled ILO estímate), generado en R Studio librería wbstats.*




*Nota*: Tomado de *Self-employed, total (% of total employment) (modeled ILO estímate)* [Map Chart], por Jesse Piburn en repositorio GitHub, 2016. (https://github.com/gshs-ornl/wbstats)

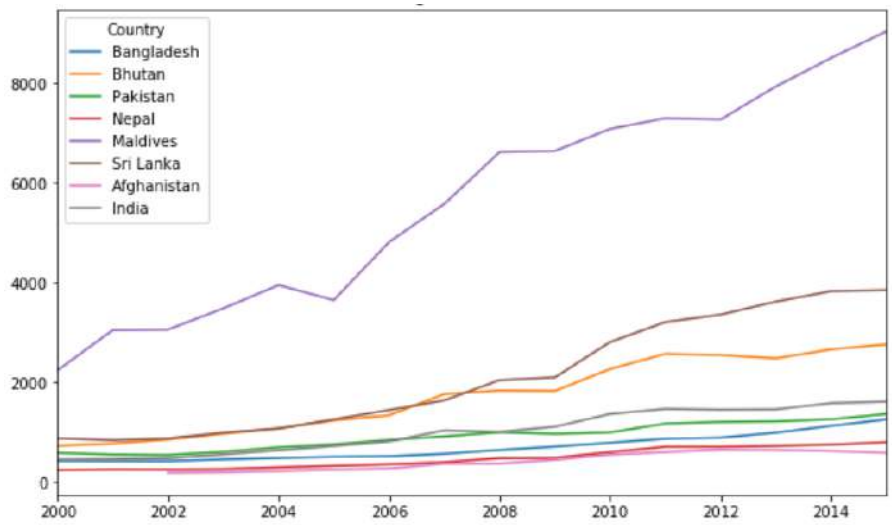
Python cuenta con el paquete wbgapi (https://pypi.org/project/wbgapi/), para acceso a datos del banco mundial, y siendo compatible con el paquete pandas, es fácil usar las funciones gráficas integradas o cualquier paquete de gráficos que prefiera (ggplot, seaborn, etc.). A continuación par salidas de visualización.

Figura 11 *Gráfico de barras que cuenta el número de economías con no-vacios en el indicador Poverty headcount ratio at national poverty lines (% of population), MRV Most Recent Value, en el periodo 1993-2019.*




*Nota*: Tomado de *Distribution of Poverty Rate, MRVs* (s.f.) [Bar Chart], por Tim Herzog en repositorio nbviewer Jupyter, Recuperado Agosto 28 de 2022. (https://nbviewer.org/github/tgherzog/wbgapi/blob/master/examples/wbgapi-cookbook.ipynb)

Figura 12 *Gráfico lineal del indicador Income growth en la región de South Asia, 2000-2016.*




*Nota*: Tomado de *Income growth in South Asia* [Line Chart], por Tim Herzog, 2021. (https://blogs.worldbank.org/opendata/introducing-wbgapi-new-python-package-accessing-world-bank-data)

Wbdata de Python (https://pypi.org/project/wbdata/) es una sencilla interfaz para buscar y solicitar información de las diversas bases de datos del Banco Mundial, ya sea como un diccionario que contiene metadatos completos o como un Data Frame del paquete pandas.

En general, hay interés en la visualización de las series (variables o indicadores), pero no se propone aun el análisis multivariante de WDI. La generalidad es presentar gráficos del tipo dispersión para comparar indicadores, mapas para localizar grupos de países, o gráficos lineales de tiempo para un indicador en uno o varios países o economías regionales.

En 2010 WDI paso de tener 339 indicadores a más de 1100, para cerca de 200 economías y 25 grupos de países, y para 2022, WDI ya contiene más de 1400 indicadores de series temporales para 217 economías y más de 40 grupos de países, con datos para muchos indicadores que se remontan a más de 50 años.

Esto hace notar que es un cubo de datos bastante grande e implica usar técnicas más sofisticadas para su análisis.

En este trabajo analizaremos la covarianza entre indicadores del desarrollo mundial WDI, para los países o grupos de ellos como economías regionales, y ver si es estable en los años, para evaluar el desarrollo de los países relacionado con diferentes indicadores y no exclusivamente con la falta de ingresos, utilizando la técnica de Sparse STATIS-dual, porque interesa aplicar restricciones para reducir dimensionalidad, analizar k-tablas de tiempo y enfatizar las relaciones entre los indicadores, y comparar resultados con la técnica STATIS-dual. A su vez se da una rápida mirada desde el dos técnicas de tres modos (PTA y MFA), y una observacino complementaria desde dos técnicas de dos modos, Sparse HJ-Biplot y desde el Biplot inducido desde STATIS, ambos para ampliar la interpretación.

Se estructura de la siguiente manera este trabajo: Se comienza con una introducción o estado del arte del caso a estudiar, se presentan los objetivos; a continuación se presenta el material y métodos, en cuyo capítulo se presenta la descripción de los datos ampliamente, se explica la selección de los países, de las características de los indicadores del desarrollo mundial y selección de los ejes temáticos de los indicadores que sean de interés, y hablaremos de los métodos estadísticos principales que utilizaremos en el estudio. Se presentarán después los resultados; y se termina con las conclusiones más relevantes.

# Objetivos

El objetivo general de este trabajo es realizar, bajo un enfoque estadístico multivariante, la comparación de técnicas para el análisis de tablas de tres vías, mediante la aplicación a los datos reales: Indicadores del Desarrollo del Banco Mundial WDI, considerando la posibilidad de fijar políticas públicas relacionadas.

Los objetivos específicos son:

- Comparar diferentes procedimientos utilizados para disminuir dimensionalidad del conjunto de indicadores del desarrollo del banco mundial WDI y analizar si proporcionan resultados equivalentes.
- Aproximar una interpretación práctica a partir de los indicadores con mayor poder discriminante, identificar los más relevantes, que diferencian a cada economía.
- Estudiar la posición de las economías con respecto a los demás y a los indicadores.
- Estudiar las modificaciones en la posición de las distintas economías a lo largo de los años.

# Material y métodos

Este trabajo de fin de Máster se centra en el análisis avanzado multivariante de los Indicadores del Desarrollo Mundial WDI, base de datos que compila información de los países relacionados con diferentes tópicos y no exclusivamente con la falta de ingresos, con el fin de estudiar la eliminación de algunos indicadores que no aporten información relevante.

## Las Economías

En WDI están clasificados los 189 países miembros del Banco Mundial, más otras 28 economías con poblaciones de más de 30.000 habitantes, para un total de 217. Las principales clasificaciones proporcionadas en la base de datos para formar grupos son por región geográfica, por ingresos, y por las categorías de préstamos operativos que fija el Banco Mundial. En WDI, el término país, usado indistintamente con economía, no implica independencia política sino que se refiere a cualquier territorio sobre el cual las autoridades reportan estadísticas sociales o económicas separadas.

En este trabajo se mantendrán los nombres en ingles tanto de las economías como de los indicadores como están descritos en la base original descargable de WDI. Estas son las tablas de clasificación actualmente usadas en WDI.

Tabla 1 *Tablas de clasificación del WDI.*

| BY REGION (7) | 218 |
|---|---|
| East Asia and Pacific | 38 |
| Europe and Central Asia | 58 |
| Latin America & the Caribbean | 42 |
| Middle East and North Africa | 21 |
| North America | 3 |
| South Asia | 8 |
| Sub-Saharan Africa | 48 |

| BY INCOME (4) | 217 |
|---|---|
| Low-income economies | 27 |
| Lower-middle-income economies | 56 |
| Upper-middle-income economies | 55 |
| High-income economies | 79 |

| BY LENDING (3) | 144 |
|---|---|
| IDA | 59 |
| Blend | 15 |
| IBRD | 70 |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

East Asia and Pacific en el listado de la web del Banco Mundial incluye a Taiwan, pero en la base de datos esta economía no está incluida. Con esto realmente son 217 países en 7 regiones geográficas principales. Venezuela se ha desclasificado temporalmente a partir de julio de 2021 (para el caso de clasificación por ingresos, Income) a la espera de la publicación de las estadísticas de cuentas nacionales revisadas. La clasificación por Regiones en WDI es como sigue.

Tabla 2 *Países de la Región Este de Asia y Pacifico*

| EAST ASIA AND PACIFIC | | 38 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| American Samoa | ASM | Korea | KOR | Papua New Guinea | PNG |
| Australia | AUS | Lao PDR | LAO | Philippines | PHL |
| Brunei | BRN | Macao SAR, China | MAC | Samoa | WSM |
| Cambodia | KHM | Malaysia | MYS | Singapore | SGP |
| China | CHN | Marshall Islands | MHL | Solomon Islands | SLB |
| Fiji | FJI | Micronesia | FSM | **Taiwan, China** | - |
| French Polynesia | PYF | Mongolia | MNG | Thailand | THA |
| Guam | GUM | Myanmar | MMR | Timor-Leste | TLS |
| Hong Kong SAR, China | HKG | Nauru | NRU | Tonga | TON |
| Indonesia | IDN | New Caledonia | NCL | Tuvalu | TUV |
| Japan | JPN | New Zealand | NZL | Vanuatu | VUT |
| Kiribati | KIR | Northern Mariana Islands | MNP | Vietnam | VNM |
| Dem. People's Rep. Korea | PRK | Palau | PLW | | |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Taiwan no está en la lista, es decir no hay indicadores para esa economía. El Banco Mundial explica la condición especial de esta economía (https://datahelpdesk.worldbank.org/knowledgebase/articles/114933-where-are-your-data-on-taiwan)

### Tabla 3 *Países de la Región Europa y Asia Central*

| EUROPE AND CENTRAL ASIA | | 58 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Albania | ALB | Gibraltar | GIB | Norway | NOR |
| Andorra | AND | Greece | GRC | Poland | POL |
| Armenia | ARM | Greenland | GRL | Portugal | PRT |
| Austria | AUT | Hungary | HUN | Romania | ROU |
| Azerbaijan | AZE | Iceland | ISL | Russia | RUS |
| Belarus | BLR | Ireland | IRL | San Marino | SMR |
| Belgium | BEL | Isle of Man | IMN | Serbia | SRB |
| Bosnia and Herzegovina | BIH | Italy | ITA | Slovak Republic | SVK |
| Bulgaria | BGR | Kazakhstan | KAZ | Slovenia | SVN |
| Channel Islands | CHI | Kosovo | XKX | Spain | ESP |
| Croatia | HRV | Kyrgyz Republic | KGZ | Sweden | SWE |
| Cyprus | CYP | Latvia | LVA | Switzerland | CHE |
| Czech Republic | CZE | Liechtenstein | LIE | Tajikistan | TJK |
| Denmark | DNK | Lithuania | LTU | Turkey | TUR |
| Estonia | EST | Luxembourg | LUX | Turkmenistan | TKM |
| Faroe Islands | FRO | Moldova | MDA | Ukraine | UKR |
| Finland | FIN | Monaco | MCO | United Kingdom | GBR |
| France | FRA | Montenegro | MNE | Uzbekistan | UZB |
| Georgia | GEO | Netherlands | NLD | | |
| Germany | DEU | North Macedonia | MKD | | |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

### Tabla 4 *Países de la Región Latinoamérica y el Caribe*

| LATIN AMERICA AND THE CARIBBEAN | | | 42 | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Antigua and Barbuda | ATG | Curaçao | CUW | Paraguay | PRY |
| Argentina | ARG | Dominica | DMA | Peru | PER |
| Aruba | ABW | Dominican Republic | DOM | Puerto Rico | PRI |
| The Bahamas | BHS | Ecuador | ECU | Sint Maarten (Dutch part) | SXM |
| Barbados | BRB | El Salvador | SLV | St. Kitts and Nevis | KNA |
| Belize | BLZ | Grenada | GRD | St. Lucia | LCA |
| Bolivia | BOL | Guatemala | GTM | St. Martin (French part) | MAF |
| Brazil | BRA | Guyana | GUY | St. Vincent and the Grenadines | VCT |
| British Virgin Islands | VGB | Haiti | HTI | Suriname | SUR |
| Cayman Islands | CYM | Honduras | HND | Trinidad and Tobago | TTO |
| Chile | CHL | Jamaica | JAM | Turks and Caicos Islands | TCA |
| Colombia | COL | Mexico | MEX | Uruguay | URY |
| Costa Rica | CRI | Nicaragua | NIC | Venezuela | VEN |
| Cuba | CUB | Panama | PAN | Virgin Islands | VIR |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Tabla 5 *Países de la Región Oriente Medio y Norte de África, Norte América, y de Sur Asia*

| MIDDLE EAST AND NORTH AFRICA | | 21 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Algeria | DZA | Jordan | JOR | Qatar | QAT |
| Bahrain | BHR | Kuwait | KWT | Saudi Arabia | SAU |
| Djibouti | DJI | Lebanon | LBN | Syrian Arab Republic | SYR |
| Egypt | EGY | Libya | LBY | Tunisia | TUN |
| Iran | IRN | Malta | MLT | United Arab Emirates | ARE |
| Iraq | IRQ | Morocco | MAR | West Bank and Gaza | PSE |
| Israel | ISR | Oman | OMN | Yemen | YEM |
| NORTH AMERICA | | 3 | | | |
| Country | Code | Country | Code | Country | Code |
| Bermuda | BMU | Canada | CAN | United States | USA |
| SOUTH ASIA | | 8 | | | |
| Country | Code | Country | Code | Country | Code |
| Afghanistan | AFG | India | IND | Pakistan | PAK |
| Bangladesh | BGD | Maldives | MDV | Sri Lanka | LKA |
| Bhutan | BTN | Nepal | NPL | | |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Tabla 6 *Países de la Región África Sub-sahariana*

| SUB-SAHARAN AFRICA | | 48 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Angola | AGO | Ethiopia | ETH | Niger | NER |
| Benin | BEN | Gabon | GAB | Nigeria | NGA |
| Botswana | BWA | The Gambia | GMB | Rwanda | RWA |
| Burkina Faso | BFA | Ghana | GHA | São Tomé and Principe | STP |
| Burundi | BDI | Guinea | GIN | Senegal | SEN |
| Cabo Verde | CPV | Guinea-Bissau | GNB | Seychelles | SYC |
| Cameroon | CMR | Kenya | KEN | Sierra Leone | SLE |
| Central African Republic | CAF | Lesotho | LSO | Somalia | SOM |
| Chad | TCD | Liberia | LBR | South Africa | ZAF |
| Comoros | COM | Madagascar | MDG | South Sudan | SSD |
| Dem. Rep. Congo | COD | Malawi | MWI | Sudan | SDN |
| Congo | COG | Mali | MLI | Tanzania | TZA |
| Côte d'Ivoire | CIV | Mauritania | MRT | Togo | TGO |
| Equatorial Guinea | GNQ | Mauritius | MUS | Uganda | UGA |
| Eritrea | ERI | Mozambique | MOZ | Zambia | ZMB |
| Eswatini | SWZ | Namibia | NAM | Zimbabwe | ZWE |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

De otro lado, está la clasificación por ingreso per cápita, Income, que basa su estimación en el Ingreso Nacional Bruto, GNI por sus siglas en inglés. Si es el interés, el Banco Mundial desarrolla ampliamente el tema de esta clasificación en su

link https://datahelpdesk.worldbank.org/knowledgebase/articles/378831-why-use-gni-per-capita-to-classify-economies-into. La clasificación de los 217 países por ingresos en WDI es como sigue.

Tabla *7 Economías de bajos ingresos per cápita.*

| LOW-INCOME ECONOMIES (GNI per cápita $1,085 OR LESS) | | | | | 27 |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Afghanistan | AFG | Guinea | GIN | Rwanda | RWA |
| Burkina Faso | BFA | Guinea-Bissau | GNB | Sierra Leone | SLE |
| Burundi | BDI | Dem. People's Rep. Korea | PRK | Somalia | SOM |
| Central African Republic | CAF | Liberia | LBR | South Sudan | SSD |
| Chad | TCD | Madagascar | MDG | Sudan | SDN |
| Dem. Rep. Congo | COD | Malawi | MWI | Syrian Arab Republic | SYR |
| Eritrea | ERI | Mali | MLI | Togo | TGO |
| Ethiopia | ETH | Mozambique | MOZ | Uganda | UGA |
| The Gambia | GMB | Niger | NER | Yemen | YEM |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Zambia cambio recientemente de clasificación, pero en la WDI descargada de la web del Banco Mundial aun no está registrada como Low-Income, sino que sigue como Lower-Middle Income.

Tabla 8 *Economías de ingresos medio-bajos per cápita.*

| LOWER-MIDDLE INCOME ECONOMIES (GNI per cápita $1,086 TO $4,255) | | | | | 56 |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Algeria | DZA | Honduras | HND | Philippines | PHL |
| Angola | AGO | India | IND | Samoa | WSM |
| Bangladesh | BGD | Indonesia | IDN | São Tomé and Principe | STP |
| Belize | BLZ | Iran | IRN | Senegal | SEN |
| Benin | BEN | Kenya | KEN | Solomon Islands | SLB |
| Bhutan | BTN | Kiribati | KIR | Sri Lanka | LKA |
| Bolivia | BOL | Kyrgyz Republic | KGZ | Tajikistan | TJK |
| Cabo Verde | CPV | Lao PDR | LAO | Tanzania | TZA |
| Cambodia | KHM | Lesotho | LSO | Timor-Leste | TLS |
| Cameroon | CMR | Mauritania | MRT | Tunisia | TUN |
| Comoros | COM | Micronesia | FSM | Ukraine | UKR |
| Congo | COG | Mongolia | MNG | Uzbekistan | UZB |
| Côte d'Ivoire | CIV | Morocco | MAR | Vanuatu | VUT |
| Djibouti | DJI | Myanmar | MMR | Vietnam | VNM |
| Egypt | EGY | Nepal | NPL | West Bank and Gaza | PSE |
| El Salvador | SLV | Nicaragua | NIC | **Zambia** | **ZMB** |
| Eswatini | SWZ | Nigeria | NGA | Zimbabwe | ZWE |

| LOWER-MIDDLE INCOME ECONOMIES (GNI per cápita $1,086 TO $4,255) | | | | | 56 |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Ghana | GHA | Pakistan | PAK | **Venezuela** | **VEN** |
| Haiti | HTI | Papua New Guinea | PNG | | |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Venezuela tiene su ultimo registro de GNI en 2011 (https://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD?locations=VE), por esto se adiciona aquí a criterio particular en este trabajo como Lower-middle income, aunque para efectos del estudio no es relevante dado que no se hace análisis de grupos o clusters, que valdría hacerlo como complemento.

Tabla 9 *Economías de ingresos medio-altos per cápita.*

| UPPER-MIDDLE-INCOME ECONOMIES (GNI per cápita $4,256 TO $13,205) | | | | | 55 |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Albania | ALB | Fiji | FJI | Moldova | MDA |
| American Samoa | ASM | Gabon | GAB | Montenegro | MNE |
| Argentina | ARG | Georgia | GEO | Namibia | NAM |
| Armenia | ARM | Grenada | GRD | North Macedonia | MKD |
| Azerbaijan | AZE | Guatemala | GTM | **Panama** | **PAN** |
| Belarus | BLR | Guyana | GUY | Paraguay | PRY |
| Bosnia and Herzegovina | BIH | Iraq | IRQ | Peru | PER |
| Botswana | BWA | Jamaica | JAM | **Romania** | **ROU** |
| Brazil | BRA | Jordan | JOR | Russia | RUS |
| Bulgaria | BGR | Kazakhstan | KAZ | Serbia | SRB |
| China | CHN | Kosovo | XKX | South Africa | ZAF |
| Colombia | COL | **Lebanon** | **LBN** | St. Lucia | LCA |
| Costa Rica | CRI | Libya | LBY | St. Vincent and the Grenadines | VCT |
| Cuba | CUB | Malaysia | MYS | Suriname | SUR |
| Dominica | DMA | Maldives | MDV | Thailand | THA |
| Dominican Republic | DOM | Marshall Islands | MHL | Tonga | TON |
| Ecuador | ECU | Mauritius | MUS | Türkiye | TUR |
| Equatorial Guinea | GNQ | Mexico | MEX | Turkmenistan | TKM |
| | | | | Tuvalu | TUV |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Belize (Lower-middle income) y Palau (High income) recientemente cambiaron de clasificación, pero en la base descargada de WDI no están

registradas como Upper-middle income. Se agrega que Lebanon (antes Lower-middle income) recientemente cambio de clasificación y aquí se refleja en esta clase en la base WDI descargada.

Tabla 10 *Economías de ingresos altos per cápita.*

| HIGH-INCOME ECONOMIES (GNI per cápita $13,205 OR MORE) | | | | | 79 |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Andorra | AND | Germany | DEU | Northern Mariana Islands | MNP |
| Antigua and Barbuda | ATG | Gibraltar | GIB | Norway | NOR |
| Aruba | ABW | Greece | GRC | Oman | OMN |
| Australia | AUS | Greenland | GRL | **Palau** | PLW |
| Austria | AUT | Guam | GUM | Poland | POL |
| The Bahamas | BHS | Hong Kong SAR, China | HKG | Portugal | PRT |
| Bahrain | BHR | Hungary | HUN | Puerto Rico | PRI |
| Barbados | BRB | Iceland | ISL | Qatar | QAT |
| Belgium | BEL | Ireland | IRL | San Marino | SMR |
| Bermuda | BMU | Isle of Man | IMN | Saudi Arabia | SAU |
| British Virgin Islands | VGB | Israel | ISR | Seychelles | SYC |
| Brunei | BRN | Italy | ITA | Singapore | SGP |
| Canada | CAN | Japan | JPN | Sint Maarten (Dutch part) | SXM |
| Cayman Islands | CYM | Korea | KOR | Slovak Republic | SVK |
| Channel Islands | CHI | Kuwait | KWT | Slovenia | SVN |
| Chile | CHL | Latvia | LVA | Spain | ESP |
| Croatia | HRV | Liechtenstein | LIE | St. Kitts and Nevis | KNA |
| Curaçao | CUW | Lithuania | LTU | St. Martin (French part) | MAF |
| Cyprus | CYP | Luxembourg | LUX | Sweden | SWE |
| Czech Republic | CZE | Macao SAR, China | MAC | Switzerland | CHE |
| Denmark | DNK | Malta | MLT | Trinidad and Tobago | TTO |
| Estonia | EST | Monaco | MCO | Turks and Caicos Islands | TCA |
| Faroe Islands | FRO | Nauru | NRU | United Arab Emirates | ARE |
| Finland | FIN | Netherlands | NLD | United Kingdom | GBR |
| France | FRA | New Caledonia | NCL | United States | USA |
| French Polynesia | PYF | New Zealand | NZL | Uruguay | URY |
| | | | | Virgin Islands | VIR |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

Panamá y Romania, ambas Upper-middle income, recientemente cambiaron de clasificación, pero en la base descargada de WDI no están registradas como High income.

Para el caso de los préstamos operativos, la clasificación se base en políticas operacionales del Banco Mundial. Es así como la categoría IDA, International

Development Association, son para países con bajo ingreso per capita y que carecen de la capacidad financiera para obtener préstamos del IBRD, International Bank for Reconstruction and Development. Los países BLEND o combinados son aquellos elegibles para préstamos tanto de IDA como también para IBRD porque son financieramente solventes. La clasificación por préstamos operativos en WDI es la siguiente, pero hay que tener presente que no son los 217 países en total susceptibles de estos préstamos, tan solo 144.

Tabla 11 *Países clasificados para préstamos IDA.*

| IDA | | 59 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Afghanistan | AFG | Haiti | HTI | Rwanda | RWA |
| Bangladesh | BGD | Honduras | HND | Samoa | WSM |
| Benin | BEN | Kiribati | KIR | São Tomé and Principe | STP |
| Bhutan | BTN | Kosovo | XKX | Senegal | SEN |
| Burkina Faso | BFA | Kyrgyz Republic | KGZ | Sierra Leone | SLE |
| Burundi | BDI | Lao PDR | LAO | Solomon Islands | SLB |
| Cambodia | KHM | Lesotho | LSO | Somalia | SOM |
| Central African Republic | CAF | Liberia | LBR | South Sudan | SSD |
| Chad | TCD | Madagascar | MDG | Sudan | SDN |
| Comoros | COM | Malawi | MWI | Syrian Arab Republic | SYR |
| Dem. Rep. Congo | COD | Maldives | MDV | Tajikistan | TJK |
| Côte d'Ivoire | CIV | Mali | MLI | Tanzania | TZA |
| Djibouti | DJI | Marshall Islands | MHL | Togo | TGO |
| Eritrea | ERI | Mauritania | MRT | Tonga | TON |
| Ethiopia | ETH | Micronesia | FSM | Tuvalu | TUV |
| The Gambia | GMB | Mozambique | MOZ | Uganda | UGA |
| Ghana | GHA | Myanmar | MMR | Vanuatu | VUT |
| Guinea | GIN | Nepal | NPL | Yemen | YEM |
| Guinea-Bissau | GNB | Nicaragua | NIC | Zambia | ZMB |
| Guyana | GUY | Niger | NER | | |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

Tabla 12 *Países clasificados para préstamos BLEND.*

| BLEND | | 15 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Cabo Verde | CPV | Grenada | GRD | St. Lucia | LCA |
| Cameroon | CMR | Kenya | KEN | St. Vincent and the Grenadines | VCT |
| Congo | COG | Nigeria | NGA | Timor-Leste | TLS |
| Dominica | DMA | Pakistan | PAK | Uzbekistan | UZB |
| Fiji | FJI | Papua New Guinea | PNG | Zimbabwe | ZWE |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

Tabla 13 *Países clasificados para préstamos IBRD.*

| IBRD | | 70 | | | |
|---|---|---|---|---|---|
| Country | Code | Country | Code | Country | Code |
| Albania | ALB | Equatorial Guinea | GNQ | North Macedonia | MKD |
| Algeria | DZA | Eswatini | SWZ | Palau | PLW |
| Angola | AGO | Gabon | GAB | Panama | PAN |
| Antigua and Barbuda | ATG | Georgia | GEO | Paraguay | PRY |
| Argentina | ARG | Guatemala | GTM | Peru | PER |
| Armenia | ARM | India | IND | Philippines | PHL |
| Azerbaijan | AZE | Indonesia | IDN | Poland | POL |
| Belarus | BLR | Iran | IRN | Romania | ROU |
| Belize | BLZ | Iraq | IRQ | Russia | RUS |
| Bolivia | BOL | Jamaica | JAM | Serbia | SRB |
| Bosnia and Herzegovina | BIH | Jordan | JOR | Seychelles | SYC |
| Botswana | BWA | Kazakhstan | KAZ | South Africa | ZAF |
| Brazil | BRA | Lebanon | LBN | Sri Lanka | LKA |
| Bulgaria | BGR | Libya | LBY | St. Kitts and Nevis | KNA |
| Chile | CHL | Malaysia | MYS | Suriname | SUR |
| China | CHN | Mauritius | MUS | Thailand | THA |
| Colombia | COL | Mexico | MEX | Trinidad and Tobago | TTO |
| Costa Rica | CRI | Moldova | MDA | Tunisia | TUN |
| Croatia | HRV | Mongolia | MNG | Turkey | TUR |
| Dominican Republic | DOM | Montenegro | MNE | Turkmenistan | TKM |
| Ecuador | ECU | Morocco | MAR | Ukraine | UKR |
| Egypt | EGY | Namibia | NAM | Uruguay | URY |
| El Salvador | SLV | Nauru | NRU | Venezuela | VEN |
| | | | | Vietnam | VNM |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

WDI tiene, además de los 217 países, otros 48 individuos que son economías que agrupan a varios países, 14 de ellos ya son individuos que representan la clasificación principal dada anteriormente, por Region (7), por Income (4) y por Lending (3), Y adicionalmente se tienen 34 individuos o filas en WDI representan agrupaciones adicionales de Region, Income, Lending asi como combinación de estas, o relacionados a aspectos como demografía, nivel de desarrollo, fragilidad, etc.

Tabla 14 *Economías adicionales (14) que se listan en WDI con sus indicadores, y que son resultado del agrupamiento básico según clasificación de los países.*

| BY REGION (7) | Code |
|---|---|
| East Asia & Pacific | EAS |
| Europe & Central Asia | ECS |
| Latin America & Caribbean | LCN |
| Middle East & North Africa | MEA |
| North America | NAC |
| South Asia | SAS |
| Sub-Saharan Africa | SSF |

| BY INCOME (4) | Code |
|---|---|
| Low income | LIC |
| Lower middle income | LMC |
| Upper middle income | UMC |
| High income | HIC |

| BY LENDING (3) | Code |
|---|---|
| IDA only | IDX |
| IBRD only | IBD |
| IDA blend | IDB |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

Tabla 15 *Economías adicionales (34) que se listan en WDI con sus indicadores, y que son resultado de otros agrupamientos de países más específicos.*

| BY REGION (+7) | Code |
|---|---|
| Africa Eastern and Southern | AFE |
| Africa Western and Central | AFW |
| Central Europe and the Baltics | CEB |
| European Union | EUU |
| Euro area | EMU |
| Arab World | ARB |
| World | WLD |

| SMALL STATES (+4) | Code |
|---|---|
| Caribbean small states | CSS |
| Pacific island small states | PSS |
| Small states | SST |
| Other small states | OSS |

| BY LEVEL OF DEVELOPMENT (+2) | Code |
|---|---|
| Least developed countries: UN classification | LDC |
| OECD members | OED |

| BY INCOME (+2) | Code |
|---|---|
| Low & middle income | LMY |
| Middle income | MIC |

| BY REGION & INCOME (+5) | Code |
|---|---|
| East Asia & Pacific (excluding high income) | EAP |
| Europe & Central Asia (excluding high income) | ECA |
| Latin America & Caribbean (excluding high income) | LAC |
| Middle East & North Africa (excluding high income) | MNA |
| Sub-Saharan Africa (excluding high income) | SSA |

| BY DEMOGRAPHIC DIVIDEND (+4) | Code |
|---|---|
| Pre-demographic dividend | PRE |
| Early-demographic dividend | EAR |
| Late-demographic dividend | LTE |
| Post-demographic dividend | PST |

| BY LENDING (+3) | Code |
|---|---|
| IDA & IBRD total | IBT |
| IDA total | IDA |
| Heavily indebted poor countries (HIPC) | HPC |

| BY REGION & IBRD-IDA (+6) | Code |
|---|---|
| East Asia & Pacific (IDA & IBRD) | TEA |
| Europe & Central Asia (IDA & IBRD) | TEC |
| Latin America & Caribbean (IDA & IBRD) | TLA |
| Middle East & North Africa (IDA & IBRD) | TMN |
| South Asia (IDA & IBRD) | TSA |
| Sub-Saharan Africa (IDA & IBRD) | TSS |

| OTHERS (+1) | Code |
|---|---|
| Fragile and conflict affected situations | FCS |

*Nota*: Adaptado de *World Bank Country and Lending Groups*, por World Bank, Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

Se puede consultar mayor información de estas economías agrupadas en detalle de requerirse ( https://data.worldbank.org/country), que países lo conforman,

los valores de algunos indicadores y gráficos del mismo. Incluye WDI para 2022 un individuo más que no agrupa países realmente, y que es nombrado como 'Not Classified', con código de país o economía 'INX', pero que no tiene valores en los indicadores, luego no es un individuo que aporte al cubo de datos, con esto para 2022 WDI tiene 265 economías como individuos, no 266. De otro lado, hecha la observación de las clasificaciones se podría crear una variable categórica a partir del nivel de desarrollo de cada país, estableciendo tres posibles: OECD, Menos Desarrollado, y Otros; y también se podría crear categorías de acuerdo a alguno de los indicadores del Tópico Poverty; esto sugiere un análisis futuro de clasificación a partir de los indicadores, que en este trabajo no son objeto de revisión.

Interesa también tener a la vista los países menos desarrollados, con código de economía LCD en WDI, y que según la clasificación de la ONU está conformado por los siguientes.

Tabla 16 *Países menos desarrollados, LCD, clasificación ONU*

| Country | Code |
|---|---|
| Afghanistan | AFG |
| Angola | AGO |
| Bangladesh | BGD |
| Benin | BEN |
| Bhutan | BTN |
| Burkina Faso | BFA |
| Burundi | BDI |
| Cambodia | KHM |
| Central African Republic | CAF |
| Chad | TCD |
| Comoros | COM |
| Dem. Rep. Congo | COD |
| Djibouti | DJI |
| Eritrea | ERI |
| Ethiopia | ETH |

| Country | Code |
|---|---|
| The Gambia | GMB |
| Guinea | GIN |
| Guinea-Bissau | GNB |
| Haiti | HTI |
| Kiribati | KIR |
| Lao PDR | LAO |
| Lesotho | LSO |
| Liberia | LBR |
| Madagascar | MDG |
| Malawi | MWI |
| Mali | MLI |
| Mauritania | MRT |
| Mozambique | MOZ |
| Myanmar | MMR |
| Nepal | NPL |

| Country | Code |
|---|---|
| Niger | NER |
| Rwanda | RWA |
| São Tomé and Principe | STP |
| Senegal | SEN |
| Sierra Leone | SLE |
| Solomon Islands | SLB |
| Somalia | SOM |
| South Sudan | SSD |
| Sudan | SDN |
| Tanzania | TZA |
| Timor-Leste | TLS |
| Togo | TGO |
| Tuvalu | TUV |
| Uganda | UGA |
| Yemen | YEM |
| Zambia | ZMB |

*Nota*: Adaptado de *Least developed countries: UN classification*, por World Bank, Recuperado 1 de Mayo de 2022 de https://data.worldbank.org/region/least-developed-countries-un-classification.

Para este trabajo, es de interés estudiar tanto los países que forman Latinoamérica y el Caribe (42), sumado a otras economías agrupadas (7): World (WLD), European Union (EUU), Latin America & Caribbean (LCN), Middle East & North Africa (MEA), OECD members (OED), Least developed countries: UN classification (LDC) y North America (NAC)

## Tópicos o grupos de Indicadores.

Hay 11 Tópicos o ejes temáticos generales en que se clasifican los indicadores del desarrollo mundial WDI:

Tabla 17 *Topicos generales de clasificación de los indicadores WDI*

| Tópicos Generales | Cantidad Indicadores | %Cantidad | %Acumulado Cantidad |
|---|---|---|---|
| **Economic Policy & Debt** | **346** | **0.23944637** | **0.2394464** |
| **Health** | **249** | **0.17231834** | **0.4117647** |
| **Private Sector & Trade** | **167** | **0.11557093** | **0.5273356** |
| Social Protection & Labor | 158 | 0.10934256 | 0.6366782 |
| Education | 147 | 0.10173010 | 0.7384083 |
| Environment | 140 | 0.09688581 | 0.8352941 |
| Public Sector | 101 | 0.06989619 | 0.9051903 |
| Financial Sector | 55 | 0.03806228 | 0.9432526 |
| Infrastructure | 39 | 0.02698962 | 0.9702422 |
| **Poverty** | **28** | 0.01937716 | 0.9896194 |
| Gender | 15 | 0.01038062 | 1.0000000 |

El tópico general que más agrupa indicadores en WDI es Economic Policy & Debt, pero sumado a los tópicos de Health, y de Private Sector & Trade, representan el 52.7% del total de indicadores en WDI. Llama la atención que los indicadores de Poverty como tópico general son tan solo 28, menos del 2% del total de indicadores. A continuación una ampliación más en detalle pero de cada tópico

especifico y la cantidad de indicadores asociados, hay varios que superan los 30 indicadores.

Tabla 18 *Cantidad de indicadores por tópicos específicos.*

| Tópicos | Cantidad de Indicadores |
|---|---|
| Economic Policy & Debt: Balance of payments: Capital & financial account | 11 |
| Economic Policy & Debt: Balance of payments: Current account: Balances | 4 |
| Economic Policy & Debt: Balance of payments: Current account: Goods, services & income | 22 |
| Economic Policy & Debt: Balance of payments: Current account: Transfers | 7 |
| Economic Policy & Debt: Balance of payments: Reserves & other items | 6 |
| Economic Policy & Debt: External debt: Debt outstanding | 10 |
| Economic Policy & Debt: External debt: Debt ratios & other items | 11 |
| Economic Policy & Debt: External debt: Debt service | 5 |
| Economic Policy & Debt: External debt: Net flows | 20 |
| Economic Policy & Debt: National accounts: Adjusted savings & income | 28 |
| Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita | 2 |
| Economic Policy & Debt: National accounts: Growth rates | 16 |
| Economic Policy & Debt: National accounts: Local currency at constant prices: Aggregate indicators | 8 |
| Economic Policy & Debt: National accounts: Local currency at constant prices: Expenditure on GDP | 11 |
| Economic Policy & Debt: National accounts: Local currency at constant prices: Other items | 3 |
| Economic Policy & Debt: National accounts: Local currency at constant prices: Value added | 5 |
| Economic Policy & Debt: National accounts: Local currency at current prices: Aggregate indicators | 11 |
| Economic Policy & Debt: National accounts: Local currency at current prices: Expenditure on GDP | 14 |
| Economic Policy & Debt: National accounts: Local currency at current prices: Value added | 5 |
| Economic Policy & Debt: National accounts: Shares of GDP & other | 25 |
| Economic Policy & Debt: National accounts: US$ at constant 2015 prices: Aggregate indicators | 5 |
| Economic Policy & Debt: National accounts: US$ at constant 2015 prices: Expenditure on GDP | 9 |
| Economic Policy & Debt: National accounts: US$ at constant 2015 prices: Value added | 7 |
| Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators | 9 |
| Economic Policy & Debt: National accounts: US$ at current prices: Expenditure on GDP | 10 |
| Economic Policy & Debt: National accounts: US$ at current prices: Value added | 4 |
| **Economic Policy & Debt: Official development assistance** | **65** |
| Economic Policy & Debt: Purchasing power parity | 13 |
| Education: Efficiency | 21 |
| **Education: Inputs** | **42** |
| **Education: Outcomes** | **40** |
| **Education: Participation** | **44** |
| Environment: Agricultural production | 13 |
| Environment: Biodiversity & protected areas | 7 |
| Environment: Density & urbanization | 12 |
| **Environment: Emissions** | **42** |
| Environment: Energy production & use | 27 |
| Environment: Freshwater | 9 |
| Environment: Land use | 24 |
| Environment: Natural resources contribution to GDP | 6 |
| Financial Sector: Access | 15 |
| Financial Sector: Assets | 14 |
| Financial Sector: Capital markets | 7 |

| Tópicos | Cantidad de Indicadores |
|---|---|
| Financial Sector: Exchange rates & prices | 10 |
| Financial Sector: Interest rates | 5 |
| Financial Sector: Monetary holdings (liabilities) | 4 |
| Gender: Agency | 2 |
| Gender: Health | 7 |
| Gender: Participation & access | 2 |
| Gender: Public life & decision making | 4 |
| Health | 2 |
| Health: Disease prevention | 27 |
| Health: Health systems | 23 |
| **Health: Mortality** | **39** |
| Health: Nutrition | 26 |
| Health: Population: Dynamics | 13 |
| **Health: Population: Structure** | **58** |
| Health: Reproductive health | 16 |
| Health: Risk factors | 32 |
| Health: Universal Health Coverage | 13 |
| Infrastructure: Communications | 13 |
| Infrastructure: Technology | 16 |
| Infrastructure: Transportation | 10 |
| Poverty: Income distribution | 9 |
| Poverty: Multidimensional poverty | 8 |
| Poverty: Poverty rates | 7 |
| Poverty: Shared prosperity | 4 |
| **Private Sector & Trade: Business environment** | **54** |
| Private Sector & Trade: Exports | 23 |
| Private Sector & Trade: Imports | 23 |
| Private Sector & Trade: Private infrastructure investment | 8 |
| Private Sector & Trade: Tariffs | 24 |
| Private Sector & Trade: Total merchandise trade | 1 |
| Private Sector & Trade: Trade facilitation | 17 |
| Private Sector & Trade: Trade indexes | 5 |
| Private Sector & Trade: Trade price indices | 2 |
| Private Sector & Trade: Travel & tourism | 10 |
| Public Sector: Conflict & fragility | 7 |
| Public Sector: Defense & arms trade | 8 |
| Public Sector: Government finance | 1 |
| Public Sector: Government finance: Deficit & financing | 10 |
| Public Sector: Government finance: Expense | 13 |
| Public Sector: Government finance: Revenue | 22 |
| **Public Sector: Policy & institutions** | **40** |
| **Social Protection & Labor: Economic activity** | **74** |
| Social Protection & Labor: Labor force structure | 28 |
| Social Protection & Labor: Migration | 5 |
| Social Protection & Labor: Performance | 27 |
| Social Protection & Labor: Unemployment | 24 |

## Indicadores

En total son 1445 indicadores en WDI, estos se pueden ver individualmente para cada uno de los 217 países o para cada uno de los 48 agrupamientos de economías, en https://data.worldbank.org/indicator. Estos indicadores tienen varias periodicidades: Anual, bienal y Trimestral (representado como Anual), y varios métodos de agregación: Gap-filled total, estimaciones del modelo lineal de efectos mixtos, medianas, sumas, promedios no ponderados y otros ponderados. Las fuentes de WDI son diversas organizaciones, entre otras: Instancias de las Naciones Unidas, oficinas estadísticas nacionales de los países, Universidades, Centros de investigación (por ejemplo Max Planck, o el JCR de Holanda, etc), Laboratorios, e inclusive de estimaciones que hace el staff del Banco Mundial. Particularmente, a continuación se listan los indicadores más relevantes que el Banco Mundial expone para LCD en su web, de los cuales veremos más adelante hecho el análisis multivariante que coinciden unos.

Tabla 19 *Indicadores que el Banco Mundial relaciona en la clasificación LDC para los objetivos del Desarrollo Sostenible SDG*

| Sustainable Development Goals (SDG) | New Indicator Name | Indicator | Most recent value | Year |
|---|---|---|---|---|
| 1: No Poverty | Po1070 | Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) | No data available | - |
| | Po1073 | Poverty headcount ratio at national poverty lines (% of population) | No data available | - |
| | Po800 | Multidimensional poverty headcount ratio (% of total population) | No data available | - |
| 2: Zero Hunger | He1108 | Prevalence of stunting, height for age (% of children under 5) | No data available | - |
| | He1112 | Prevalence of undernourishment (% of population) | 21 | 2019 |
| | He1104 | Prevalence of severe food insecurity in the population (%) | No data available | - |
| | He1110 | Prevalence of stunting, height for age, female (% of children under 5) | No data available | - |
| | He1111 | Prevalence of stunting, height for age, male (% of children under 5) | No data available | - |
| 3: Good Health | He737 | Maternal mortality ratio (modeled estimate, per 100,000 | 415 | 2017 |

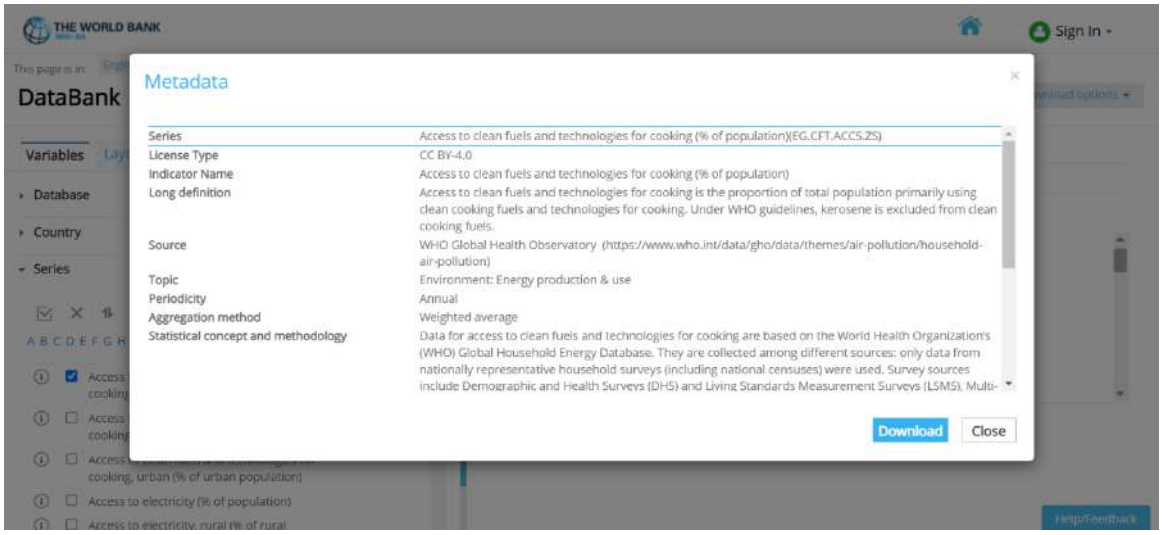
| Sustainable Development Goals (SDG) | New Indicator Name | Indicator | Most recent value | Year |
|---|---|---|---|---|
| and Well-Being | | live births) | | |
| | He797 | Mortality rate, under-5 (per 1,000 live births) | 61 | 2020 |
| | He603 | Incidence of HIV, ages 15-49 (per 1,000 uninfected population ages 15-49) | 0.9 | 2020 |
| | He798 | Mortality rate, under-5, female (per 1,000 live births) | 56 | 2020 |
| | He799 | Mortality rate, under-5, male (per 1,000 live births) | 65 | 2020 |
| | He51 | Adolescent fertility rate (births per 1,000 women ages 15-19) | 90 | 2020 |
| 4: Quality Education | Ed1205 | School enrollment, primary and secondary (gross), gender parity index (GPI) | 0.95 | 2020 |
| | Ed722 | Lower secondary completion rate, total (% of relevant age group) | 47.3 | 2019 |
| | Ed709 | Literacy rate, youth total (% of people ages 15-24) | 79 | 2020 |
| | Ed720 | Lower secondary completion rate, female (% of relevant age group) | 45.9 | 2019 |
| | Ed721 | Lower secondary completion rate, male (% of relevant age group) | 48.6 | 2019 |
| 5: Gender Equality | Ge1148 | Proportion of seats held by women in national parliaments (%) | 25 | 2021 |
| | Ge1433 | Women Business and the Law Index Score (scale 1-100) | No data available | - |
| | So420 | Female share of employment in senior and middle management (%) | No data available | - |
| | Pr444 | Firms with female top manager (% of firms) | 17 | 2020 |
| 6: Clean Water and Sanitation | He962 | People using at least basic drinking water services (% of population) | 67 | 2020 |
| | He971 | People using safely managed sanitation services (% of population) | 26 | 2020 |
| | He965 | People using at least basic sanitation services (% of population) | 37 | 2020 |
| | En695 | Level of water stress: freshwater withdrawal as a proportion of available freshwater resources | No data available | - |
| | He968 | People using safely managed drinking water services (% of population) | No data available | - |
| 7: Affordable and Clean Energy | En1177 | Renewable energy consumption (% of total final energy consumption) | 72.95 | 2015 |
| | En4 | Access to electricity (% of population) | 54.7 | 2020 |
| | En1 | Access to clean fuels and technologies for cooking (% of population) | 17 | 2020 |
| | En5 | Access to electricity, rural (% of rural population) | 44.1 | 2020 |
| | En6 | Access to electricity, urban (% of urban population) | 78.4 | 2020 |
| 8: Decent Work and Economic Growth | So479 | GDP per person employed (constant 2017 PPP $) | 8,278 | 2021 |
| | Ec471 | GDP growth (annual %) | 2.3 | 2021 |
| | Ec476 | GDP per capita growth (annual %) | 0 | 2021 |
| | Fi7 | Account ownership at a financial institution or with a mobile-money-service provider (% of population ages 15+) | No data available | - |
| 9: Industry, Innovation and Infrastructure | En194 | CO2 emissions (metric tons per capita) | 0.3 | 2019 |
| | Ec728 | Manufacturing, value added (% of GDP) | 15 | 2021 |
| | In1183 | Research and development expenditure (% of GDP) | No data available | - |
| 10: Reduced Inequalities | Po87 | Annualized average growth rate in per capita real survey mean consumption or income, bottom 40% of population (%) | No data available | - |
| | Po88 | Annualized average growth rate in per capita real survey mean consumption or income, total population (%) | No data available | - |
| | Po1143 | Proportion of people living below 50 percent of median income (%) | No data available | - |
| | Fi104 | Average transaction cost of sending remittances to a specific country (%) | No data available | - |

| Sustainable Development Goals (SDG) | New Indicator Name | Indicator | Most recent value | Year |
|---|---|---|---|---|
| | Fi103 | Average transaction cost of sending remittances from a specific country (%) | No data available | - |
| 11: Sustainable Cities and Communities | En1058 | Population living in slums (% of urban population) | 59 | 2018 |
| | En1418 | Urban population growth (annual %) | 3.9 | 2021 |
| | En993 | PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total) | 100 | 2017 |
| 12: Responsible Consumption and Production | Ec29 | Adjusted net savings, excluding particulate emission damage (% of GNI) | 19.8 | 2020 |
| | En1352 | Total natural resources rents (% of GDP) | 6.7 | 2020 |
| | En321 | Droughts, floods, extreme temperatures (% of population, average 1990-2009) | No data available | - |
| 13: Climate Action | En307 | Disaster risk reduction progress score (1-5 scale; 5=best) | No data available | - |
| | En1351 | Total greenhouse gas emissions (kt of CO2 equivalent) | 1,593,780 | 2019 |
| 14: Life Below Water | En734 | Marine protected areas (% of territorial waters) | No data available | - |
| | En1349 | Total fisheries production (metric tons) | 14,245,744 | 2018 |
| | En91 | Aquaculture production (metric tons) | 4,182,897 | 2018 |
| | En135 | Capture fisheries production (metric tons) | 10,062,847 | 2018 |
| 15: Life On Land | En459 | Forest area (% of land area) | 26.3 | 2020 |
| | En1324 | Terrestrial and marine protected areas (% of total territorial area) | 12.8 | 2018 |
| | En1325 | Terrestrial protected areas (% of total land area) | 14.1 | 2018 |
| 16: Peace, Justice and Strong Institutions | Pu1128 | Primary government expenditures as a proportion of original approved budget (%) | No data available | - |
| | Pr129 | Bribery incidence (% of firms experiencing at least one bribe payment request) | 28 | 2020 |
| | He218 | Completeness of birth registration (%) | 44 | 2017 |
| | He221 | Completeness of birth registration, rural (%) | No data available | - |
| | He222 | Completeness of birth registration, urban (%) | No data available | - |
| 17: Partnerships For the Goals | In618 | Individuals using the Internet (% of population) | 24 | 2020 |
| | Ec986 | Personal remittances, received (% of GDP) | 4.6 | 2020 |
| | Ec455 | Foreign direct investment, net inflows (% of GDP) | 1.9 | 2020 |
| | Ec296 | Debt service (PPG and IMF only, % of exports of goods, services and primary income) | 10.2 | 2020 |
| | Pu1305 | Tax revenue (% of GDP) | 10.2 | 2018 |
| | Ec397 | Exports of goods and services (% of GDP) | 19.5 | 2021 |

*Nota*: Adaptado de *Least developed countries: UN classification by SDG Goal*, por World Bank, Recuperado 20 de Agosto de 2022 de https://data.worldbank.org/region/least-developed-countries-un-classification.

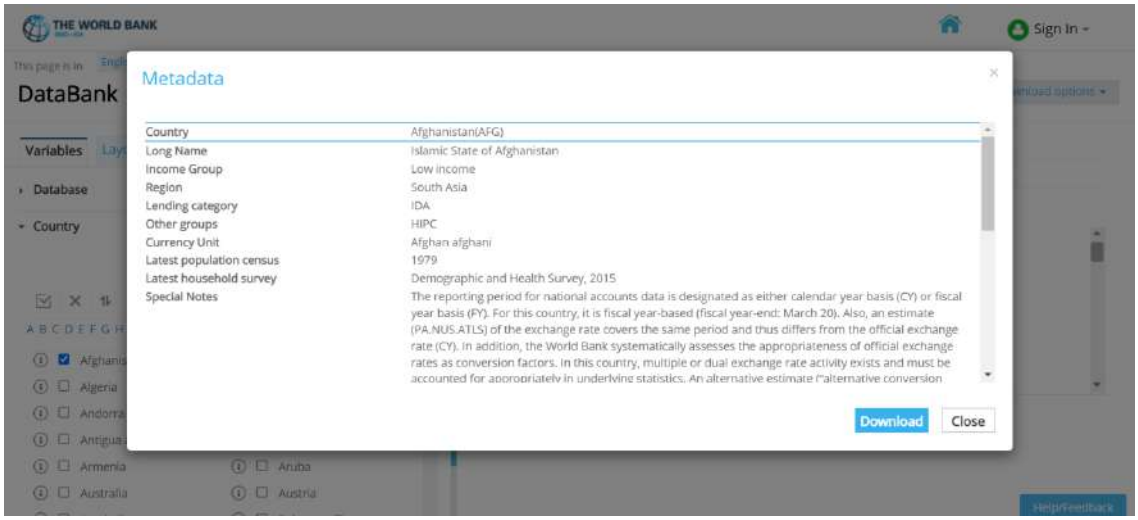
La información completa de cada indicador se puede obtener al descargar WDI en formato Excel, y se tienen las hojas para las series (indicadores) con las explicaciones en detalle. También desde los metadatos de WDI (https://databank.worldbank.org/reports.aspx?source=2&type=metadata) seleccionando un país, un año y el indicador de interés, y seleccionar el botón de Metadata, y se desplegara toda la información del indicador, también se puede ver la del país o economía agrupada.

Figura 13 *Metadata del indicador Access to clean fuels and technologies for cooking (% of population).*




*Nota*. Tomado de *DataBank: World Development Indicators, Metadata Series*. Recuperado 20 Agosto de 2022 de (https://databank.worldbank.org/reports.aspx?source=2&type=metadata).

Figura 14 *Metadata de la economia o pais Afghanistan (AFG)*




*Nota*. Tomado de *DataBank: World Development Indicators, Metadata Country*. Recuperado 20 Agosto de 2022 de (https://databank.worldbank.org/reports.aspx?source=2&type=metadata).

Para 2022 WDI compila la información de los años 1960 a 2021, para un total de 62. Muchos de estos indicadores no se tienen para cada país o para cada año, por lo que hay una gran mayoría de datos faltantes o perdidos, NAs.

Tabla 20 *Valoración de cantidad de NAs por Tópicos generales*

| Tópicos Generales | %NAs | %Cantidad Indicadores | Valoracion combinada (1-%NAs)*%Qty |
|---|---|---|---|
| **Economic Policy & Debt** | 0.7049507 | **0.23944637** | **0.070648484** |
| Education | **0.6642371** | **0.10173010** | **0.034157193** |
| Environment | 0.7589654 | 0.09688581 | 0.023352832 |
| Financial Sector | 0.7761747 | 0.03806228 | 0.008519301 |
| Gender | **0.6266068** | 0.01038062 | 0.003876053 |
| **Health** | **0.590369** | **0.17231834** | **0.070586934** |
| Infrastructure | **0.6184646** | 0.02698962 | 0.010297495 |
| Poverty | 0.8472528 | 0.01937716 | 0.002959807 |
| **Private Sector & Trade** | **0.665924** | **0.11557093** | **0.038609474** |
| Public Sector | **0.6463356** | 0.06989619 | 0.024719794 |
| Social Protection & Labor | **0.6086668** | **0.10934256** | **0.042789374** |

El tópico general con menos NAs es el de Health 59%, seguido de Social Protection & Labour 60.8%, Infrastructure con 61.84%, Gender 62.6%, Public Sector 64.63%, Education 66.4% y Private Sector & Trade 66.6%, el resto >70%. Un valoración combinada de menos NAs y mayor numero de indicadores, nos deja que para 2022 WDI los tópicos generales de interés son Economic Policy & Debt, Health y Private Sector & Trade, estos tres representan el 52.7% del total de indicadores en WDI.

El proceso de exclusión de variables explicativas o indicadores determina en gran medida los resultados, dado que depende en primera instancia de lo que se quiera analizar. Para WDI es qué ocasiones o años (k-tabla), que tópicos (grupos de indicadores) y que indicadores (variables) dentro de ellos, son los de interés para el estudio; no es lo mismo aplicar técnicas a los indicadores de Education solos, o compararlos abiertamente con los de Environment en un mismo estudio, o porque no comparar Education, Environment, Health, y Poverty al tiempo, pero en esos escenarios tan densos ya la cantidad es bastante representativa y compleja de manejar desde el punto de vista de cálculo computacional. De hecho, la selección de individuos, filas, o sea Economías o países ayuda a optimizar los recursos físicos y lógicos con que contamos para el análisis. Pasar de 265 economías, con 1445

indicadores agrupados en 11 tópicos generales, y para 62 años, a solo una matriz reducida de los tres modos amerita una revisión clara y consistente, y en este trabajo se usa en principio una valoración combinada de menor cantidad de datos perdidos con una mayor cantidad de indicadores por tópico.

## Momentos o años

WDI del 2022, tiene indicadores entre 1960 y 2021, pero hay muchos indicadores por economías que no se tienen, esto hace que tengamos un alto porcentaje de datos perdidos, faltantes o NAs. Del dataset WDI inicial que se descarga (sea CSV o Excel), donde las 266 economías multiplicadas por 1445 forman las filas (384370), y 62 años forman las columnas, se puede hacer un rápido cálculo de los % de NAs a partir del conteo de los mismos.

Tabla 21 *Valoración de cantidad de NAs por años.*

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| 2010 | 165258 | 0.429945105 |
| 2014 | 166164 | 0.432302209 |
| 2012 | 168182 | 0.437552358 |
| **2015** | **169485** | **0.440942321** |
| **2016** | **170887** | **0.444589848** |
| 2011 | 172052 | 0.447620782 |
| 2013 | 173177 | 0.450547649 |
| **2017** | **173265** | **0.450776595** |
| **2018** | **177930** | **0.462913339** |
| 2007 | 178650 | 0.464786534 |
| 2009 | 178788 | 0.465145563 |
| 2008 | 180075 | 0.468493899 |
| 2005 | 183130 | 0.47644197 |
| 2006 | 183301 | 0.476886854 |
| 2004 | 194826 | 0.506870984 |
| **2019** | **197942** | **0.514977756** |
| 2000 | 198827 | 0.517280225 |
| 2002 | 199901 | 0.520074407 |
| 2003 | 200014 | 0.520368395 |

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| 1998 | 231382 | 0.601977261 |
| 1997 | 232682 | 0.605359419 |
| 1995 | 234814 | 0.610906158 |
| 1996 | 234954 | 0.611270391 |
| 1994 | 242658 | 0.631313578 |
| 1993 | 244937 | 0.637242761 |
| 1992 | 245442 | 0.638556599 |
| **2020** | **248347** | **0.646114421** |
| 1991 | 250450 | 0.651585712 |
| 1990 | 257889 | 0.670939459 |
| 1987 | 282911 | 0.736038192 |
| 1989 | 282916 | 0.736051201 |
| 1988 | 284467 | 0.740086375 |
| 1986 | 284548 | 0.74029711 |
| 1985 | 285354 | 0.742394047 |
| 1984 | 286668 | 0.745812628 |
| 1982 | 286922 | 0.74647345 |
| 1983 | 287379 | 0.747662409 |
| 1981 | 288670 | 0.751021151 |

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| 1977 | 296143 | 0.77046 |
| 1978 | 296608 | 0.77167 |
| 1976 | 300076 | 0.7807 |
| 1975 | 302393 | 0.78672 |
| 1974 | 306088 | 0.79634 |
| 1972 | 307068 | 0.79889 |
| 1973 | 307579 | 0.80022 |
| 1971 | 310192 | 0.80701 |
| 1970 | 317287 | 0.82547 |
| 1969 | 335530 | 0.87293 |
| 1967 | 335897 | 0.87389 |
| 1968 | 336323 | 0.875 |
| 1966 | 337665 | 0.87849 |
| 1965 | 337692 | 0.87856 |
| 1964 | 339874 | 0.88424 |
| 1962 | 340232 | 0.88517 |
| 1963 | 340392 | 0.88558 |
| 1961 | 342475 | 0.891 |
| 1960 | 347409 | 0.90384 |

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| 2001 | 205301 | 0.534123371 |
| 1999 | 225878 | 0.587657726 |

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| 1980 | 290989 | 0.757054401 |
| 1979 | 295736 | 0.76940448 |

| Year | Qty.NAs | %Nas.per.Col (384370 filas tot) |
|---|---|---|
| **2021** | **374344** | **0.97392** |

Entre los años 2000 y 2019 se tienen porcentajes de NAs entre el 42.99% y el 53.4%, el resto los supera. Los años 2020 y 2021 tienen porcentajes de NAs de 64.61% y 97.39% respectivamente. Claramente el 2021 no nos aporta mucha información. En aras de lograr los años adecuados para aplicar el análisis multivariante con suficiente información, se toma 2015 a 2019 con % menores al 51.4%, y además incluir el 2020.

## Los métodos estadísticos

Estamos asistiendo ya a la era del Big Data, donde trabajamos con grandes cantidades de datos que se producen desde diversas fuentes, involucran además muchas disciplinas, y esto complica de manera contundente las interpretaciones. Esta complejidad en los datos obliga a un salto de la estadística tradicional a la estadística multivariante, donde hay la preocupación por la gran variabilidad sumado a variables íntimamente relacionadas y que lleva a mantener la hipótesis teórica que no es generalmente la que quiere contrastar un investigador, y esto nos exige el uso de técnicas capaces de simplificar la información original y proporcionar significado a los resultados obtenidos. Este escenario dinámico demanda a las escuelas de pensamiento del análisis multivariante, el desarrollo de técnicas apropiadas para enfrentar el reto; es así como el grupo de Salamanca, la escuela Salmantina, desde sus líneas prioritarias como es el Biplot, Sparse, y técnicas relacionadas, aporta nuevos desarrollos en tal sentido.

Entrando en contexto, partimos de múltiples variables en WDI, los indicadores, que no son fáciles para la interpretación final, buscamos una metodología que nos aproxime y nos permita reducir esa cantidad de variables y mejorar la interpretación, y lograr algo con significado.

Figura 15 *Arreglo de Matrices y Esquema de arreglo de tres vías*

Dataset Original WDI

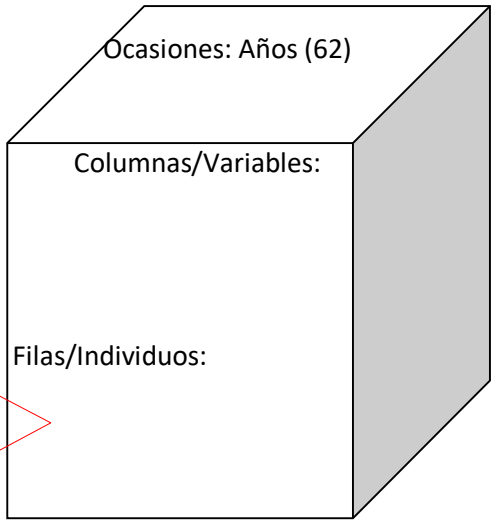
| k-tabla: Economia | Filas: Indicadores | Columnas: años | | | | |
|---|---|---|---|---|---|---|
| | | 1960 | 1961 | ... | 2020 | 2021 |
| AFG | AG.AGR.TRAC.NO | 23.72 | 21.68 | ... | 21.87 | 22.23 |
| AFE | AG.CON.FERT.PT.ZS | 66363.42 | 63422.20 | ... | 75073.49 | 79651.76 |
| AFW | AG.CON.FERT.ZS | 33.19 | 33.37 | ... | 30.04 | 26.90 |
| ... | ... | ... | ... | ... | ... | ... |
| ZWE | VC.IHR.PSRC.P5 | 6.28 | 5.34 | ... | 4.18 | 4.13 |

Arreglo de datos para STATIS (enfasis en los individuos)

| k-tabla: años | Filas: Indicadores | Columnas: Economias | | | | |
|---|---|---|---|---|---|---|
| | | AFG | ... | YEM | ZMB | ZWE |
| 1960 | AG.AGR.TRAC.NO | 2.74 | ... | 2.68 | 3.05 | 3.05 |
| 1961 | AG.CON.FERT.PT.ZS | 2.79 | ... | 2.76 | 3.42 | 3.29 |
| ... | AG.CON.FERT.ZS | 3.24 | ... | 3.24 | 3.26 | 3.29 |
| 2020 | ... | ... | ... | ... | ... | ... |
| 2021 | VC.IHR.PSRC.P5 | 3.18 | ... | 3.26 | 3.19 | 3.22 |

**Arreglo de datos para STATIS-dual (enfasis en las variables)**

| k-tabla: años | Filas: Economias | Columnas: Indicadores | | | | |
|---|---|---|---|---|---|---|
| | | AG.AGR.TRAC.NO | AG.CON.FERT.PT.ZS | AG.CON.FERT.ZS | ... | VC.IHR.PSRC.P5 |
| 1960 | AFG | 2.95 | 2.94 | 2.89 | ... | 2.89 |
| 1961 | ... | ... | ... | ... | ... | ... |
| ... | YEM | 2.91 | 2.94 | 2.91 | ... | 2.90 |
| 2020 | ZMB | 3.24 | 3.33 | 3.26 | ... | 3.29 |
| 2021 | ZWE | 2.76 | 2.78 | 2.74 | ... | 2.68 |




WDI se organiza para que tenga una estructura de datos de la forma anterior, en tres vías, para aplicar el STATIS-dual. La versión WDI descargable trae como columnas variables los años, y los individuos filas son los indicadores por cada país, siendo este ultimo entonces la ocasión, momento o k-tabla.

Los métodos de tres modos se clasifican de la siguiente forma:

Tabla 22 *Clasificación de los métodos de tres modos.*

| Estructura de datos | |
|---|---|
| Datos en tres vías | Modo A, modo B, modo C (cubo de datos), son entradas de datos completas, se tienen todas las filas / individuos, todas las columnas / variables y todos los momentos / años o escenarios, en un cubo completo. |
| Conjuntos múltiples | • Mismas variables en diferentes ocasiones/momentos (años), pero no necesariamente los mismos individuos |
| | • Diferentes momentos, pero las variables son diferentes, pero los individuos si son los mismos |

*Nota.* Adaptado de *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado*, por Rodriguez-Martinez, 2020. [Tesis de Doctorado, Universidad de Salamanca] Repositorio Institucional – Universidad de Salamanca

WDI es un conjunto múltiple, con las mismas variables o indicadores del desarrollo, en diferentes ocasiones (años), pero no necesariamente los mismos individuos (economías o países) tienen registros de esos indicadores para cada año. Esto se ve reflejado en el alto porcentaje de datos perdidos, faltantes o NAs en el cubo de datos. En general se entiende a WDI como un cubo de datos (tres vías) pero se da un tratamiento a los NAs usando las técnicas para tal fin.

Al seguir el enfoque, de estos métodos de la escuela Salmantina de la Universidad de Salamanca, para el análisis de datos de tres modos y tablas de conjuntos múltiples, se propone en este trabajo aplicar solo uno de los siguientes que se listan, pero merecería y dado el tipo de información que es WDI que ayuda a las políticas globales contra la pobreza y a las naciones para su desarrollo, darle una mirada posible en todos y cada uno de estos métodos, lo que demandaría aun un trabajo mayor y expedito.

Tabla 23 *Métodos de tres modos de la Escuela Salmantina.*

| Autor | Año | Directores Tesis Doctoral | Método | Descripción |
|---|---|---|---|---|
| Martin-Rodriguez et al. | 1996 | Purificación Galindo-Villardón & Jose Luis Vicente-Villardón | METABIPLOT | Compara las estructuras resultantes de un análisis Biplot, luego de inspeccionar la máxima congruencia |
| Baccala | 2004 | Purificación Galindo-Villardón & María José Fernánez-Gómez | Biplot Múltiple | Alternativa Biplot al análisis factorial múltiple |
| Vallejo-Arboleda | 2004 | Jose Luis Vicente-Villardón | STATIS Canónico | Las filas tienen estructura de grupo. los grupos de individuos son los mimos, las variables pueden ser diferentes |
| Cortes-Saud | 2005 | Jose Luis Vicente-Villardón & María José Fernández-Gómez | ACPR a tres vías | Extiende la metodología del Análisis de Componentes Principales Restringido ACPR, a tres vías |
| Basso | 2006 | Jose Luis Vicente-Villardón & Vicente-Tavera | Biplot tríadico | Representaciones Biplot del análisis tríadico |
| Vallejo-Arboleda | 2008 | Jose Luis Vicente-Villardón & Purificacion Galindo-Villardón | STATIS Dual Canónico | Busca un subespacio de referencia común para el análisis de variables canónicas de todas las ocasiones |
| Mendes / Mendes et al. | 2011 / 2017 | Purificacion Galindo-Villardón & María José Fernández-Gómez | Co-Tucker (que debería ser llamado realmente como Tucker3-Co) | Combina técnicas STATICO y Tucker3 |
| Pinzón / Pinzón y Villardón | 2011 / 2012 | Jose Luis Vicente-Villardón | Biplot consenso | Obtiene un subespacio de referencia común para todas las tablas, a partir de la proyección sobre subespacios de dimensión reducida |
| Frutos-Bernal | 2014 | Jose Luis Vicente-Villardón & María José Fernández-Gómez | Análisis de datos acoplados T3-PCA | Analiza un bloque de datos de tres vías y un bloque de datos de dos vías, ambos de tipo continuo, que tienen un modo en común, el modelo tiene como restricción que la matriz de componentes del modo común ha de ser igual en ambos submodelos |
| Egido-Miguelez / Egido-Miguelez, Galindo-Villardon | 2015 / 2015 | Jose Luis Vicente-Villardón & María José Fernández-Gómez | Biplot dinámico | Obtiene un conjunto de trayectorias de los individuos y de las variables sobre un Biplot de situación de referencia |
| Rodriguez-Rosa / Rodriguez-Rosa et al. | 2016 / 2019 | Jose Luis Vicente-Villardón & Gallego-Alvarez | Co-Tucker3 | Combina los métodos Tucker3 y en el análisis de la co-inercia. Resuelve el problema de describir no solo la parte estable de estructura de datos, sino también la posibilidad de extraer la estructura latente |
| Gonzalez-Garcia | 2019 | Purificacion Galindo-Villardón & Nieto-Librero | C_enet-Tucker | Método de descomposición para datos tensoriales que hace la restricción Elastic Net |
| **Rodriguez-Martinez / Rodriguez-Martinez et al.** | **2020 / 2021** | Purificación Galindo-Villardón & Purificación Vicente Galindo | **Sparse STATIS-dual** | **Implementa la técnica de penalización elastic net en el modelo STATIS-dual y busca retener las variables más importantes del modelo, y obtener resultados más precisos e interpretables,** |

| Autor | Año | Directores Tesis Doctoral | Método | Descripción |
|---|---|---|---|---|
| | | | | **propone un software lenguaje R** |
| Gonzalez-Narvaez | 2021 | María José Fernández-Gómez & Susana Mendes | HJ STATICO & MIX STATICO | HJ STATICO: Combina STATICO y HJ Biplot), MIX STATICO: combina DISTATIS y STATICO (datos mixtos) |
| Martin-Barreiro / Martin-Barreiro et al | 2021 / 2021 | Purificación Galindo-Villardón & Ana Martín Casado | Disjoint TUCKERs | Permite calcular componentes disjuntos en el modelo PARAFAC y en los modelos TUCKER con el propósito de obtener matrices de cargas interpretables |
| Ballesteros-Espinoza et al | 2021 | Purificación Vicente Galindo, Miguel Rodríguez-Rosas & Ana Sánchez-García | STATIS-dual dicotómico | Alternativa para un análisis más eficiente de datos dicotómicos de múltiples tablas, en particular para una secuencia de matrices, todas ellas con las mismas variables en columnas pero con diferentes individuos en filas. se puede utilizar con todo tipo de datos dicotómicos, proponen un software en lenguaje R |

*Nota.* Adaptado de *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado*, por Rodriguez-Martinez, 2020. [Tesis de Doctorado, Universidad de Salamanca] Repositorio Institucional – Universidad de Salamanca

Dado el arreglo dispuesto de la matriz de entrada a las herramientas (librerías) de R que hemos de usar en este trabajo, y donde los indicadores son las variables columnas, y las economías o países son los individuos filas, y las ocasiones son los años, es que estamos aplicando el STATIS-dual, y es porque interesa precisamente enfatizar las relaciones entre variables, bajo la premisa que estas no han de cambiar. Por otro lado, si se hubiese hecho el arreglo matricial de entrada de tal manera que las variables columnas fuesen las economías o países, y los individuos filas los indicadores, manteniendo las ocasiones como los años, aplicaríamos un STATIS, donde nos interesaría enfatizar las posiciones de las economías.

En este trabajo centraremos la atención en la análisis de 3 vías Sparse STATIS-dual de la escuela Salmantina, y una comparación con el método STATIS-dual de la escuela francesa.

### El análisis de los NAs y la técnica usada

Los datos perdidos o faltantes, también llamados NA, Not Available por sus siglas en inglés, son recurrentes en todo tipo de data frames, en el caso de estudio de este trabajo WDI se tiene un alto porcentaje de NAs, esto hace que resolver esto sea fundamental.

Tabla 24 *Algoritmos de imputación de valores perdidos categorizados dentro de diferentes clases.*

| Algoritmo | Clase | Observaciones |
|---|---|---|
| **SVDimpute** | Global | Singular value decomposition based |
| BPCA | Global | Bayesian principle component analysis |
| **KNNimpute** | Local | K nearest neighbor based |
| SKNNimpute | Local | Sequencial KNN |
| IKNNimpute | Local | Iterative KNN |
| GMCimpute | Local | Gaussian mixture dustering based |
| LSimpute | Local | Single linear regression |
| LLSimpute | Local | Multiple linear regression |
| SLLSimpute | Local | Sequencial multiple linear regression |
| ILLSimpute | Local | Iterated multiple linear regression |
| RLSP | Local | Least square regression with principal components |
| BGSregress | Local | Linear and non-linear regression with Bayesian gene selection |
| CMVE | Local | Linear regression with multiple parallel imputations |
| AMVI | Local | CMVE with automatic determination of number of reference genes |
| ARLSimpute | Local | AR modeling with least square regression |
| LinCmb | Hybrid | Combining local and global approaches |
| POCSimpute | Knowledge | Using knowledge agout microarray experiment process |
| GOimpute | Knowledge | Using Gene Ontology information |
| HAlimpute | Knowledge | Using histone acetylation information |
| WeNNI | Knowledge | Using spot quality information in weighted nearest neighbor |
| WeNNLBC | Knowledge | Using one-channel depletion information for bias correction |
| iMISS | Knowledge | Using multiple external reference data sets |
| metaMISS | Knowledge | Using database matrix obtained from databases of microarray data |

*Nota*: Tomado de *Imputación de datos perdidos (s.f.), en Análisis de Datos Procedentes de Microarrays (I)* (p.37) [Diapositiva de PowerPoint], por Vicente-Villardón. Recuperado 22 Abril de 2022. [Clase del Master: Omicos. Universidad de Salamanca].

A pesar que hay muchos algoritmos para imputar datos a los NAs, se usa la media y que no requiere de una complejidad algorítmica como las de la tabla arriba, que esto se traduciría en tiempo computacional quizás. En todo caso de los

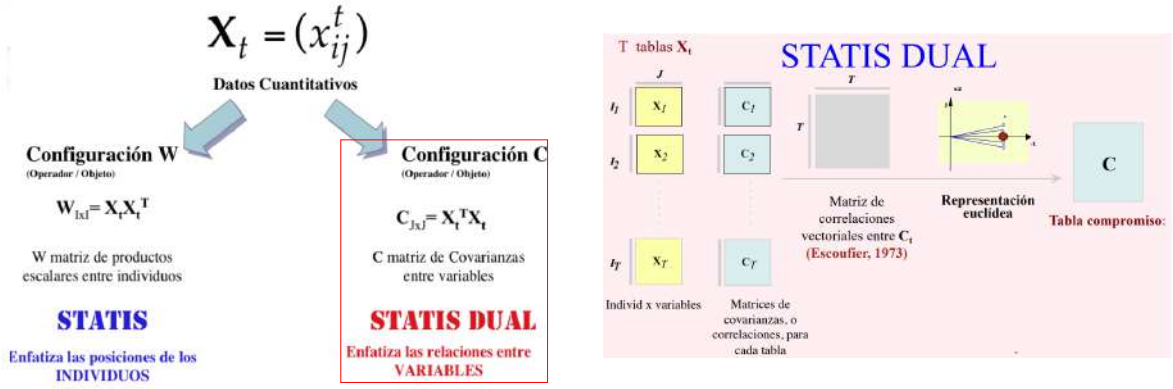
algoritmos más usados están KNNimpute o SVDimpute, que son una alternativa a evaluar en nuevos estudios.

### Sparse STATIS-dual

Las técnicas multivariantes para el análisis de datos se basan en el método de descomposición matricial que busca explorar grandes volúmenes de información, aprovechando su alta dimensionalidad. La representación más común es el Análisis de Componentes Principales PCA realizado a través de la Descomposición en Valores Singulares (DVS). Cada componente principal PC se expresa como una combinación lineal de las variables originales y establecen su contribución a cada PC.

Los métodos STATIS y STATIS-dual, de origen en la escuela francesa, son una generalización de las componentes principales, y permiten exploración simultánea de varias tablas de datos. Cuando son los mismos individuos / filas en todas las ocasiones / situaciones pero para diferentes variables / columnas o no, se habla de STATIS, y se quiere estudiar las posiciones relativas de los individuos. Cuando son las mismas variables en todas las ocasiones pero para diferentes individuos / filas o no, se habla de STATIS-dual, y se quiere estudiar las relaciones entre las variables. En el caso que sean los mismos individuos y las mismas variables en todas las ocasiones sirven ambos métodos, aunque cada uno parte de matrices diferentes como se ilustra a continuación.

Figura 16 *Representación de las matrices de datos*




*Nota*: Tomado de *Representación de las matrices de datos* (s.f.), en Análisis Multivariante de tres vías: Métodos de la familia STATIS (p.9 y 15) [Diapositiva de PowerPoint], por Galindo-Villardón. Recuperado 24 Febrero de 2022. [Clase de Master: Tablas de 3 entradas/STATIS. Universidad de Salamanca].

Tanto en el STATIS como en el STATIS-dual, se llevan a cabo a partir de cuatro etapas, siguiendo lo expuesto por Rodríguez-Martínez (2020).

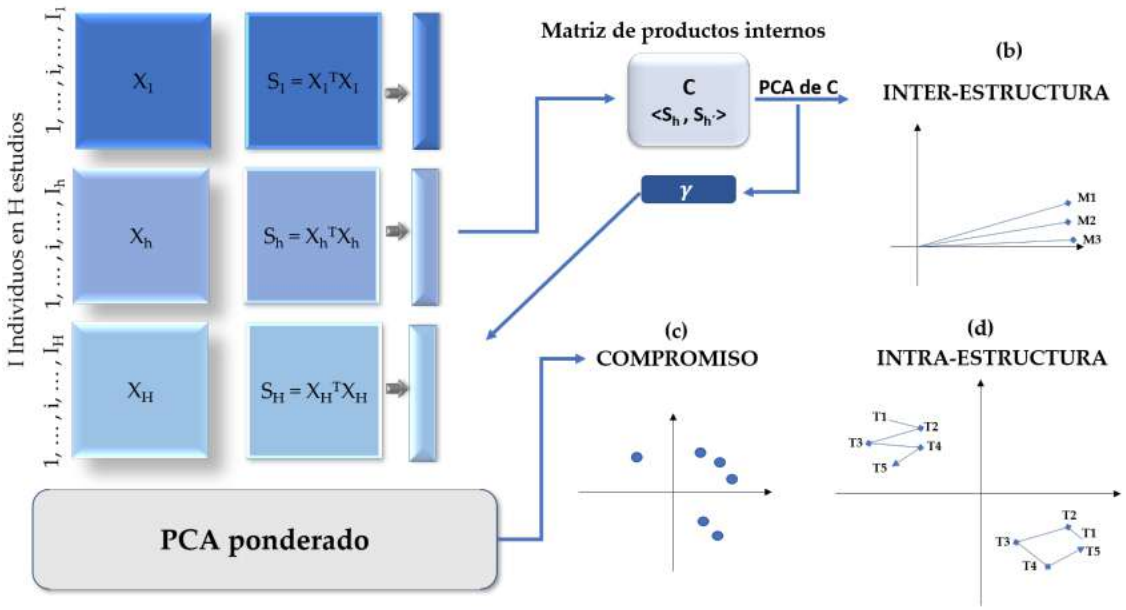
Tabla 25 *Etapas del STATIS-dual*

| Etapa | | Objetivo | Pasos |
|---|---|---|---|
| 1 | Encontrar la inter-estructura | Comparar y analizar la relación entre los diferentes conjuntos de datos | • Hallar matriz **Ct** de correlaciones vectoriales Hilbert-Schmidt (HS) entre matrices<br>○ Proceder a una comparación global de la estructura de las R matrices de datos<br>○ La inter-estructura destaca las diferencias y similitudes entre las tablas<br>• Representar la inter-estructura en un subespacio de dimensión reducida. Esto es representar las matrices de datos correspondientes a las diferentes ocasiones como puntos en un espacio vectorial de baja dimensión.<br>○ Descomposición espectral de la matriz de correlaciones vectoriales, PCA(**Ct**), y proyección sobre el subespacio de baja dimensión<br>○ Interpretación del diagrama factorial resultante del PCA (imagen euclídea): La distancia entre puntos se interpreta en términos de similitud y, por ende, en semejanza entre estructura de varianza/covarianza y congruencia entre estructuras factoriales. Las estructuras serán semejantes si los ángulos formados por los vectores de la imagen euclídea se aproximan a cero. |
| 2 | Construcción matriz compromiso **C** o estructura consenso **C** | Combinar los datos en una estructura común llamada compromiso o consenso | • Búsqueda de una estructura común a las matrices en estudio.<br>○ Calcular la matriz consenso **C** como medias ponderadas de las matrices de partida y los pesos son las componentes del primer vector propio. |
| 3 | Estudio de la intra-estructura | Analizar el compromiso mediante ACP para revelar la estructura común entre las observaciones | • Representar la estructura de cada matriz de datos en un espacio de baja dimensión |

| Etapa | | Objetivo | Pasos |
|---|---|---|---|
| 4 | Representación de las trayectorias de los individuos o de las variables | Proyectar cada uno de los conjuntos de datos originales en el compromiso para analizar las comunalidades y las discrepancias | • Se trazan las trayectorias de los individuos (en STATIS) o las variables (en el STATIS-dual) en la imagen euclídea compromiso de los individuos o de las variables.<br>• La observación de las trayectorias de los individuos o de las variables ayuda a explicar estas diferencias a nivel individual |

*Nota.* Adaptado de *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado*, por Rodriguez-Martinez, 2020. [Tesis de Doctorado, Universidad de Salamanca] Repositorio Institucional – Universidad de Salamanca

Figura 17 *Esquema de los procedimientos del STATIS-dual*



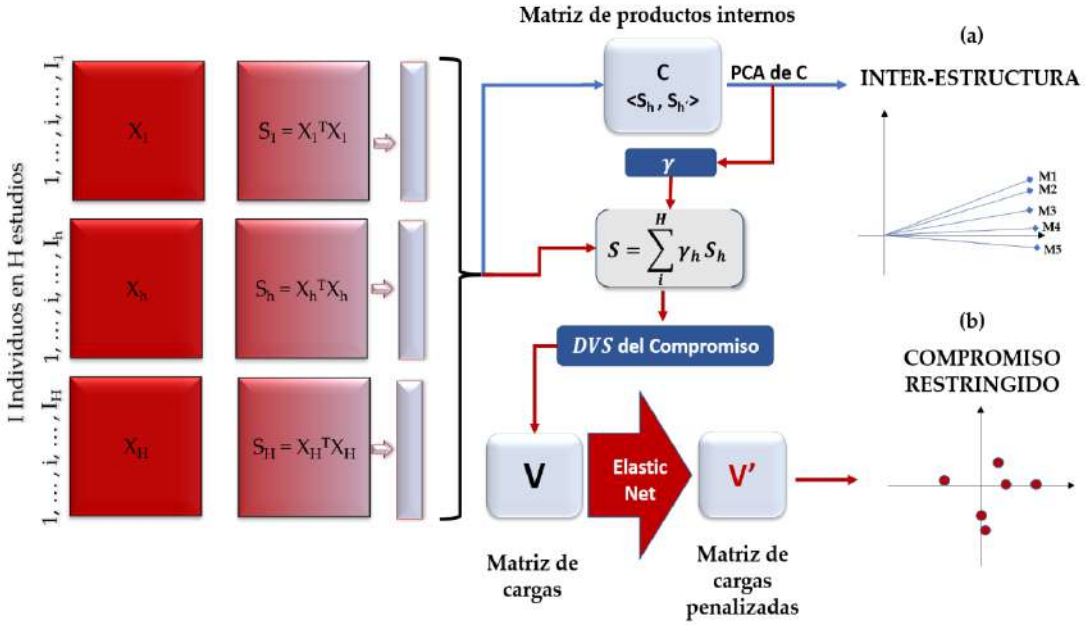

*Nota.* Tomado de *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado* (p.69), por Rodriguez-Martinez, 2020. [Tesis de Doctorado, Universidad de Salamanca] Repositorio Institucional – Universidad de Salamanca

Los coeficientes de las combinaciones lineales de las variables, denominados cargas y habitualmente distintos de cero, generan el principal inconveniente de estas técnicas multivariantes, su interpretación. Se han propuesto varios métodos para modificar el PCA, con el fin de mejorar la interpretación de sus resultados, que van desde las técnicas de rotación hasta la imposición de restricciones sobre las cargas factoriales del PCA.

Elastic Net es un método que, a partir de las restricciones en base a los métodos de regularización Ridge y Lasso, realiza simultáneamente contracción de los coeficientes y selección de variables, es decir se reducen las cargas factoriales y se hace selección de variables porque las cargas se van anulando. A las cargas modificadas se les llama cargas sparse. Elastic Net es la base del algoritmo sparse PCA que busca matrices de carga con muy pocos elementos distintos de cero (no nulos), para facilitar la interpretación de las componentes.

En la práctica, hecha la Descomposición en Valores Singulares (DVS), se penaliza la matriz de cargas V para producir componentes modificadas o componentes sparse. Esta restricción tiene entonces el efecto de contraer a cero las estimaciones de las cargas sin anularlas, con las cargas modificadas se procede a calcular las nuevas coordenadas de individuos y de variables. Cada componente es una combinación solo de las variables realmente relevantes, es decir que quedan solamente las variables más importantes para el análisis de la información que se está evaluando.

Figura 18 *Diagrama de los pasos del Sparse STATIS-dual*




*Nota.* Adaptado de *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado* (p.162), por Rodriguez-Martinez, 2020. [Tesis de Doctorado, Universidad de Salamanca] Repositorio Institucional – Universidad de Salamanca

Rodríguez-Martínez (2020) realiza la tesis doctoral que propone Sparse STATIS-dual como nueva alternativa STATIS que consiste en adaptar restricciones para contraer y/o producir cargas nulas en las componentes, con base en la teoría de regularización de Elastic Net. Esto es, penalizar o contraer las cargas de las componentes principales, que permitan la interpretación de la información que aportan los datos de alta dimensionalidad. Elastic Net se aplica después del PCA ponderado y antes de construir el compromiso. Propone además los valores de los parámetros para contraer cargas por default en Elastic Net, Alpha = 0.001 y Lambda = 0.001.

# Desarrollo y Resultados

Los datos iniciales de WDI, agrupados en 11 tópicos generales, contienen los valores de indicador para cada economía, sea un país o grupos de estos, y que dan una medida de diferentes aspectos de dicha economía relacionada con el desarrollo. Esta medida de desarrollo es actualizada periodicamente.

- Cubo de datos: Indicadores del desarrollo mundial WDI
- Filas / Individuos: **265** economías (países o agrupaciones de ellos)
- Ocasiones / momentos: **62** años, 1960 a 2021
- Columnas / Variables: **1445** Indicadores del desarrollo mundial

Antes de aplicar el método estadístico, hay que hacer algunas consideraciones al conjunto de datos inicial de WDI. En primer lugar, se cambian los nombres de los indicadores a una forma reducida que combina las dos primeras letras del tópico general al que pertenece, seguido de la posición en el vector de nombres de indicadores. Debido a que este listado consta de 1445 indicadores, a continuación un ejemplo como ilustración, pero se puede consultar el listado completo de ser necesario esto, vía el código que se ha preparado en R para este trabajo.

Tabla 26 *Ejemplo cambio de nombres del indicador.*

| **New Indicator Name** | WDI Indicator code | WDI Topic | WDI Indicator Name |
|---|---|---|---|
| **Ec**123 | DT.DOD.DIMF.CD | **Economic Policy & Debt**: External debt: Debt outstanding | Use of IMF credit (DOD, current US$) |
| Ec696 | NY.GDP.PCAP.PP.KD | Economic Policy & Debt: Purchasing power parity | GDP per capita, PPP (constant 2017 international $) |
| **He**962 | SH.H2O.SMDW.UR.ZS | **Health**: Disease prevention | People using safely managed drinking |

| New Indicator Name | WDI Indicator code | WDI Topic | WDI Indicator Name |
|---|---|---|---|
| | | | water services, urban (% of urban population) |
| He965 | SH.HIV.1524.FE.ZS | Health: Risk factors | Prevalence of HIV, female (% ages 15-24) |
| Pr421 | IC.FRM.FEMO.ZS | Private Sector & Trade: Business environment | Firms with female participation in ownership (% of firms) |
| Pr465 | IE.PPN.WATR.CD | Private Sector & Trade: Private infrastructure investment | Public private partnerships investment in water and sanitation (current US$) |

Ya se ha dicho que, en el proceso de exclusión de variables explicativas o indicadores, después de una valoración combinada de menor cantidad de datos perdidos / faltantes / NAs y con un mayor número de indicadores, para 2022 WDI los tópicos de interés son Economic Policy & Debt (346 indicadores), Health (249 indicadores) y Private Sector & Trade (167 indicadores), estos tres representan el 52.7% (762) del total (1445) de indicadores en WDI, y adicionamos además Poverty (28 indicadores), porque es un tema central que precisamente es el interés del Banco Mundial.

Esta disminución a solo 4 tópicos generales para analizar se debe a la limitante computacional, dado que demandaría mucho tiempo el algoritmo para los cálculos que precisa al generar las variables con mayor carga, aquellas que nos interesan porque proporcionan mayor información, y en consecuencia nos proveen una mejor interpretación, ubican mejor los individuos, y ayudan a ver qué características los representa mejor.

Bajo esa misma premisa, reduciremos la cantidad de momentos o años en el análisis, precisamente por alto consumo en tiempo que implicaría incluirlos todos, y como dicho antes, años 2015 a 2019 es un lapso de tiempo con menor porcentaje de NAs, además se incluye 2020 para seguir la línea de tiempo, pero no el 2021 dado su casi 98% de NAs.

De otro lado, de las 265 economías solo se considera para este trabajo los países que hacen parte de Latinamerica & Caribbean (42), sumados a economías agrupadas (7) y denotadas como se ha dicho antes : World, European Union, Latin America & Caribbean, Middle East & North Africa, OECD members, LDC Least developed countries: UN classification, y North America.

Con lo dicho, el data frame de trabajo que reduce el WDI inicial está formado de la siguiente manera:

- Filas / Individuos: **49** economías (países o agrupaciones de ellos)
- Ocasiones / momentos: **6** matrices de datos correspondientes a los años 2015 a 2020.
- Columnas / Variables: **790** Indicadores del desarrollo mundial

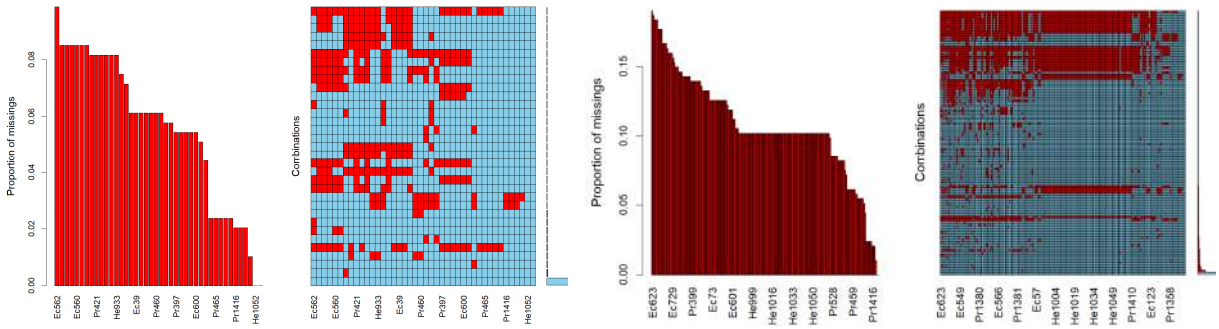
Siguiendo lo expuesto por Mallo-Fernández (2011), se excluyen variables que desde el punto de vista estadístico presentan una calidad inferior a la deseable para no ser tenidas en cuenta en el análisis cuando es elevado el porcentaje de datos faltantes, aunque hay otros aspectos a considerar en los cuales se insta a hacer un estudio más profundo, tales como: La concentración de la distribución en pocos valores, los valores extraños que ponen de manifiesto errores en la construcción de los indicadores, la existencia de colinealidad cuando haga parte importante del análisis sobre todo en la búsqueda del poder explicativo de las variables, y la no existencia de poder predictivo y/o baja asociación con la variable objetivo que se defina en casos en que se pretenda hacer análisis predictivos (ejemplo modelos predictivos para clasificar si una economía para futuros puede o no ser considerada como desarrollada).

Es necesario entonces no considerar los indicadores / columnas que contengan un % de NAs superior a un valor predeterminado, se pretende dejar las variables con una mayor cantidad de informacion, y que por criterio particular se ha fijado para analizar comparativamente dos escenarios, al 20% para 193 indicadores y al 10% para 42 indicadores. Tener presente que hasta aquí estas reducciones no obedecen en ningún caso a imputar valores por NAs aun. El data frame de trabajo queda con los siguientes escenarios:

- Filas / Individuos: 49 economías (países o agrupaciones de ellos)
- Ocasiones / momentos: 6 matrices de datos correspondientes a los años 2015 a 2020.
- Columnas / Variables: **193** o **42** Indicadores del desarrollo mundial

Al hacer la validación del estado de datos perdidos o NAs sobre el data frame encontramos desde R aun muchos NAs. En el siguiente gráfico, se muestran los indicadores ordenados de mayor a menor % de datos perdidos y la combinación de NAs y los no perdidos. A este data frame se le imputa la media por indicador / columna a los NAs para completarlo.

Figura 19 *Validación NAs*



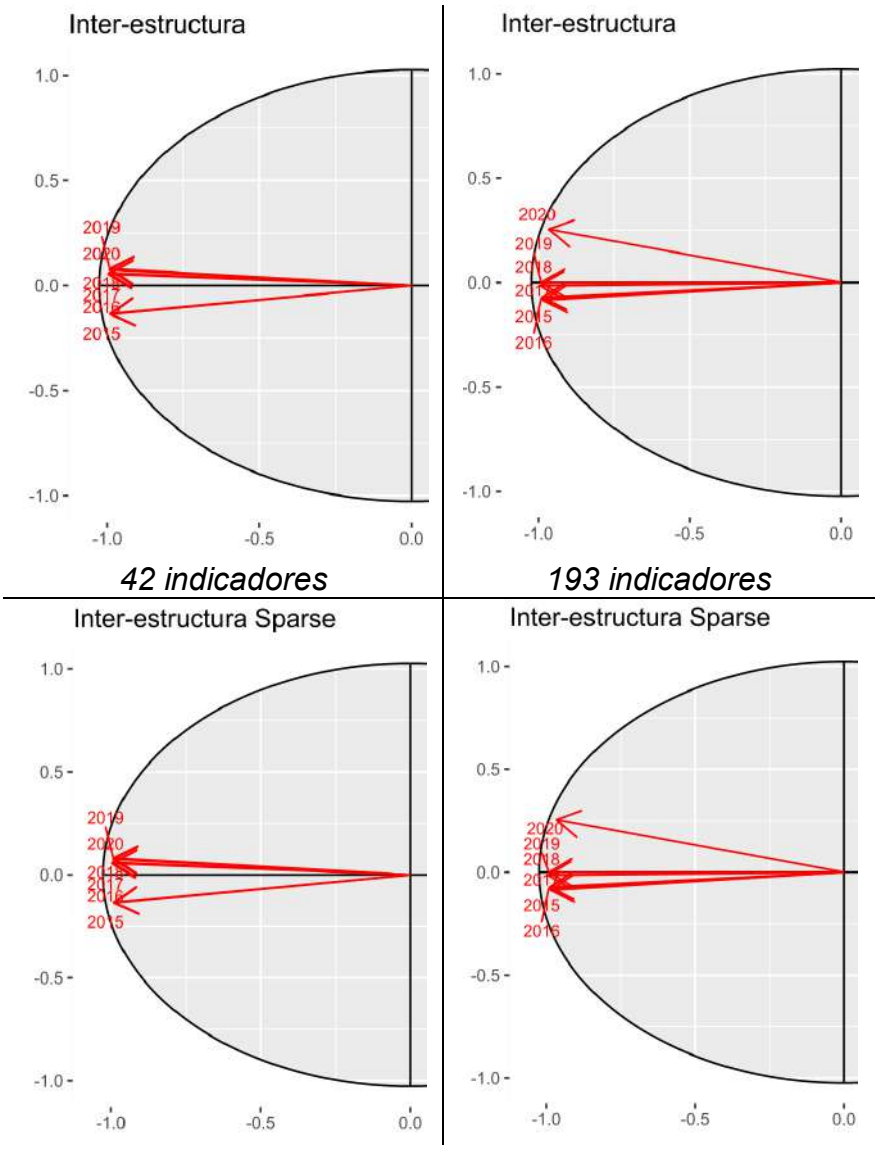

Caso para 42 indicadores

Caso para 193 indicadores

Teniendo ya la base con los indicadores completos, sea solo 42 o los 193, sin NAs, que son los mismos, y considerando que se tienen las mismas economías o individuos, para todas las ocasiones o años, se aplica el método STATIS-dual y también el Sparse STATIS-dual, y se comparan.

El análisis de las inter-estructuras es el primer paso y el fin es determinar si las estructuras de covariancias son similares en los diferentes momentos (años), esto es si las condiciones del desarrollo mundial son similares.

Figura 20 *Interestructura del análisis STATIS-dual y Sparse STATIS-dual.*

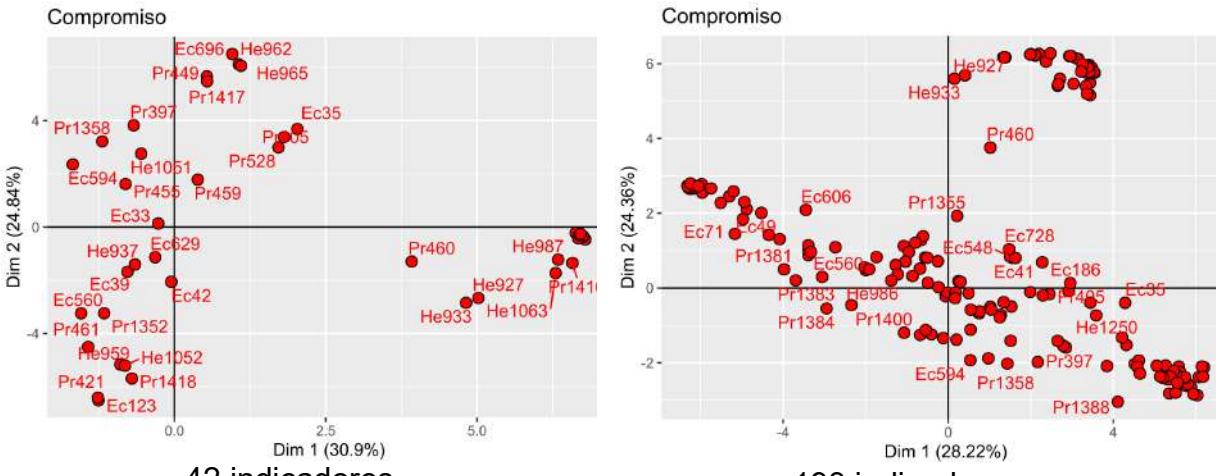



*Nota*: Solo se ilustran dos cuadrantes, los restantes dos no aportan mayor información.

Se da cuenta, tanto para STATIS-dual como para Sparse STATIS-dual, que todos los años correlacionan con el primer eje, 2017 y 2018 son los que mejor lo definen, no obstante, los años 2019 y 2020 tienen una estructura de covariancias diferentes respecto de los años 2015 y 2016 al estar en cuadrantes diferentes, lo cual puede ser sinónimo que los indicadores del desarrollo analizados han presentado cambios importantes que valdría analizar, y se intuye que es entre 2017 y 2018. Las tablas de años tienen una representación euclídea importante sobre la componente principal 1, dado que, a mayor longitud del vector en el grafico, mejor es la representación.

Seguida de la inter-estructura, para el caso de STATIS-dual, esta la construcción del compromiso entre indicadores, es decir muestra las variables sobre el compromiso, visibiliza las correlaciones dada la cercanía o no entre indicadores (correlación directa, indirecta, o nula entre indicadores), esta proyección se obtiene con la descomposición en valores y vectores singulares, que proporciona una imagen euclídea, en el que es posible proyectar las variables.

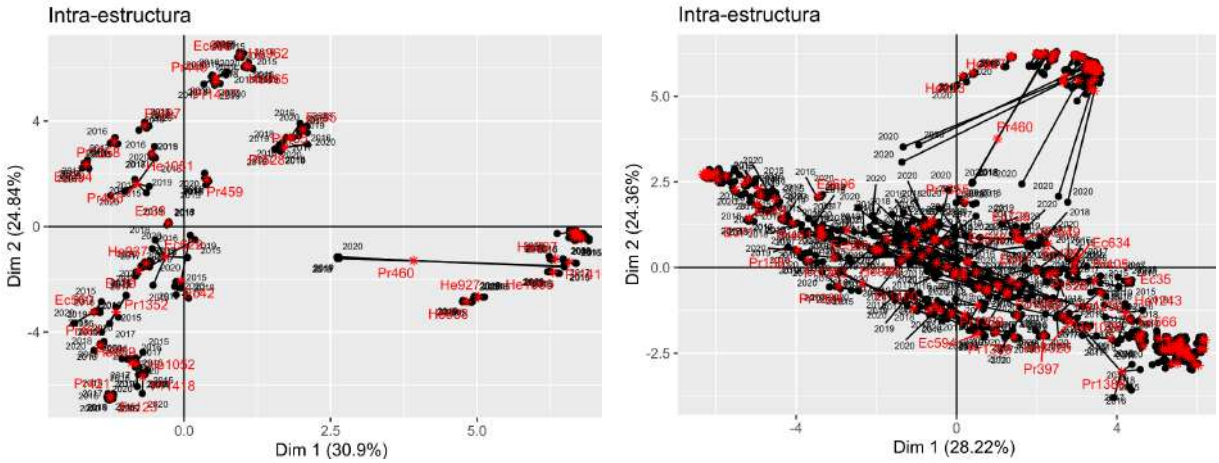
Figura 21 *Proyección del espacio Compromiso, análisis STATIS-dual*




*Nota.* Los % no corresponden a los cálculos, es una errata de la herramienta en R que se creó para Sparse STATIS-dual.

Del ejemplo con 42 indicadores, para todos los años, hay un grupo importante de indicadores que definen la segunda componente principal (por ejemplo, parte positiva están Ec696, He962 He965, Pr449, y en la parte negativa estan Ec123, Pr421, Pr1418), y hay otro más reducido de indicadores que definen la primera componente (por ejemplo: Pr1416, He987, He1063), esto de alguna manera sugiere que la segunda componente puede ser una variable latente que determina el comportamiento de los indicadores en Economías desarrolladas o no. Por el contrario, en el escenario de 193 indicadores, la contribución es mayor hacia la primera componente. En general, analizando en detalle los coeficientes de correlación se puede encontrar las relaciones que puedan existir.

Figura 22 *Intraestructura STATIS-dual*

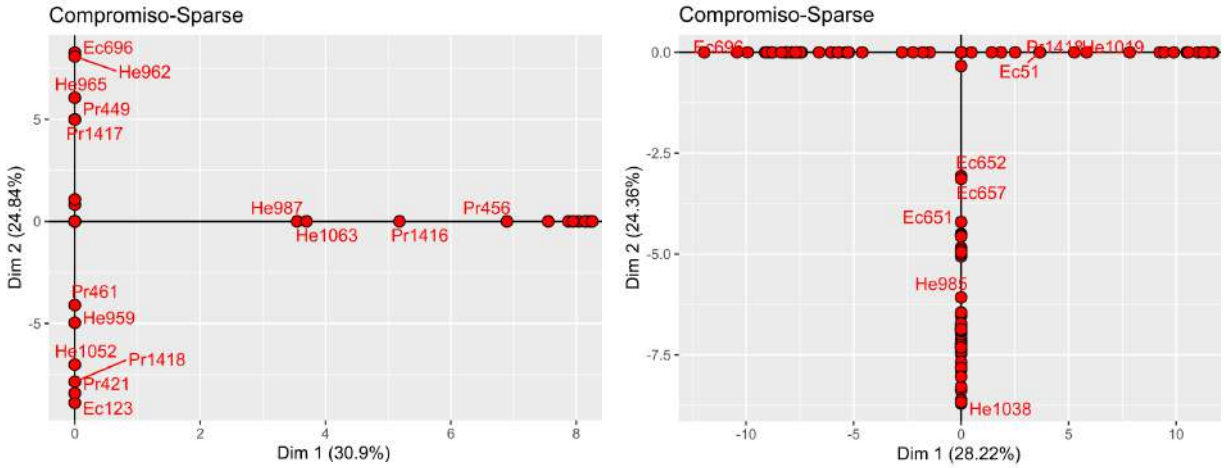



42 indicadores 193 indicadores

En el caso de Sparse STATIS-dual, nos interesa además visibilizar a que componente principal aporta más realmente cada variable, ya que va ser una

combinación solo de las relevantes, quedando solamente las variables más importantes para el análisis de la información que se está evaluando.

Figura 23 *Proyección de las variables sobre el compromiso penalizado, Sparse STATIS-dual.*



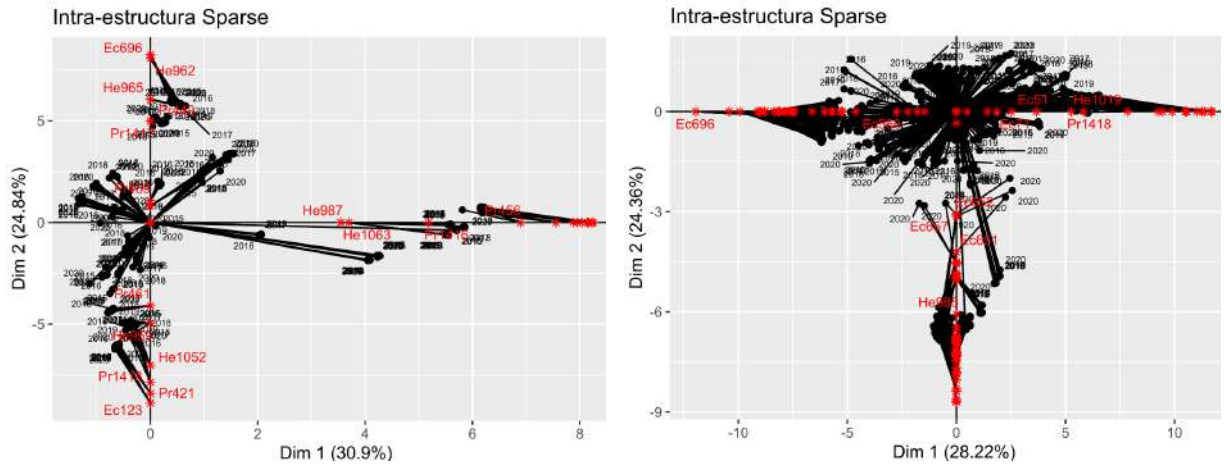

42 indicadores

193 indicadores

.

Esto nos deja que la segunda componente, en el caso de los 42 indicadores, está definida por: Ec696, He962, He965, Ec123, Pr421 y Pr1418. En la primera componente gráficamente no vemos los identificadores de los indicadores con mayores cargas, pero como ejemplo está Pr456.

Figura 24 *Intraestructura Sparse STATIS-dual.*

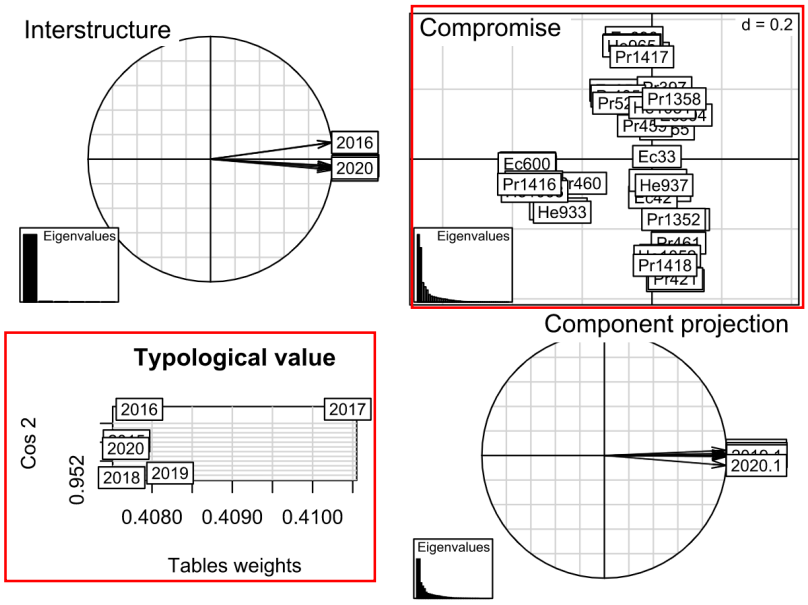



42 indicadores

193 indicadores.

Las trayectorias en la herramienta creada de Sparse STATIS-dual en R aun no están habilitadas y está pendiente de ser empaquetada la librería correspondiente, por lo cual, y solo para ilustrar la funcionalidad de las trayectorias, se usa aquí la herramienta de la escuela francesa Ade4. Hay que tener presente que hecho hasta ahora con Sparse STATIS-dual se considera como k-tabla el año, o sea cada año define cada matriz, el factor es Year, y los individuos son las economías o países.  Se puede entonces pensar la posibilidad para un análisis alternativo, y que no se hace en este trabajo con Sparse STATIS-dual, y seria tomar como k-tabla a las economías, lo que haría que el factor seria Country, y los individuos serian los años, Year.

Figura 25 *Compromiso STATIS en Ade4, caso 42 indicadores, Factor usado: Year.*

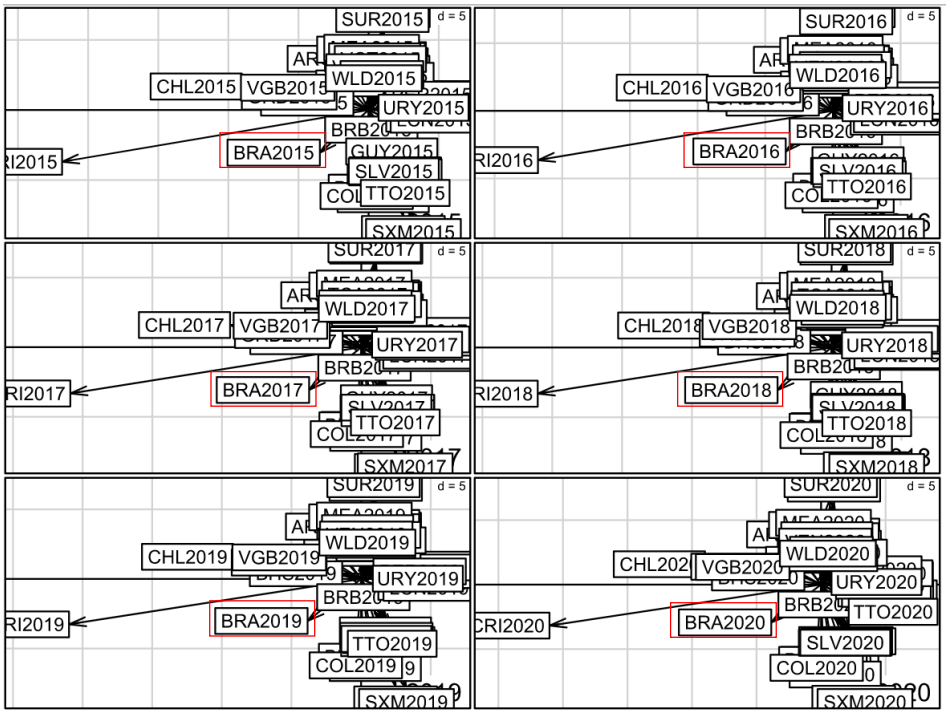



Se suma de este grafico el hecho que está informando el valor tipológico, Typological value, donde la k-tabla del año 2017 es la que mas información aporta

para construir la matriz consenso dado que tiene alto peso, y a la vez es la que está mejor representada en la matriz consenso dado su coseno al cuadrado alto. La k-tabla del 2016 está bien representada pero aporta poca información, y la k-tabla del 2018 aporta poca información al compromiso y este explica de una manera muy pobre los indicadores del desarrollo para ese año dado su cos2 bajo.

Del compromiso para indicadores, se notan 3 agrupamientos: primero en el tercer cuadrante Pr1416-He933-Otros, segundo sobre la parte positiva de la segunda componente H965-Pr1417-Otros, y tercero sobre la parte negativa de la segunda componente Pr421-Pr1418-Otros. Analizando en detalle los coeficientes de correlación encontramos las relaciones que puedan existir.

Figura 26 *Trayectorias STATIS en Ade4, caso 42 indicadores, Factor usado: Year.*

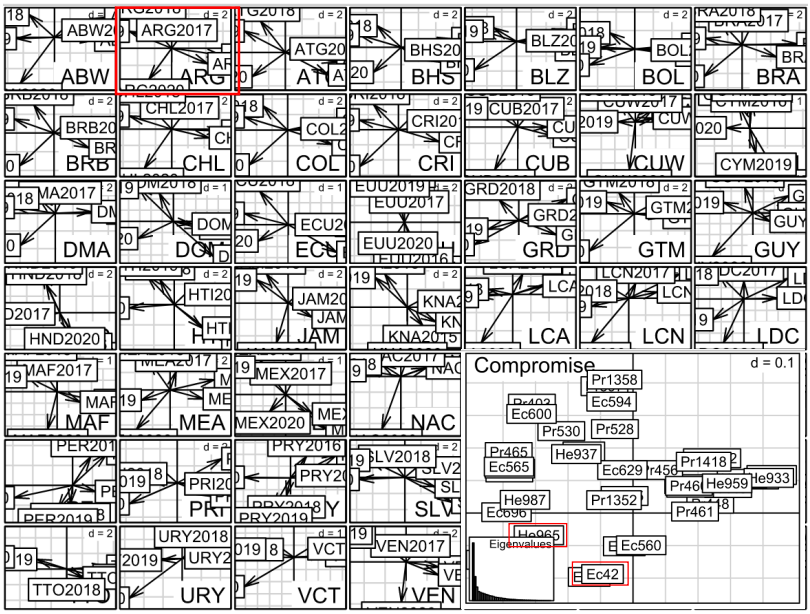



Estos mapas de trayectorias de la herramienta Ade4 dada la cantidad de individuos por cada categoría del factor Year (que son 6), es muy denso y no se

alcanza a diferenciar comportamientos para analizar en conjunto con las posiciones de los indicadores del compromiso en cada cuadrante del plano bifactorial. Se logra ver que Brasil BRA y Costa Rica CRI, ambos entre 2015 y 2020, en el STATIS apuntan siempre hacia el 3er cuadrante, y del compromiso logrado, los indicadores con mayor variabilidad en ese cuadrante son: He933 y Pr1416.

De otro lado, dando una mirada diferente, en Ade4 usando factor Country (o sea las economías seleccionadas), visualmente es denso el grafico. Se alcanza a ver por ejemplo que la economía ARG (Argentina) para 2020, en el STATIS apunta hacia el 3er cuadrante, y del compromiso logrado, los indicadores en ese cuadrante con mayor variabilidad en la misma dirección y que se logran ver son: Ec42 y He965. En cualquier caso revisando los parámetros del objeto clase statis creado después de aplicar el comando statis de la herramienta Ade4 en R, se pueden conocer las cargas y así entrar en el detalle correspondiente de cuáles son los indicadores en específico que coinciden en el cuadrante, hacer aproximaciones, y concluir las relaciones con cada economía.

Figura 27 *Compromiso y Trayectorias STATIS en Ade4, caso 42 indicadores. Factor usado: Country.*




Este análisis obedece al supuesto que no se quiera analizar la evolución en el tiempo, como hemos hecho hasta ahora, donde la k-tabla u ocasión son los años, es decir el factor es Year, sino que se quiera analizar por economía, pasando a ser estas ahora las k-tablas u ocasiones, la tercera vía, o sea ahora el factor es Country; y todo esto con el ánimo de analizar similitudes, o la estructura espacial de cada economía, para ver cómo se comporta el tiempo. En general, esto implica una estructura de datos de la siguiente forma:

Tabla 27 *Arreglo de datos para Análisis de estructura espacial por Economía (énfasis en las variables)*

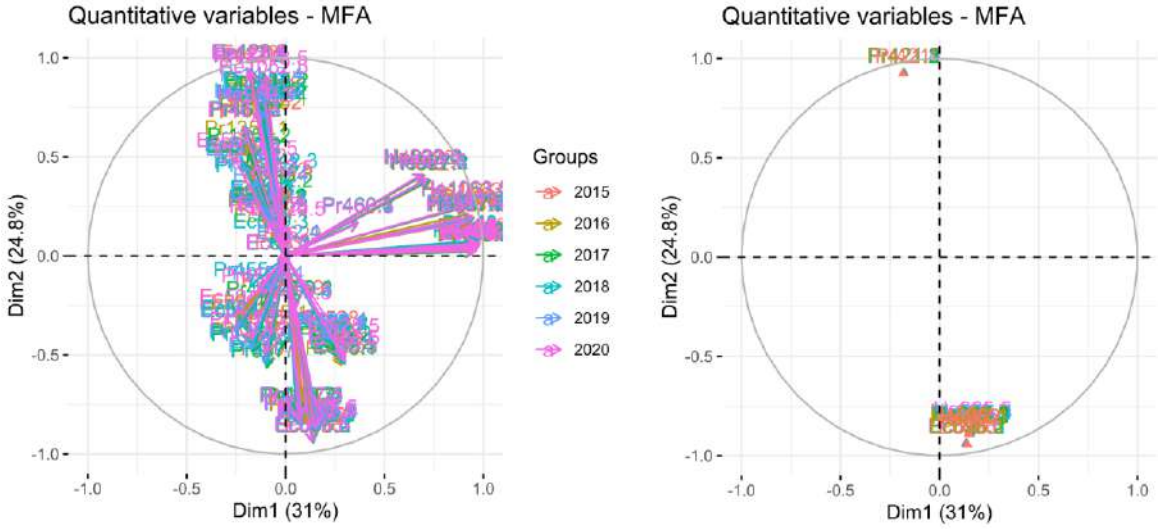
| k-tabla: Economias | Filas: años | Columnas: Indicadores | | | | |
|---|---|---|---|---|---|---|
| | | AG.AGR.TRAC.NO | AG.CON.FERT.PT.ZS | AG.CON.FERT.ZS | … | VC.IHR.PSRC.P5 |
| AFG | 1960 | 2.95 | 2.94 | 2.89 | … | 2.89 |
| … | 1961 | … | … | … | … | … |
| YEM | … | 2.91 | 2.94 | 2.91 | … | 2.90 |
| ZMB | 2020 | 3.24 | 3.33 | 3.26 | … | 3.29 |
| ZWE | 2021 | 2.76 | 2.78 | 2.74 | … | 2.68 |

Hacer cambios en la estructura de los datos nos permite usar variadas técnicas, como es el caso de un Análisis Factorial Múltiple, MFA por sus siglas en ingles, donde la estructura necesaria seria de la siguiente forma, concatenando las k-tablas de manera horizontal, pero manteniendo el énfasis en las variables o indicadores, y siendo aun el factor los años.

Tabla 28 *Arreglo de datos para Análisis Factorial Múltiple. Factor Year.*

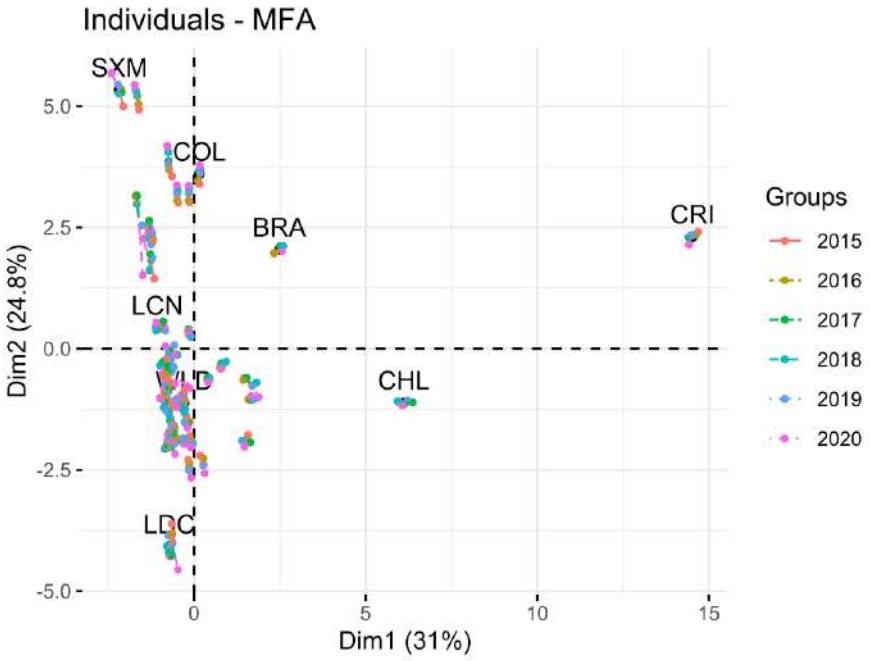
| | k-tabla: años ==> | 1960 | | | año n… | | | 2021 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Columnas: Indicadores ==> | AG.AGR.TRAC.NO | … | VC.IHR.PSRC.P5 | AG.AGR.TRAC.NO | … | VC.IHR.PSRC.P5 | AG.AGR.TRAC.NO | … | VC.IHR.PSRC.P5 |
| Filas: años | AFG | 2.95 | … | 2.89 | 21.87 | … | 22.23 | 2.89 | … | 2.89 |
| | … | … | … | … | 75073.49 | … | 79651.76 | … | … | … |
| | YEM | 2.91 | … | 2.91 | 30.04 | … | 26.90 | 2.91 | … | 2.90 |
| | ZMB | 3.24 | … | 3.26 | … | … | … | 3.26 | … | 3.29 |
| | ZWE | 2.76 | … | 2.74 | 4.18 | … | 4.13 | 2.74 | … | 2.68 |

Figura 28 *Proyección de las variables de cada k-tabla en el plano factorial. Caso 42 indicadores.*



Después de hecho el MFA, en la figura de la izquierda, donde se proyectan todas las variables, o indicadores, para todas las k-tablas, en general es validar si se forman grupos diferentes; para el caso de estudio de los 42 indicadores seleccionados, se forman 3 agrupaciones en el gráfico, y esto es que cada uno tienen comportamientos diferentes, los indicadores en esos agrupamientos califican diferente. Filtrando un poco las variables, se ve la figura anterior lado izquierdo, dejando solo unas cuantas, por ejemplo del año 2015 (Ec696, Pr421, He965, He962), del 2016 (Ec696.1, Pr421.1, He965.1, He962.1), del 2017 (Ec696.2, Pr421.2, He965.2, He962.2), del 2018 (Ec696.3, Pr421.3, He965.3, He962.3), del 2019 (Ec696.4, Pr421.4, He965.4, He962.4), y del 2020 (Ec696.5, Pr421.5, He965.5, He962.5). En principio seria tomar las variables con mayores cargas sobre las componentes y analizar, aquí ya hemos visto que estas variables coinciden con las ya vistas cuando hemos aplicado las restricciones en la técnica de Sparse STATIS-dual, pero se entiende que no debería ser la generalidad.

Figura 29 *Trayectorias o Starplot para las Economías, desde MFA. Caso 42 indicadores. Factor Year.*



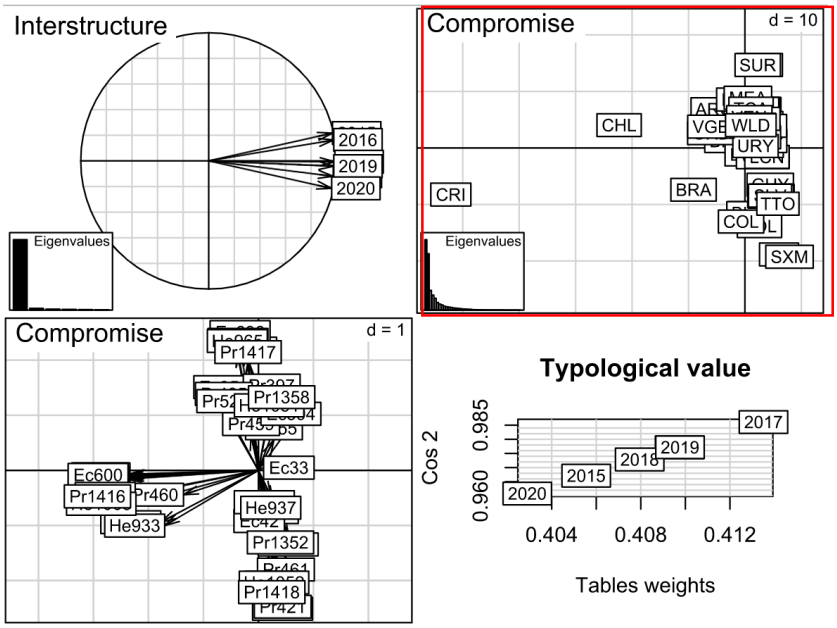
Del MFA también se logran las trayectorias de los individuos, o sea las Economías, del resultado se lee que World WLD está cerca del origen, cerca pero por debajo de la media, y LatinAmerica & The Caribbean LCN también cerca pero por encima superando la media; y estas Economías están sobre el mismo plano factorial donde se pueden comparar con las variables o indicadores si se tuviesen superpuestas allí. Por ejemplificar un caso, tomando el indicador Firms with female participation in ownership (% of firms) Pr421.5 esta posicionado prácticamente en la segunda componente del plano factorial del lado positivo y siendo el más extremo de los indicadores, ergo Latinamerica & The Caribbean LCN tiene un indicador de esa índole para el año 2021 ligeramente superior a la media, y Less Development Countries UN Classfication LDC tiene ese indicador en sus valores más bajos y lejos de la media.

Esto equivaldría a la comparación que se hace en STATIS-dual con el compromiso y las trayectorias que se obtienen en dicha técnica. En general puede que haya desde MFA una estructura factorial diferente al compromiso de STATIS-dual, pero de lo visto en estas incursiones en las técnicas y para el caso de 42 indicadores, en términos generales se ven muy similares. De hecho haciendo los giros adecuados horizontal y verticalmente a las imagenes, serian comparables; claro es que dependiendo de la técnica que se aplique se puede encontrar estructuras diferentes.

De otro lado, vale la mirada hecha desde el análisis tríadico parcial de la escuela francesa y que no utiliza operadores matriciales, PTA de sus siglas en ingles, conocido como X-STATIS, también de la herramienta Ade4 en R, manteniendo el año como factor (k-tabla), se puede ver el compromiso para los indicadores o variables y evidenciar gráficamente, y tal como ya se ha visto en

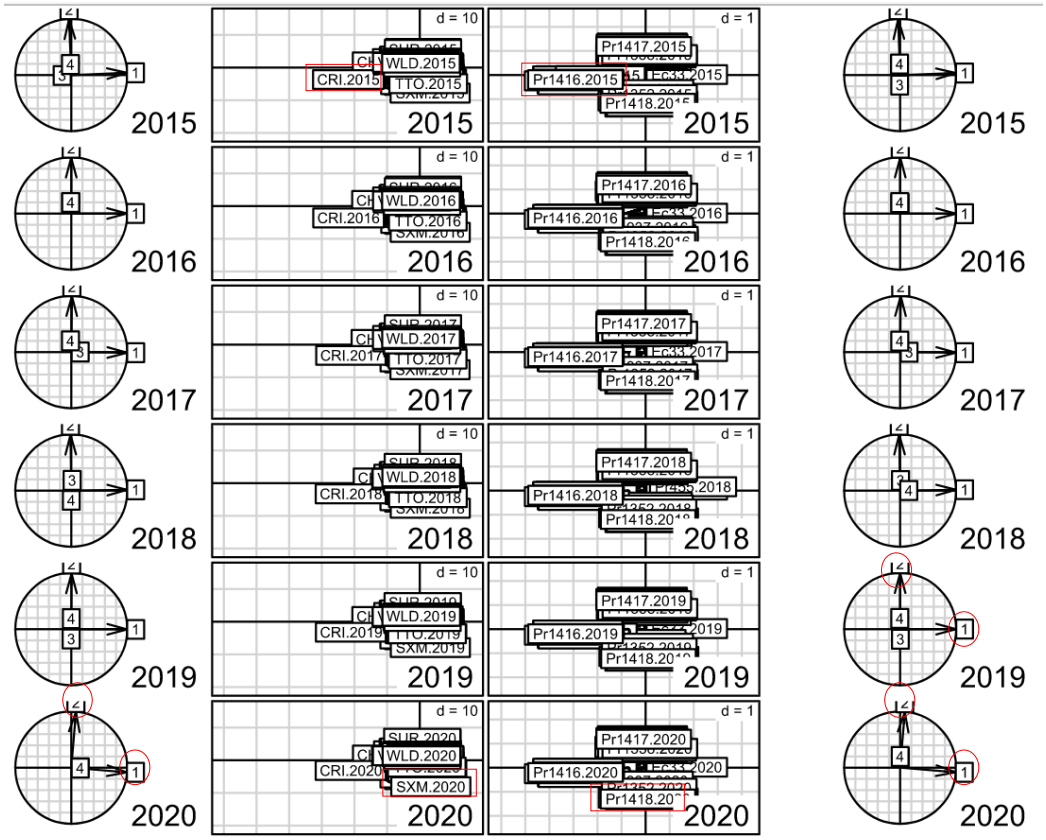
anteriores compromisos, que indicadores están correlacionados;  y desde el compromiso para las economías o individuos, apuntamos a que países son mas similares entre sí según el desarrollo.

Figura 30 *Interestructura PTA, caso 42 indicadores. Factor Year.*




En PTA se logra ver el compromiso para las economías o individuos, se nota un grupo denso alrededor del WLD (World) sobre el lado positivo de la segunda componente principal, indicativo que han mostrado niveles de desarrollo similares a través de los años; hay otro agrupamiento en la parte negativa de la segunda componente cercano a COL (Colombia), y de otro lado, un poco separados de todos están CHL (Chile) en el segundo cuadrante, BRA (Brasil) con CRI (Costa Rica) en el tercer cuadrante, que es sinónimo que tienen condiciones de desarrollo diferentes a los demás.

Figura 31 *IntraEstructura PTA, caso 42 indicadores. Factor Year.*




De la intraestructura PTA, se nota para todos los años, o sea para todas las k-tablas, que las componentes principales más relevantes son la primera y la segunda, tanto para las variables o indicadores, como para las economías o individuos. Y la relación que guardan los indicadores con las economías, se puede ver por ejemplo que SXM (Sint Maarten Dutch part) está en posición similar de Pr1418 (Ores and metals exports, % of merchandise exports), y esto lo hace que se caracterice por tener valor superior a la media de las otras economías; o como que Pr1416 (Insurance and financial services,  % of commercial service exports) tiene valor mayor a la media en el caso de CRI (Costa Rica).

En general, lo más seguro es que exista diferencias en los resultados de aplicar STATIS-dual o aplicar el PTA, ya que en uno se utiliza las matrices de

covariaciones (o en su defecto correlaciones, similitud), y en el otro solo las matrices originales. Hay que considerar que en el caso del PTA, cuando se crea el compromiso, si las variables tienen igual importancia en cada k-tabla, y si dicha k-tabla tiene un peso importante, las variables van a tenerlo por igual en el consenso, y esto no pasa en el STATIS-dual.

Volviendo a nuestra herramienta Sparse STATIS-dual de interés, de la escuela Salmantina, tomando los scores resultantes, para el caso de 42 indicadores, se tiene:

Tabla 29 *Cargas de las tres primeras componentes principales obtenidas mediante el STATIS-dual y el método de regularización Elastic Net, para el caso de 42 indicadores seleccionados*

| Indicador | STATIS-dual | | | Elastic Net (Sparse) STATIS-dual | | |
|---|---|---|---|---|---|---|
| | CP1 | CP2 | CP3 | CP1 | CP2 | CP3 |
| Ec123 | -1.253263141 | -6.509210706 | -1.168108115 | 0 | **-8.887714867** | 0 |
| Ec33 | -0.266350192 | 0.138137809 | 0.741909593 | 0 | 0 | 0 |
| Ec35 | 2.031785028 | 3.688656827 | -0.567463207 | 0 | 0.816352789 | 0 |
| Ec39 | -0.775784947 | -1.682148059 | -2.64829107 | 0 | 0 | 0 |
| Ec42 | -0.051119808 | -2.056666896 | 0.684362791 | 0 | 0 | 0 |
| Ec560 | -1.541723948 | -3.238261603 | 2.052991876 | 0 | 0 | 0 |
| Ec562 | 6.672374564 | -0.415611848 | 0.075831002 | **8.141479774** | 0 | 0 |
| Ec565 | 6.69087642 | -0.29517706 | 0.088510953 | **8.255812978** | 0 | 0 |
| Ec594 | -1.677690777 | 2.354233697 | -4.919847539 | 0 | 0 | 6.496993096 |
| Ec600 | 6.708651348 | -0.251724378 | -0.234792926 | **7.948195012** | 0 | 0 |
| Ec629 | -0.317771362 | -1.132422265 | 0.381998381 | 0 | 0 | 0 |
| Ec696 | 0.955961889 | 6.503103659 | 0.655658696 | 0 | **8.26676002** | 0 |
| He1051 | -0.547806504 | 2.760365816 | -4.918027147 | 0 | 0 | 5.677528966 |
| He1052 | -0.817074276 | -5.205855064 | -2.382279196 | 0 | **-7.018832674** | 0 |
| He1063 | 6.298387188 | -1.727316197 | -0.704499574 | 3.698550493 | 0 | 0 |
| He927 | 5.020980331 | -2.665337527 | -1.158123377 | 0 | 0 | 0 |
| He933 | 4.818506842 | -2.840844787 | -1.17890565 | 0 | 0 | 0 |
| He937 | -0.649662271 | -1.400278098 | -3.07953172 | 0 | 0 | 0 |
| He959 | -0.892792434 | -5.160602105 | -0.720304626 | 0 | -4.961093009 | 0 |
| He962 | 1.064199015 | 6.115075302 | 0.881629414 | 0 | **8.080846288** | 0 |
| He965 | 1.100118588 | 6.059234668 | 1.37132648 | 0 | 6.055800755 | 0 |
| He987 | 6.338739532 | -1.2214812 | -0.776990562 | 3.547224864 | 0 | 0 |
| Pr1352 | -1.159980402 | -3.235440812 | -2.590782543 | 0 | 0 | 0 |
| Pr1358 | -1.193304924 | 3.22312531 | -5.248712363 | 0 | 0 | **7.076601227** |
| Pr1416 | 6.575707531 | -1.346903886 | -0.519585442 | 5.179981066 | 0 | 0 |
| Pr1417 | 0.541651154 | 5.476849836 | -0.334145791 | 0 | 4.987221258 | 0 |
| Pr1418 | -0.701782198 | -5.689745754 | -1.729896147 | 0 | **-7.854921338** | 0 |
| Pr397 | -0.673514811 | 3.824097909 | -5.246910721 | 0 | 0 | 6.667405474 |
| Pr403 | 6.687755112 | -0.217158178 | -0.267152722 | **7.875555818** | 0 | 0 |
| Pr405 | 1.813401779 | 3.382122741 | -1.413647796 | 0 | 0 | 0 |
| Pr421 | -1.2617366 | -6.405977015 | -1.045221537 | 0 | **-8.417413739** | 0 |
| Pr429 | 6.700801224 | -0.274913286 | 0.047391496 | **8.159338906** | 0 | 0 |
| Pr448 | 6.786560179 | -0.464011528 | -0.147240703 | **7.554778121** | 0 | 0 |

| Indicador | STATIS-dual | | | Elastic Net (Sparse) STATIS-dual | | |
|---|---|---|---|---|---|---|
| | CP1 | CP2 | CP3 | CP1 | CP2 | CP3 |
| Pr449 | 0.538192457 | 5.663880659 | 0.761307087 | 0 | 5.000878734 | 0 |
| Pr455 | -0.807135457 | 1.618549624 | -4.346824385 | 0 | 0 | 1.758846112 |
| Pr456 | 6.629311056 | -0.217869538 | -0.518054526 | 6.895559854 | 0 | 0 |
| Pr459 | 0.38580134 | 1.785717524 | 0.378840788 | 0 | 1.077035502 | 0 |
| Pr460 | 3.920314806 | -1.29405387 | 0.063071092 | 0 | 0 | 0 |
| Pr461 | -1.423859842 | -4.501129109 | -0.606894115 | 0 | -4.100627265 | 0 |
| Pr465 | 6.754017348 | -0.348196148 | -0.019299871 | **8.224222494** | 0 | 0 |
| Pr528 | 1.719720842 | 2.995239512 | -2.111283679 | 0 | 0 | 0 |
| Pr530 | 6.745671057 | -0.337237336 | -0.120010486 | **8.036414582** | 0 | 0 |
| Sparsity | 0 | 0 | 0 | 30 | 29 | 37 |
| % Varianza Explicada | 30.90% | 24.84% | 8.83% | 34.81% | 29.62% | 9.72% |
| %Total | 64.56% | | | 74.15% | | |

Resumiendo el caso de los 193 indicadores, a continuación la varianza calculada.

Tabla 30 *Varianza explicada de las tres primeras componentes principales obtenidas mediante el STATIS-dual y el método de regularización Elastic Net, para el caso de 193 indicadores seleccionados*

| Indicador | STATIS-dual | | | Elastic Net (Sparse) STATIS-dual | | |
|---|---|---|---|---|---|---|
| | CP1 | CP2 | CP3 | CP1 | CP2 | CP3 |
| Sparsity | 0 | 0 | 0 | 138 | 141 | 178 |
| % Varianza Explicada | 28.22% | 24.36% | 7.56% | 43.68% | 30.10% | 12.44% |
| %Total | 60.13% | | | 86.22% | | |

Se observa, para tanto para el caso comparativo con 42 indicadores como para el de 193, la variabilidad explicada por cada eje en el Elastic Net (o sea el Sparse STATIS-dual) varía con relación a las del STATIS-dual. Esto es porque el método de regularización Elastic Net busca quedarse al final con los indicadores más importantes para cada componente principal, haciendo nulas las cargas menos importantes. En las primeras componentes principales se espera se concentre la mayor proporción de varianza explicada, y esto está relacionado con los valores de penalización aplicados en Elastic Net, para el caso se uso el default propuesto en la herramienta en R con Alpha = 0.001 y Lambda = 0.001.

Recordemos que bajo el principio de parsimonia se busca en este proceso seleccionar la mejor hipótesis y menos compleja entre varias igualmente soportadas por los datos, y en aras de la mejorar la interpretación final.

De la tabla anterior de scores, consideramos para comparación los indicadores (15 de los 42) con valores altos que corresponden a tres tópicos de los cuatro iniciales propuestos, Poverty no hace parte de estos. Además hay mayor protagonismo del tópico Private Sector & Trade y en lo relativo a Business Environment, no siendo el único, pero si determinante en la CP1. También el tópico de Economic Policy & Debt, del que llama la atención lo relacionado a Purchasing power parity, sobre CP2, donde se suma a los dos únicos indicadores de Health, siendo llamativo el relacionado a Number of people pushed below the $1.90 ($ 2011 PPP) poverty line by out-of-pocket health care expenditure.

Tabla 31 *Los 15 Scores altos seleccionados, del escenario para 42 indicadores, aplicando Elastic Net para las CP.*

| Indicator | Indicator code | Topic | Indicator Name | CP |
|---|---|---|---|---|
| **Ec123** | DT.DOD.DIMF.CD | Economic Policy & Debt: **External debt: Debt outstanding** | Use of IMF credit (DOD, current US$) | 2 |
| Ec562 | NE.CON.TOTL.KN | Economic Policy & Debt: National accounts: Local currency at constant prices: Expenditure on GDP | Final consumption expenditure (constant LCU) | 1 |
| Ec565 | NE.DAB.TOTL.CD | Economic Policy & Debt: National accounts: US$ at current prices: Expenditure on GDP | Gross national expenditure (current US$) | 1 |
| Ec600 | NE.RSB.GNFS.CN | Economic Policy & Debt: National accounts: Local currency at current prices: Expenditure on GDP | External balance on goods and services (current LCU) | 1 |
| **Ec696** | NY.GDP.PCAP.PP.KD | Economic Policy & Debt: **Purchasing power parity** | **GDP per capita, PPP (constant 2017 international $)** | 2 |
| He1052 | SH.UHC.NOP1.TO | Health: Universal Health Coverage | **Number of people pushed below the $1.90 ($ 2011 PPP) poverty line by out-of-pocket health care expenditure** | 2 |
| He962 | SH.H2O.SMDW.UR.ZS | Health: Disease prevention | People using safely managed drinking water services, urban (% of urban population) | 2 |
| Pr1358 | TM.TAX.MANF.BC.ZS | Private Sector & Trade: Tariffs | **Binding coverage, manufactured products (%)** | **3** |
| Pr1418 | TX.VAL.MMTL.ZS.UN | Private Sector & Trade: Exports | Ores and metals exports (% of merchandise exports) | 2 |

| | | | | |
|---|---|---|---|---|
| Pr403 | IC.CRD.PUBL.ZS | Private Sector & Trade: **Business environment** | Public credit registry coverage (% of adults) | 1 |
| Pr421 | IC.FRM.FEMO.ZS | Private Sector & Trade: **Business environment** | **Firms with female participation in ownership (% of firms)** | 2 |
| Pr429 | IC.GOV.DURS.ZS | Private Sector & Trade: **Business environment** | Time spent dealing with the requirements of government regulations (% of senior management time) | 1 |
| Pr448 | IC.TAX.DURS | Private Sector & Trade: **Business environment** | **Time to prepare and pay taxes (hours)** | 1 |
| Pr465 | IE.PPN.WATR.CD | Private Sector & Trade: Private infrastructure investment | Public private partnerships investment in water and sanitation (current US$) | 1 |
| Pr530 | LP.LPI.OVRL.XQ | Private Sector & Trade: Trade facilitation | Logistics performance index: Overall (1=low to 5=high) | 1 |

.

Esto ya nos permitiría de alguna manera dejar una línea de trabajo para fijar una política pública tomando en cuenta estos indicadores, ya que en las políticas públicas como instrumento es donde los Estados buscan racionalmente apoyar la toma de decisiones eficaces sobre temas de gran envergadura para el desarrollo, ya que se persigue en términos generales el bienestar social. Estos indicadores del desarrollo ayudan a los Estados a establecer que objetivos procura conseguir, que líneas de programas y proyectos deben ser implementados, y con qué recursos.

Complementando esta revisión de los indicadores, para el caso de 193, seleccionando valores altos de los scores en Elastic Net para las CP, se reduce los indicadores a 15 pero pudiendo ser mas.

Tabla 32 *Los 15 Scores altos seleccionados, del escenario para 193 indicadores, aplicando Elastic Net para las CP.*

| Indicator | Indicator code | Topic | Indicator Name | CP |
|---|---|---|---|---|
| **Ec123** | DT.DOD.DIMF.CD | Economic Policy & Debt: External debt: Debt outstanding | **Use of IMF credit (DOD, current US$)** | 1 |
| **Ec696** | NY.GDP.PCAP.PP.KD | Economic Policy & Debt: Purchasing power parity | **GDP per capita, PPP (constant 2017 international $)** | 1 |
| He999 | SH.STA.AIRP.P5 | Health: Mortality | Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population) | 1 |
| He1000 | SH.STA.ANVC.ZS | Health: Reproductive health | **Pregnant women receiving prenatal care (%)** | 1 |
| He1001 | SH.STA.ARIC.ZS | Health: Disease prevention | ARI treatment (% of children under 5 taken to a health provider) | 1 |
| He1003 | SH.STA.BASS.UR.ZS | Health: Disease prevention | **People using at least basic sanitation** | 1 |

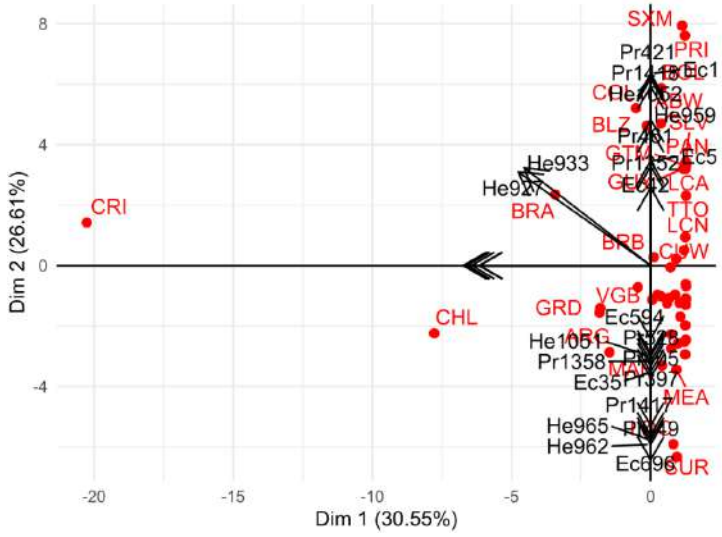
| Indicator | Indicator code | Topic | Indicator Name | CP |
|---|---|---|---|---|
| | | | **services, urban (% of urban population)** | |
| He1005 | SH.STA.BFED.ZS | Health: Nutrition | **Exclusive breastfeeding (% of children under 6 months)** | 1 |
| He1007 | SH.STA.BRTW.ZS | Health: Nutrition | **Low-birthweight babies (% of births)** | 1 |
| He1008 | SH.STA.DIAB.ZS | Health: Risk factors | Diabetes prevalence (% of population ages 20 to 79) | 1 |
| He1009 | SH.STA.FGMS.ZS | Health: Risk factors | **Female genital mutilation prevalence (%)** | 1 |
| He1021 | SH.STA.ORCF.ZS | Health: Disease prevention | Diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding) | **3** |
| He1022 | SH.STA.ORTH | Health: Disease prevention | Diarrhea treatment (% of children under 5 who received ORS packet) | **3** |
| He1023 | SH.STA.OWGH.FE.ZS | Health: Nutrition | **Prevalence of overweight, weight for height, female (% of children under 5)** | **3** |
| He1024 | SH.STA.OWGH.MA.ZS | Health: Nutrition | Prevalence of overweight, weight for height, male (% of children under 5) | **3** |
| He1290 | SP.POP.2024.MA.5Y | Health: Population: Structure | Population ages 20-24, male (% of male population) | 1 |

En este caso (escenario 193) hay muchos más indicadores del tópico Health, y ahora tan solo dos de Economic Policy & Debt: External debt coinciden con los del escenario con 42 indicadores anterior, cosa que deja visto que hay que promover líneas de trabajo sobre ellos. En el caso de Health, se nota una constante relacionada con la mujer, lo que insinúa desarrollar políticas aun más fuertes que genere el beneficio necesario a dicha población. En general, la prevalencia de uno u otro indicador, grupos de ellos y sus relaciones, cambian dependiendo del número de indicadores que se seleccionan como insumo para aplicar la técnica Sparse STATIS-dual, y esto da diversas interpretaciones a los resultados que se logran.

Tomando un solo año, como ejemplo, para 2020, y sobre la base de 42 indicadores, y en aras de complementar este trabajo, se trae a colación el método de Sparse HJ-Biplot que desarrolla Cubilla-Montilla (2019), también de la escuela Salmantina y basada en el seminal paper sobre HJ-Biplot de Galindo-Villardón (1986), logramos ver como algunas economías (países o grupos de ellos) tienen mayor o menor relación con algunos indicadores. Por ejemplo, los países LDC

(Menos desarrollados según la ONU) están más relacionados con dos indicadores de Health (He965: Prevalence of HIV, **female** (% ages 15-24), y He962: People using safely managed drinking water services, urban (% of urban population)), y con uno de Economic Policy & Debt (Ec696: Purchasing power parity - GDP per capita, PPP (constant 2017 international $)), también con Private Sector & Trade (Pr449: Business Environment - Firms expected to give gifts in meetings with tax officials (% of firms)), pero en sentido contrario sobre el mismo CP2, se tiene otro indicador relativo a Business Environment (Pr421: Firms with **female** participation in ownership (% of firms)). Por lo menos en dos indicadores, está expresamente claro el papel de la mujer.

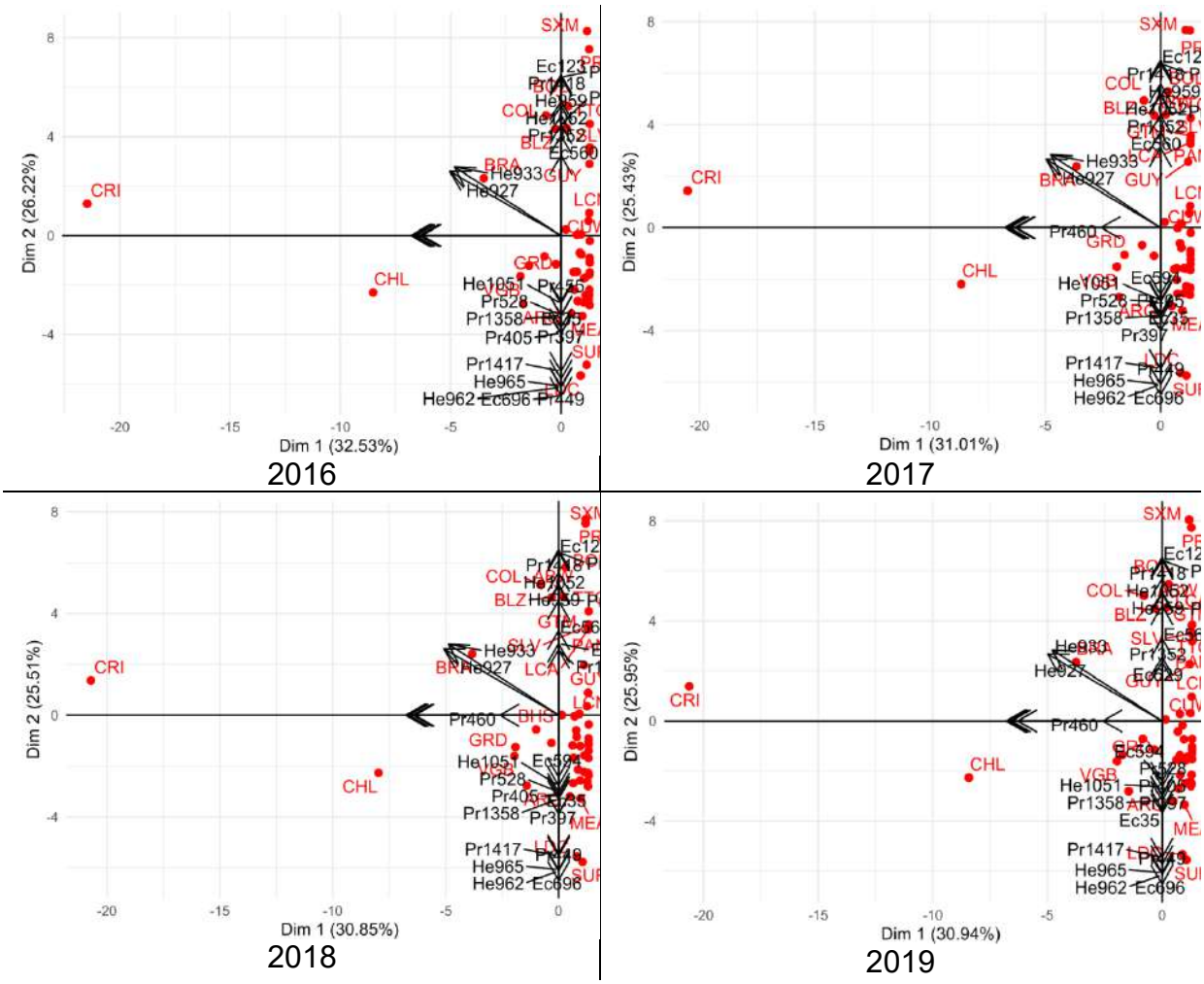
Figura 32 *Sparse HJ-Biplot con restricción Lasso del año 2020.*




Comparando con los otros años, no se ve mayor diferencia entre ellos a nivel general, pero vale la pena hacer una validación en detalle por cada economía o grupo de ellos, esto es ver cómo los países se posicionan en el plano bifactorial

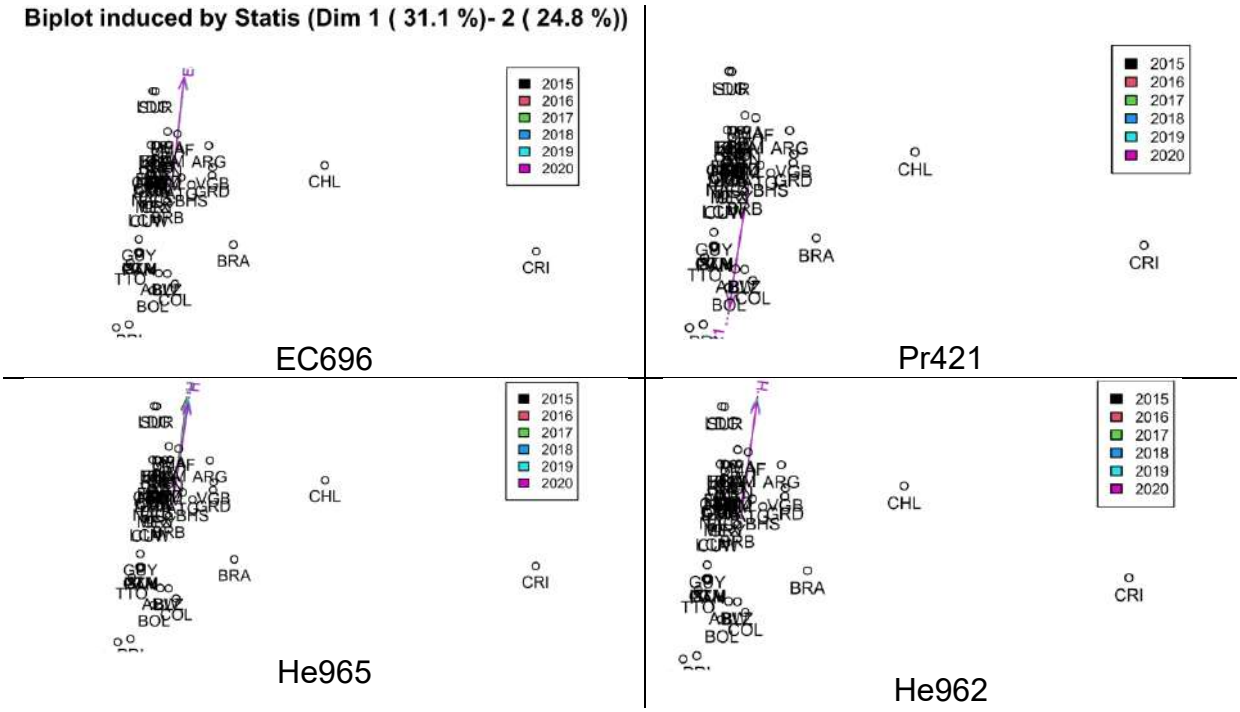
respecto a los indicadores para cada año. A continuación los Sparse HJ-Biplot de 2016, 2017, 2018 y 2019.

Figura 33 *Sparse HJ-Biplot con restricción Lasso de los años 2016 a 2019*




Otra vista es desde la herramienta MultBiplotR en R, Vicente-Villardón (2017), también de la escuela Salmantina, y se logra, para el escenario de 42 indicadores, el Biplot inducido por STATIS para un solo indicador cada vez; para tal fin se toman los indicadores de interés: Ec696, Pr421, He965, He962. Como visto antes, no hay mayor variación de la dirección de mayor variabilidad de los indicadores año tras año.

Figura 34 *Biplot inducido por STATIS*




Desde esta vista de los indicadores, de manera individual, o combinación de varios, se puede hacer una mejor reflexión del comportamiento en los diferentes años, y la relación que guardan con los países. Es el caso, a modo de reflexión, LDC países menos desarrollados y SUR Suriname está en el sentido contrario de la dirección de máxima variabilidad del indicador Pr421 para todos los años (2015-2020), y además de ser estas dos economías las más alejadas del centro del plano bifactorial; esto es que el indicador que mide el % de participación de mujeres en la propiedad, para LDC y SUR es mucho menor a la media, para todos los años entre 2015-2020. Y otra reflexión a partir del indicador He965, nos deja que los países menos desarrollados como agrupamiento y Surinam individualmente, ambas economías, mantienen altos % en mujeres con edades entre 15 y 24 años con prevalencia del VIH, para todos los años 2015-2020. Similares evaluaciones se pueden hacer para cada indicador.

Los indicadores que el Banco Mundial relaciona en la clasificación LDC son 71 para los 17 Objetivos del Desarrollo Sostenible, SDG por sus siglas en ingles, que agrupándolos por tópicos el mayor conjunto son los relacionados con Environment, y le sigue Health, ambos suman más del 50% de indicadores a considerar. De Health precisamente coinciden dos de ellos que hemos observado en este trabajo, He962 y He965. Valdría realizar un futuro análisis incluyendo variables de la dimensión Environment.

Tabla 33 *Cantidad de indicadores LDC-SDG*

| Sustainable Development Goals (SDG) | Qty indicators per SDG | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | En | He | Ec | Po | Ed | Fi | Pu | Pr | So | Ge | In | Tot | % |
| 01: No Poverty | | | | 3 | | | | | | | | 3 | 4% |
| 02: Zero Hunger | | 5 | | | | | | | | | | 5 | 7% |
| 03: Good Health and Well-Being | | 6 | | | | | | | | | | 6 | 8% |
| 04: Quality Education | | | | | 5 | | | | | | | 5 | 7% |
| 05: Gender Equality | | | | | | | | 1 | 1 | 2 | | 4 | 6% |
| 06: Clean Water and Sanitation | 1 | 4 | | | | | | | | | | 5 | 7% |
| 07: Affordable and Clean Energy | 5 | | | | | | | | | | | 5 | 7% |
| 08: Decent Work and Economic Growth | | | 2 | | | 1 | | | 1 | | | 4 | 6% |
| 09: Industry, Innovation and Infrastructure | 1 | | 1 | | | | | | | | 1 | 3 | 4% |
| 10: Reduced Inequalities | | | | 3 | | 2 | | | | | | 5 | 7% |
| 11: Sustainable Cities and Communities | 3 | | | | | | | | | | | 3 | 4% |
| 12: Responsible Consumption and Production | 2 | | 1 | | | | | | | | | 3 | 4% |
| 13: Climate Action | 2 | | | | | | | | | | | 2 | 3% |
| 14: Life Below Water | 4 | | | | | | | | | | | 4 | 6% |
| 15: Life On Land | 3 | | | | | | | | | | | 3 | 4% |
| 16: Peace, Justice and Strong Institutions | | 3 | | | | | 1 | 1 | | | | 5 | 7% |
| 17: Partnerships For the Goals | | | 4 | | | | 1 | | | | 1 | 6 | 8% |
| Qty indicators per Topic | 21 | 18 | 8 | 6 | 5 | 3 | 2 | 2 | 2 | 2 | 2 | 71 | |
| % | 30% | 25% | 11% | 8% | 7% | 4% | 3% | 3% | 3% | 3% | 3% | | |
| %Acum | 30% | 55% | 66% | 75% | 82% | 86% | 89% | 92% | 94% | 97% | 100% | | |

Con el análisis anterior hecho usando diversas técnicas, aun de manera muy general, se puede llegar a pensar en una política pública que empodere a las

mujeres en entornos de negocios, y también promover la salud preventiva de las mujeres desde su adolescencia mediante agentes educativos, además de las ya clásicas políticas públicas relacionadas solo con los aspectos económicos o del sector privado.

En 2010, el entonces presidente del Banco Mundial Robert B. Zoellick explicaba la importancia del open-data y el pensamiento detrás de la iniciativa del open-data: “I believe it’s important to make the data and knowledge of the World Bank available to everyone. Statistics tell the story of people in developing and emerging countries and can play an important part in helping to overcome poverty. They are now easily accessible on the Web for all users, and can be used to create new apps for development.” (World Bank Data Team, 2010).

# Conclusiones.

El actual estudio de WDI marca un interés en la visualización de las series, que son las variables o los indicadores, pero no se propone el análisis multivariante de ellos. La generalidad es presentar gráficos del tipo dispersión para comparar indicadores, mapas para localizar grupos de países, o gráficos lineales de tiempo para un indicador en uno o varios países o economías regionales.

Se han comparado resultados obtenidos de las técnicas multivariantes STATIS-dual y Sparse STATIS-dual, aplicados a datos sobre indicadores del desarrollo mundial, compilados anualmente por el Banco Mundial en el World Development Indicators WDI. Y Se ha demostrado que pasar del STATIS-dual al Sparse STATIS-dual reduce el número de indicadores, lo que mejora la interpretación de las variables.

El proceso de exclusión de variables explicativas o indicadores determina en gran medida los resultados, dado que depende en primera instancia de lo que se quiera analizar y también de la cantidad de datos perdidos. Esta disminución vela por dejar aquellos indicadores que nos interesan porque proporcionan mayor información, y en consecuencia nos proveen una mejor interpretación, ubican mejor los individuos, y ayudan a ver qué características los representa mejor. WDI es un conjunto múltiple, con las mismas variables o indicadores del desarrollo, en diferentes ocasiones (años), pero no necesariamente los mismos individuos (economías o países) tienen registros de esos indicadores para cada año, se esperaría que si se tuvieran datos completos pero eso no ocurre en la realidad.

En general, la prevalencia de uno u otro indicador, grupos de ellos y sus relaciones, cambian dependiendo del número de indicadores que se seleccionan como insumo para aplicar la técnica Sparse STATIS-dual, y esto da diferentes resultados y en consecuencia diversas interpretaciones a los mismos.

Interpretar es complejo y requiere del concurso de expertos en varios temas y una mirada transversal, pero estas técnicas nos ayudan a delimitar los aspectos importantes que pueden fijar una política pública consistente y que apunte al desarrollo de las economías.

Al comparar STATIS-dual con otras técnicas como PTA y MFA, da pie, al igual que al variar la cantidad de indicadores usados, a diversas interpretaciones de los resultados. Claro es que dependiendo de la técnica que se aplique se puede encontrar estructuras diferentes, y en principio seria tomar las variables con mayores cargas sobre las componentes y analizarlas. De el caso de comparativo aplicado en las tres técnicas, sobre 42 indicadores, ya hemos visto que hay coincidencias pero se entiende que no debería ser la generalidad, y cuando se comparan con Sparse STATIS-dual es más optima la selección de variables este ultimo.

El uso de las técnicas de regularización Sparse en el STATIS-dual, permite obtener soluciones eficientes a problemas de alta dimensionalidad de los datos. Sparse es una opción que podría ser aplicada a otras técnicas de la misma familia STATIS u otras familias de análisis para dos o tres vías.

La estructura de datos toma relevancia de acuerdo a la técnica que se aplica, no es igual el arreglo original de WDI, con el necesario en Sparse STATIS-dual, que difiere además del de un MFA. Y de otro lado, toma relevancia en una misma

técnica dependiendo que se quiere analizar: si la evolución en el tiempo de los indicadores y la relación con las economías, o si por el contrario es analizar similitudes de estructura espacial de cada economía, para ver cómo se comporta el tiempo.

Al complementar Sparse STATIS-dual, que es de 3 vías, con otras técnicas de 2 vías, como Sparse JH-Biplot, y el Biplot inducido por STATIS, se logra afinar la interpretación desde la vista de las variables, y de manera individual, o combinación de varias, y así se puede hacer una mejor reflexión más profunda del comportamiento en las diferentes k-tablas, y la relación que guardan con los individuos.

Para tener una más amplia interpretación lo recomendable es analizar otras perspectivas a partir de los otros métodos de la escuela Salmantina, la francesa y la anglosajona, y que no ha sido parte del presente trabajo. Valdría hacer matrices que comparen enteramente dos tópicos de WDI, es decir y como ejemplo separar el dataSet WDI en par matrices cada vez, como ejemplo una matriz X_ijk con solo con indicadores de Health, y otra matriz Y_ijk con solo indicadores de Poverty; luego aplicarles la técnica HJ-STATICO de la escuela Salmantina, y en el proceso detectar en tal interestructura o similaridad o disimilaridad en cuanto a la coestructura entre el par de tópicos de WDI que se estén analizando, y a la vez para las Economías o países, o sea los individuos comparados entre sí individualmente y por grupos que sobresalgan en planos factoriales, ya sea que presenten leves semejanzas o que por el contrario tengan marcadas diferencias en los indicadores del tópico Poverty (X_ijk) pero bajo indicadores aproximados o cercanos digamos en el tópico de Health (Y_ijk); para al final obtener un grafico HJ-Biplot combinando los indicadores más relevantes de ambos tópicos en estudio e lograr la mejor interpretación. Otros

casos interesantes a implementar para WDI pueden ser Tucker3-Co (Co-Tucker) que estudia la parte dinámica de la estructura, a diferencia de los STATIS que capturan la parte estable y explican lo que ha permanecido; métodos adicionales a considerar son C_enet-Tucker, Co-Tucker3, y Disjoint TUCKERs, etc. En general, al cubo de datos se le puede aplicar cualquier procedimiento, solo que los análisis son diferentes, estamos capturando partes diferentes de la información que aportan esas matrices, no se está distorsionándolas.

De la conclusión última, al estilo quizás del otrora Siglo de Oro cuando la Studii Salmantini propuso la doctrina sobre derechos humanos, se plantea desde este trabajo que se debe focalizar en una política pública, y que ya es bastante conocida, que empodere a las mujeres en entornos de negocios, y otra en paralelo que promueva la salud sexual preventiva de las mujeres desde su adolescencia, temas que desde la perspectiva de este análisis multivariante del desarrollo mundial son determinantes para que potencialmente una economía salga del subdesarrollo, y estas son alternativas para justificar cada aspecto de un presupuesto y programas públicos.

# Bibliografía

Arel-Bundock, V. (s.f.). *World Bank data in R*. Repositorio GitHub. Recuperado en Agosto 10 de 2022 de https://github.com/vincentarelbundock/WDI.

Azevedo, J. P. (2011). *WBOPENDATA: Stata module to access World Bank databases*. Repositorio GitHub. Recuperado en Agosto 11 de 2022 de https://github.com/jpazvd/wbopendata.

Cubilla-Montilla, M. I. (2019). *Contribuciones al análisis Biplot basadas en soluciones factoriales disjuntas y en soluciones Sparse*. [Tesis de Doctorado, Universidad de Salamanca]. Repositorio Institucional – Universidad de Salamanca.

Drupal (s.f.). *World Bank API*. Recuperado en Julio 1 de 2022 de https://www.drupal.org/project/wbapi

Galindo-Villardón, M. P. (1986). *An alternative for simultaneous representation: HJ-Biplot*. Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa.

Galindo-Villardón, M. P. (s.f.). *Análisis Multivariante de tres vías: Métodos de la familia STATIS* [Diapositiva de PowerPoint]. Recuperado 24 Febrero de 2022. [Clase de Master: Tablas de 3 entradas/STATIS. Universidad de Salamanca]

Google's Public Data Explorer (s.f.). *World Development Indicators*. Recuperado 15 Julio de 2022 de https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_.

Herzog , T. (2021). *Introducing WBGAPI: A new python package for accessing World Bank data*. World Bank Blogs. https://blogs.worldbank.org/opendata/introducing-wbgapi-new-python-package-accessing-world-bank-data.

Herzog , T. (s.f.). *A Cookbook of WBGAPI Recipes*. Repositorio nbviewer Jupyter. Recuperado Agosto 2 de 2022 de https://nbviewer.org/github/tgherzog/wbgapi/blob/master/examples/wbgapi-cookbook.ipynb.

Hurtado, F. (s.f.). *La Escuela de Salamanca: nacen los derechos humanos y la economía de mercado*. Recuperado en Agosto 25 de 2022. https://www.geografiainfinita.com/2020/09/la-escuela-de-salamanca-nacen-los-derechos-humanos-y-la-economia-de-mercado/

IDA (s.f). *What is IDA?*. Recuperado Julio 3 de 2022 de https://ida.worldbank.org/en/what-is-ida

IFM (2021). Debt Relief Under the Heavily Indebted Poor Countries (HIPC) Initiative. Recuperado Julio 10 de 2022. https://www.imf.org/en/About/Factsheets/Sheets/2016/08/01/16/11/Debt-Relief-Under-the-Heavily-Indebted-Poor-Countries-Initiative

IISD (s.f.). International Bank for Reconstruction and Development: Partial Credit Guarantees PRGs. Recuperado Julio 7 de 2022. https://www.iisd.org/credit-enhancement-instruments/institution/world-bank-international-bank-for-reconstruction-and-development/

Mallo-Fernández, F. (2011). *Modelos multivariantes internos de medición de riesgos de crédito, acorde con Basilea I*. [Tesis de Doctorado, Universidad de Salamanca]. Repositorio Institucional – Universidad de Salamanca.

Miller, S. V. (2021). Grab *World Bank Data in R with {WDI}*. Steven V. Miller Blog. Recuperado en Julio 15 de 2022. http://svmiller.com/blog/2021/02/gank-world-bank-data-with-wdi-in-r/

Piburn, J. (s.f.). *wbstats: An R package for searching and downloading data from the World Bank API*. Repositorio GitHub. Recuperado en Agosto 12 de 2022 de https://github.com/gshs-ornl/wbstats.

PyPI Python Package Index (s.f.-a). wbdata *0.3.0: pip install wbdata*. Released: Jun 27, 2020. Recuperado en Agosto 9 de 2022 de https://pypi.org/project/wbdata/.

PyPI Python Package Index (s.f.-b). *wbgapi 1.0.12: pip install wbgapi*. Released: Jul 5, 2022. Recuperado en Agosto 20 de 2022 de https://pypi.org/project/wbgapi/.

Rodríguez-Martínez, C. C. (2020). *Contribuciones a los métodos STATIS basados en técnicas de aprendizaje no supervisado*. [Tesis de Doctorado, Universidad de Salamanca]. *Repositorio* Institucional – Universidad de Salamanca.

United Nations (s.f.-a). *About Us*. Recuperado 5 Mayo de 2022 de https://www.un.org/development/desa/socialperspectiveondevelopment/what-we-do.html

United Nations (s.f.-b). *Issues*. Recuperado 5 Mayo de 2022 de https://www.un.org/development/desa/socialperspectiveondevelopment/issues.html

United Nations (s.f.-c). *Third United Nations Decade for the Eradication of Poverty 2018-2027*. Recuperado 5 Mayo de 2022 de https://www.un.org/development/desa/socialperspectiveondevelopment/2022/07/08/third-un-decade-poverty/

United Nations (s.f.-d). *World Summit for Social Development 1995*. Recuperado 5 Mayo de 2022 de https://www.un.org/development/desa/dspd/world-summit-for-social-development-1995.html

Vicente-Villardon, J. L. (s.f.). *Datos Procedentes de Microarrays (I)* [Diapositiva de Power Point]. Recuperado 22 Abril de 2022. [Clase de Master: Omicos. Universidad de Salamanca].

Vicente-Villardón, J. L. (2010). *MULTBIPLOT: package for multivariate analysis using biplots.* Departamento *de* Estadística. Universidad de Salamanca. Recuperado de http://biplot.usal.es/multbiplot/introduction.html.

Vicente-Villardón, J. L. (2017). *MultBiplotR: Multivariate Analysis using Biplot*. R package *version* 0.1.0. Recuperado de http://biplot.usal.es/classicalbiplot/multbiplot-in-r/.

World Bank (Octubre 2015). *United Nations and Its Relationship to the World Bank Group*. Recuperado 15 Mayo de 2022 de https://elibrary.worldbank.org/doi/10.1596/978-1-4648-0484-7_united_nations_and

World Bank (s.f.-a). *Data: Developer Information*. Recuperado el 15 julio de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/topics/125589-developer-information.

World Bank (s.f.-b). *Ending Extreme Poverty and Promoting Shared Prosperity*. Recuperado Abril 24 de 2022 de https://www.worldbank.org/en/news/feature/2013/04/17/ending_extreme_poverty_and_promoting_shared_prosperity

World Bank (s.f.-c). International Bank for Reconstruction and Development IBRD. Recuperado en Agosto 20 de 2022. https://www.worldbank.org/en/who-we-are/ibrd

World Bank (s.f.-d). *Least developed countries: UN classification*. Recuperado 1 de Mayo de 2022 de https://data.worldbank.org/region/least-developed-countries-un-classification.

World *Bank* (s.f.-e). *Open Knowledge Repository: World Development Indicators*. Recuperado 28 de Abril de 2022 de https://openknowledge.worldbank.org/handle/10986/2135/discover?filtertype=supportedlanguage&filter_relational_operator=equals&filter=en

World Bank (s.f.-g). *wbopendata: Stata module to access World Bank databases*. Recuperado el 15 julio de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/889464-wbopendata-stata-module-to-access-world-bank-data.

World Bank (s.f.-g). *World Bank Country and Lending Groups*. Recuperado 1 de Mayo de 2022 de https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

World Bank (s.f.-h). *World Development Indicators*. Recuperado el 27 de Abril de 2022 de https://datatopics.worldbank.org/world-development-indicators/.

World *Bank* (s.f.-i). *World Development Indicators 2014*. Recuperado 28 de Abril de 2022 de https://openknowledge.worldbank.org/handle/10986/18237

World Bank *Data* Team (2010). *Open data measures progress*. Recuperado en Mayo 10 de 2022 de https://blogs.worldbank.org/opendata/open-data-measures-progress

# Anexo Código en R

```
#-------------------------------------------------------------------------------
#Consideraciones generales
#-------------------------------------------------------------------------------
#1) En cada parte de codigo, siempre se llama un archivo de
entrada con el que
#se trabaja

#2) Por cada cambio en la estructura de datos, se renombra el
dataSet con un
#consecutivo y se graba como RDS, y sera el(los) archivo(s) de
salida de esa
#parte, y para su lectura dentro de otro apartado del codigo R



#-------------------------------------------------------------------------------
#INTRO: WDI, wbstats, WBAPI
#-------------------------------------------------------------------------------

#====== WDI
#install.packages('WDI')
#To install the development version of the package, use remotes:
#library(remotes)
#install_github('vincentarelbundock/WDI')
library(WDI)

#Searching for data
WDI::WDIsearch('gdp')
WDIsearch('gdp')[1:10,]
WDIsearch('gdp.*capita.*constant')

#Download and use the data
dat = WDI(indicator='NY.GDP.PCAP.KD',
country=c('MX','CA','US'),
        start=1960, end=2012)
head(dat)
dat #se nota que hay NAs...
#You can use country='all' to download data for all available
countries. You
#can also feed a vector of indicator strings if you want to download
multiple
#indicators at once.
datALL = WDI(indicator='NY.GDP.PCAP.KD', country='all',
start=1960, end=2012)
head(datALL)  #son muchos paises... bajar la cantidad... sino no
sale nada
datALL #hay NAs...

#Plot
library(ggplot2)
ggplot(dat, aes(year, NY.GDP.PCAP.KD, color=country)) +
geom_line() +
  xlab('Year') + ylab('GDP per capita')
#Warning message:Removed 37 row(s) containing missing values
(geom_path).
ggplot(datALL, aes(year, NY.GDP.PCAP.KD, color=country)) +
geom_line() +
  xlab('Year') + ylab('GDP per capita')
#con el caso de datALL no sale nada de grafico... son muchos...

#Monthly or quarterly data
WDI(indicator = 'DPANUSSPB', country = 'CHN', start =
'2012M01',
    end = '2012M05')

#Automatic rename
dat <- WDI(indicator = c("gdp_per_capita" = "NY.GDP.PCAP.KD",
                         "population" = "SP.POP.TOTL"))
head(dat)

#Updating series list
#To speed up search, WDI ships with a local list of all available
WDI series.
#This list will be updated semi-regularly, but you may still want to
update it
#manually to get access to the very latest data series. To do so,
use the cache
#function:
new_cache = WDIcache()
WDIsearch('gdp', cache=new_cache)


#Otras formas de consultas:
library(tidyverse)  # for most things
library(stevemisc)  # for my own vanity
library(WDI)        # for World Bank goodness
library(kableExtra) # for tables, where appropriate
WDI() %>% as_tibble()
WDIsearch("GDP") %>% as_tibble()
WDIsearch("ease of") %>% as_tibble()

WDI(indicator = c("EG.ELC.ACCS.ZS", # access to electricity
                  "BN.CAB.XOKA.GD.ZS", # current account balance
                  "IC.BUS.DFRN.XQ", # ease of doing business
                  "FP.CPI.TOTL.ZG", # CPI
                  "FR.INR.LNDP"), # interest rate spread
    start = 1960, end = 2020) %>% as_tibble() -> dat
dat

dat %>%
  dplyr::rename(elecperpop = 5,
         cab = 6,
         edb = 7,
         cpi = 8,
         ratespread = 9) -> dat
dat


dat %>%
  filter(country == "Mexico") %>%
  mutate(cpiprop = cpi/100) %>% # going somewhere with this...
  ggplot(.,aes(year, cpiprop)) +
  #theme_steve_web() +
  #Error in theme_steve_web() : could not find function
"theme_steve_web"
  geom_bar(stat="identity", alpha=.8, fill="#619cff", color="black") +
  scale_x_continuous(breaks = seq(1960, 2020, by = 10)) +
  # Below is why I like proportions
  scale_y_continuous(labels = scales::percent) +
  labs(x = "", y = "Consumer Price Index (Annual %)",
       caption = "Data: International Monetary Fund, via {WDI}",
       title = "The Consumer Price Index (Annual %) in Mexico,
1960-2020",
       subtitle = "Debt crises and currency devaluations will account
for the spikes you see.")



#====== wbstats
#install.packages("wbstats")
library(wbstats)
# Population for every country from 1960 until present
d <- wb_data("SP.POP.TOTL")
head(d)

#Hans Rosling's Gapminder using wbstats
library(tidyverse)
library(wbstats)
```

```
my_indicators <- c(
  life_exp = "SP.DYN.LE00.IN",
  gdp_capita ="NY.GDP.PCAP.CD",
  pop = "SP.POP.TOTL"
)

d <- wb_data(my_indicators, start_date = 2016)

d %>%
  left_join(wb_countries(), "iso3c") %>%
  ggplot() +
  geom_point(
    aes(
      x = gdp_capita,
      y = life_exp,
      size = pop,
      color = region
    )
  ) +
  scale_x_continuous(
    labels = scales::dollar_format(),
    breaks = scales::log_breaks(n = 10)
  ) +
  coord_trans(x = 'log10') +
  scale_size_continuous(
    labels = scales::number_format(scale = 1/1e6, suffix = "m"),
    breaks = seq(1e8,1e9, 2e8),
    range = c(1,20)
  ) +
  theme_minimal() +
  labs(
    title = "An Example of Hans Rosling's Gapminder using
wbstats",
    x = "GDP per Capita (log scale)",
    y = "Life Expectancy at Birth",
    size = "Population",
    color = NULL,
    caption = "Source: World Bank"
  )

#Using ggplot2 to map wbstats data
library(rnaturalearth)
library(tidyverse)
library(wbstats)

ind <- "SL.EMP.SELF.ZS"
indicator_info <- filter(wb_cachelist$indicators, indicator_id == ind)

ne_countries(returnclass = "sf") %>%
  left_join(
    wb_data(
      c(self_employed = ind),
      mrnev = 1
    ),
    c("iso_a3" = "iso3c")
  ) %>%
  filter(iso_a3 != "ATA") %>% # remove Antarctica
  ggplot(aes(fill = self_employed)) +
  geom_sf() +
  scale_fill_viridis_c(labels = scales::percent_format(scale = 1)) +
  theme(legend.position="bottom") +
  labs(
    title = indicator_info$indicator,
    fill = NULL,
    caption = paste("Source:", indicator_info$source_org)
  )



#===== APIs

#usando APIs en R:
#install.packages(c("httr", "jsonlite"))
library(httr)
library(jsonlite) #JSON: JavaScript Object Notation

urlAPI1 <-
'http://api.worldbank.org/v2/source/11/indicator/NY.GDP.MKTP.CD
'
urlAPI2 <-
"http://api.worldbank.org/v2/indicators/NY.GDP.MKTP.CD?format=
json"
urlAPI3 <- "http://api.worldbank.org/V2/incomeLevel/LIC/country"
urlAPI4 <- "http://api.worldbank.org/V2/country?incomeLevel=LIC"
urlAPI5 <- "http://api.worldbank.org/v2/country/br?format=json"
urlAPI6 <- "http://api.worldbank.org/v2/topic?format=json"
urlAPI7 <- "http://api.worldbank.org/v2/topic/5/indicator"
urlAPI8 <- "http://api.worldbank.org/v2/indicator?topic=5"
base_raw_data<-GET(urlAPI8)
#View(base_raw_data)
head(base_raw_data)
str(base_raw_data)
base_raw_data$content
rawToChar(base_raw_data$content)
#help("rawToChar")
WDIfromJSON <- fromJSON(rawToChar(base_raw_data$content),
flatten = T)
#Error: lexical error: invalid char in json text.
#                                 Ã¯Â»Â¿<?xml version="1.0" encod
#                 (right here) ------^
df <- WDIfromJSON$actualsTimeseries
head(df[ , c("1960","2021")])



#------------------------------------------------------------------------
#Descarga y lectura de WDI
#------------------------------------------------------------------------

setwd("~/R")

#Archivo de Entrada:
#descargando directamente de la web de World Bank:
download.file("https://databank.worldbank.org/data/download/WDI
_excel.zip",
              destfile="WDI_excelv2.zip")
unzip(zipfile='WDI_excelv2.zip')

#si se quiere cambiar a la version inicial del 27 de Abril de 2022
#unzip(zipfile="WDI_excel_v27Abr2022.439pm.zip")


#lectura y visualizacion del archivo descargado
library(readxl)
WDIEXCELData <- read_excel("WDIEXCEL.xlsx", sheet = "Data")
#TIME...

#visualizacion por consola
head(WDIEXCELData)
print(WDIEXCELData)
library(dplyr)
#dplyr::glimpse(WDIEXCELData)
glimpse(WDIEXCELData) #otra forma de ver vbles/columnas, tipo
dato, y vlres

#visualizacion por ventanas
View(WDIEXCELData) #o lo es es lo mismo: pararse encima del
objeto
#(WDIEXCELData) y darle tecla F2

#visualizacion en 'RGui' en ventana emergente tipo tabla
fix(WDIEXCELData)


#Archivo de salida:
#grabo el archivo de uso en R una vez y evito cargar desde
WDIEXCEL,
#que ademas de tiempo ocupa espacio memoria
saveRDS(object = WDIEXCELData, file = "WDIEXCELData")
```

```
#-------------------------------------------------------------------------
#Ordenar las k-tablas: estructura datos para aplicar STATIS-dual
#-------------------------------------------------------------------------

#Archivo de Entrada:
setwd("~/R")
WDIEXCELData <- readRDS(file = "WDIEXCELData")

#quitar 2 columnas 1:Country Name y 3:Indicator Name, y renombrar
dataSet1<-WDIEXCELData[,c(-1,-3)]
#dataSet1 contiene ahora 62 cols:yrs,+ 2vbles:Country Code, Indicator Code


#fix(dataSet1)
dim(dataSet1)
#[1] 384370 64: Abril 27 2022, 266 economias * 1445 indicadores * 62 years
#[1] 383572 64: Agosto 31 2022, 266 economias * 1442 indicadores * 62 years

#Resumen descriptivo: min, 1st Qu., Median, Mean, 3rd Qu. y max,
#y CONTEO de los NAs para cada columna (que son los Years)
resumeSet1vbleYr<-summary(dataSet1)
str(resumeSet1vbleYr) #table
resumeSet1vbleYr
#ejemplo de parte de la salida
#      2019              2020              2021
#Min.   :-1.276e+15   Min.   :-1.883e+15   Min.   :-4.525e+14
#1st Qu.: 6.000e+00   1st Qu.: 5.000e+00   1st Qu.: 8.000e+00
#Median : 4.400e+01   Median : 4.500e+01   Median : 5.200e+01
#Mean   : 4.437e+12   Mean   : 6.221e+12   Mean   : 7.330e+12
#3rd Qu.: 6.351e+04   3rd Qu.: 2.256e+06   3rd Qu.: 3.368e+08
#Max.   : 2.756e+16   Max.   : 3.940e+16   Max.   : 1.697e+16
#NA's   :192925       NA's   :231437       NA's   :321290


#Estudio de los NAs por YEAR
#--Guardar solo el CONTEO de los NAs por Year
yearListNAs<-resumeSet1vbleYr[7,3:ncol(resumeSet1vbleYr)]
str(yearListNAs) #named chr
namesYearListNAs<-names(yearListNAs)
namesYearListNAs
max(yearListNAs) #[1] "NA's   :374344  " tab equivale a 2 caracteres
min(yearListNAs) #[1] "NA's   :165258  " info CONTEO entre pos 9-14 de la cadena
yearListNAs[1] #1960 "NA's   :347409  " ejemplo
nchar(yearListNAs[1]) #1960 16 = tiene 16 caracteres
yearListNAs<-substr(yearListNAs,9,14) #dejando los caracteres del CONTEO
yearListNAs<-as.numeric(yearListNAs) #convertido en numerico
str(yearListNAs)
names(yearListNAs)<-namesYearListNAs #reasignar otra vez los nombres years
View(yearListNAs)
BiocGenerics::which.min(yearListNAs) #which se aplica sobre vlrs numericos
#CONCLUSION: el Year 2010 es el que menos NAs tiene para todas las vbles-filas

#--ordenar de menor NAs a mayor NAs...
i<-order(yearListNAs, decreasing = TRUE)
ii<-order(yearListNAs) #default: decreasing FALSE, mayor a menor
yearListNAs[ii]
#En la descarga WDI del Abril 27 de 2022:
# 2010     2014     2012     2015     2016     2011     2013     2017
#165258   166164   168182   169485   170887   172052   173177   173265
# 2018     2007     2009     2008     2005     2006     2004     2019
#177930   178650   178788   180075   183130   183301   194826   197942
#En la descarga WDI del Agosto 31 de 2022, hay menos NAs, WDI se actualiza:
# 2010     2014     2012     2015     2016     2011     2017     2013
#164844   165797   167763   169070   170430   171638   172484   172810
# 2018     2007     2009     2008     2005     2006     2019     2004
#176941   178150   178280   179584   182692   182838   192925   194578
#CONCLUSION: de 2010-2019 son las columnas que menos NAs tiene para todas las
#filas (Economia-indicadores) de este arreglo


#--Archivo de salida: Grabar como df para sacarlo como csv, tabla excel
yearListNAsdf <- data.frame(yearListNAs[ii])
colnames(yearListNAsdf)<-c("Qty.NAs") #cambiar nombre de la columna
yearListNAsdf$Por.NAs<-yearListNAsdf$Qty.NAs/dim(dataSet1)[1] #calculo %NAs
write.csv(yearListNAsdf, "yearListNAsdf.csv")


#--estableciendo una medida de NAs con respecto a todo el dataSet1, % global
sum(yearListNAs[ii]) #[1] 15954415 en Abril / [1] 15864834 en Agosto
((dim(dataSet1[,3:ncol(dataSet1)])[2])*(dim(dataSet1[,3:ncol(dataSet1)])[1]))
#[1] 23830940 en Abril / [1] 23781464 en Agosto
porcentajeNAs<-(sum(yearListNAs[ii]))/((dim(dataSet1[,3:ncol(dataSet1)])[2])*
                                       (dim(dataSet1[,3:ncol(dataSet1)])[1]))
porcentajeNAs #[1] 0.6694832 en Abril / [1] 0.6671092 en Agosto
#nota: En la descarga de Agosto habia menos indicadores: 1442, antes 1445


#Identificar los indicadores en una lista, tambien paises o economias, y Years
class(dataSet1) #[1]"tbl_df" "tbl" "data.frame"
dataSet1$`Country Code`<-as.factor(dataSet1$`Country Code`)
CountriesList<-levels(dataSet1$`Country Code`) #lista de paises
class(CountriesList) #[1] "character"
YearsList<-colnames(dataSet1[,c(-1,-2)]) #lista de years
class(YearsList) #[1] "character"
cantInd<-dim(dataSet1)[1]/length(CountriesList)
IndicatorsList<-c(dataSet1[1:cantInd,2]) #lista de indicadores
class(IndicatorsList) #[1]"character"
length(IndicatorsList) #[1] 1445 Abril / [1] 1442 Agosto


#creacion del dataSet2 (estructura STATIS-dual): traspuesta de dataSet1
#--se crea el dataSet vacio que va recibir las traspuestas para cada pais,
#quedando filas con years y columnas con indicadores:
cantCntry_Yr<-length(CountriesList)*dim(dataSet1[,3:ncol(dataSet1)])[2]
cantCntry_Yr
dataSet2<-data.frame(matrix(vector(), cantCntry_Yr, cantInd,
                            dimnames =
list(c(rep(YearsList,length(CountriesList))),
                                 IndicatorsList)),
                     stringsAsFactors=F)
```

```
class(dataSet2) #[1] "data.frame"
colnames(dataSet2) #las columnas quedan con los nombres de
indicadores: ok
dim(dataSet2) #[1] 16492  1445 o 1442
head(dataSet2,5)
rownames(dataSet2) # [1]"X1960" "X1961"... "X1962.16"
"X1963.16"

#--llenado de dataSet2 con info de traspuesta de Indicadores *
Year, para cada
#pais o Economia
qtyvblsO=cantInd #1445 o 1442 qty de Indicadores: ver
WDIEXCEL sheet Series
qtyvblsPrev=1
qtyvblsFin=qtyvblsO

nFilaO=length(YearsList)
nFilaPrev=1
nFilaFin=nFilaO

qtypaises=dim(dataSet1)[1]/qtyvblsO
(dim(dataSet1[,3:ncol(dataSet1)])[2])*qtypaises
#[1] 16492 = 62*266: qty de filas necesarias

for (i in 1:qtypaises){
  dataSet2[nFilaPrev:nFilaFin,1:cantInd]<-
t(dataSet1[qtyvblsPrev:qtyvblsFin,
                                    3:dim(dataSet1)[2]])
  nFilaPrev<-nFilaFin+1
  nFilaFin<-nFilaFin+nFilaO
  qtyvblsPrev<-qtyvblsFin+1
  qtyvblsFin<-qtyvblsFin+qtyvblsO
}

#---creando el factor Country en el dataSet2
VecCountries<-c()
for (i in 1:qtypaises)
  VecCountries<-c(VecCountries,c(rep(CountriesList[i],nFilaO)))
VecCountries
dataSet2$Country<-as.factor(VecCountries)

#---creando el factor years en el dataSet2
VecYears<-c()
VecYears
for (i in 1:qtypaises)
  VecYears<-c(VecYears,YearsList)
VecYears
dataSet2$Year<-as.factor(VecYears)

#---cambiando el nombre a las filas en el dataSet2
#se crea una nueva columna en el dataSet2 llamada Nombrefilas
que une columnas
#1446:pais y 1447:year
library(tidyr)
numcol<-ncol(dataSet2)
dataSet2<-unite(dataSet2,Nombrefilas,c((numcol-
1):numcol),sep="",remove=FALSE)
rownames(dataSet2)<-dataSet2$Nombrefilas
dataSet2$Nombrefilas<-NULL #se quita la columna, ahora ya son
rownames


#Archivo de salida: Grabo el archivo de uso en R una vez y evito
volver a
#procesar lo de la traspuesta del dataSet1 para conseguir a
dataSet2, que ocupa
#espacio memoria
saveRDS(object=dataSet2, file= "dataSet2")


#Estudio de los NAs por INDICADOR
#Se aplica mismo analisis de NAs ahora con dataSet2, para ver
que indicadores
#que ahora son columnas son los que tienen mayor/menor NAs
#Recordar que dataSet2 viene de dataSet1 bruto, no se ha
eliminado nada aun, ni
#indicadores ni years
#--hacer un resumen descriptivo con el min, 1st Qu., Median,
Mean, 3rd Qu. y max,
#y CONTEO de los NAs para cada INDICADOR
summary(dataSet2)
resumeSet2vbleYr<-summary(dataSet2)
indicatorsListNAs2<-resumeSet2vbleYr[7,1:cantInd]
namesIndicatorsListNAs2<-names(indicatorsListNAs2)
max(indicatorsListNAs2) #[1] "NA's   :9983  "
min(indicatorsListNAs2) #[1] "NA's   :10007  "<- posiciones 8 a 14
de la cadena
indicatorsListNAs2[1] #"NA's   :11498  "
nchar(indicatorsListNAs2[1]) #EG.CFT.ACCS.ZS 15 <- tiene 15
caracteres
nchar(min(indicatorsListNAs2[1])) #[1] 15
indicatorsListNAs2<-substr(indicatorsListNAs2,9,14)
indicatorsListNAs2<-as.numeric(indicatorsListNAs2)
names(indicatorsListNAs2)<-namesIndicatorsListNAs2
indicatorsListNAs2
BiocGenerics::which.min(indicatorsListNAs2)
#CONCLUSION: indicadores con menos NAs : (ejemplo)
#EG.CFT.ACCS.ZS         EG.CFT.ACCS.RU.ZS
EG.CFT.ACCS.UR.ZS
#         11498                  11515                  11515
#EG.ELC.ACCS.ZS         EG.ELC.ACCS.RU.ZS
EG.ELC.ACCS.UR.ZS
#          9397                   9771                   9439

#--ordenar de menor NAs a mayor NAs...
iii<-order(indicatorsListNAs2, decreasing = TRUE)
iii
iv<-order(indicatorsListNAs2) #default: decreasing FALSE, mayor
a menor
iv
indicatorsListNAs2[iv]
#tomando varias (de 1445 0 1442 indicadores) lineas, hay una
gran mayoria
#de indicadores del tipo 'SP', que en WDIEXCEL.xlsx sheet
'Series' estos
#corresponden a indicadores de : Health (varios: Mortality,
#Population: Dynamics, Population: Structure, Reproductive
health),
#Environment: Density & Urbanization, Gender: Agency, and
#Infrastructure: Technology...
#PERO hay que confirmar cuales topicos realmente son
#SP.POP.TOTL           SP.RUR.TOTL.ZS
SP.URB.TOTL.IN.ZS
#        105                    196                    196
#SP.RUR.TOTL            SP.URB.TOTL            SP.POP.GROW
#        229                    229                    372
#... y hasta 1445 o 1442 indicadores...

#---estableciendo una medida de NAs con respecto a todo el
dataSet2
sum(indicatorsListNAs2[iv]) #[1] 15954415 en Abril, [1] 15864834
en Agosto
sum(indicatorsListNAs2) #[1] 15954415
((dim(dataSet2[,1:ncol(dataSet2)])[2])*(dim(dataSet2[,1:ncol(dataS
et2)])[1]))
#[1] 23830940 en Abril, [1] 23814448 en Agosto
porcentajeNAs<-(sum(indicatorsListNAs2[iv]))/

((dim(dataSet2[,1:ncol(dataSet2)])[2])*(dim(dataSet2[,1:ncol(dataS
et2)])[1]))
porcentajeNAs #[1] 0.6694832 en Abril, [1] 0.6661853 en Agosto

#---convierto por c/indicador su qty NAs en un %, teniendo en
cuenta que el
#total de filas de dataSet2
indicatorsListNAs2porc <-indicatorsListNAs2/dim(dataSet2)[1]
indicatorsListNAs2porc
```

```
#---ordenar de menor a mayor % de NAs para todas los
indicadores:
v<-order(indicatorsListNAs2porc, decreasing = TRUE)
v
vi<-order(indicatorsListNAs2porc) #default: decreasing FALSE,
mayor a menor
vi
indicatorsListNAs2porc[vi]
#Hay indicadores con % de NAs <= 5%: pero son pocas, 9
#      SP.POP.TOTL          SP.RUR.TOTL.ZS
SP.URB.TOTL.IN.ZS
#      0.006366723          0.011884550          0.011884550
#      SP.RUR.TOTL            SP.URB.TOTL
SP.POP.GROW
#      0.013885520          0.013885520          0.022556391
#      SP.URB.GROW           AG.LND.TOTL.K2
EN.POP.DNST
#      0.029893282          0.032743148          0.034683483
#   SP.RUR.TOTL.ZG          SP.DYN.CBRT.IN
SP.DYN.CDRT.IN
#      0.060271647          0.081918506          0.083373757
#   AG.SRF.TOTL.K2         SP.POP.0004.FE.5Y
SP.POP.0004.MA.5Y
#      0.083798205          0.090831919          0.090831919
#SP.POP.0014.TO.ZS         SP.POP.0014.FE.ZS
SP.POP.0014.MA.ZS
#      0.090831919          0.090831919          0.090831919
#SP.POP.0509.FE.5Y         SP.POP.0509.MA.5Y
SP.POP.1014.FE.5Y
#      0.090831919          0.090831919          0.090831919
#... y hasta 1445 0 1442 indicadores

#Archivo de Salida: Se utiliza adelante en el anallisis cruzado NAs
por Topico
saveRDS(indicatorsListNAs2porc,file="indicatorsListNAs2porc")



#al no haber ninguna al 0%, o sea que no tenga NAs, que es lo
mismo que tenga
#100% en vlrs, se intuye no hay una vble que pueda ser
catalogada para en un
#escenario de prediccion llamarla como la respuesta Y, y sea
usada
#como tal, pero en general se puede tomar una y ver el
comportamiento del resto
#en un escenario de regresiones y clusters
#PERO en general se deben revisar estos indicadores con bajos
% NAs a ver que
#significan y que puedan ser un indicador global, quizas un vlr de
clasificacion
#si es posible asi de interpretar, etc...
#AUNQUE por la clasificacion de pais desarrollado o no, se puede
crear una
#vble a partir del la clasificacion que establece la ONU (grupo
LDC) y sumar a
#eso otro nivel que sea pais OCDE y el ultimo nivle es no es LDC
ni es OCDE.

#Se anotaba en el MODULO 10 del master en clase de Big Data
Business Intelligence
#BI que Cuando nos encontramos con un porcentaje de valores
ausentes superior al
#5% consideramos el valor "faltante" como una categoría más de
la variable de
#cara a la construcción del modelo, ya que puede constituir un
grupo homogéneo
#frente al incumplimiento (MODEL.MltVARIANT MEDICION
RIESGO CREDITO II_2022.ppt
#slide 26)

#usando estas 9 vbles con % de NAs tan (mas) bajos (o usar las
indicadores del
#tipo .TOTL que son TOTALES), se podria hacer los analisis y
aplicar desde ya el
#na.omit y asi tener una matriz para 2 years por cada pais y 9
vbles, y sin
#datos faltantes o perdidos, y de ahi aplicable la permutacion y el
bootstrap de
#omicos




#-------------------------------------------------------------------------
#Analisis de Seleccion de TOPICOS
#-------------------------------------------------------------------------

#Archivo de entrada
setwd("~/R")
library(readxl)
WDIEXCELSeries <- read_excel("WDIEXCEL.xlsx", sheet =
"Series")

#--A que topico corresponde cada indicador?, se crea el dataSet
con Topico
dataSet3Topics<-WDIEXCELSeries[,1:2]
head(dataSet3Topics,5)
#       Series Code                          Topic
#1    AG.AGR.TRAC.NO Environment: Agricultural production
#2 AG.CON.FERT.PT.ZS Environment: Agricultural production
#3    AG.CON.FERT.ZS Environment: Agricultural production
#4    AG.LND.AGRI.K2              Environment: Land use
#5    AG.LND.AGRI.ZS              Environment: Land use
names(dataSet3Topics)
dataSet3TopicsShortList<-strsplit(dataSet3Topics$Topic,": ")
dataSet3Topics$TopicFull<-dataSet3Topics$Topic
for (i in 1:dim(dataSet3Topics)[1])
  dataSet3Topics$Topic[i]<-dataSet3TopicsShortList[[i]][1]
dataSet3Topics$Topic<-as.factor(dataSet3Topics$Topic)
dataSet3Topics<-as.data.frame(dataSet3Topics)
dataSet3Topics
levels(dataSet3Topics$Topic)
#[1] "Economic Policy & Debt"    "Education"
#[3] "Environment"               "Financial Sector"
#[5] "Gender"                    "Health"
#[7] "Infrastructure"            "Poverty"
#[9] "Private Sector & Trade"    "Public Sector"
#[11] "Social Protection & Labor"


#--Separacion de los indicadores por Topico General
EconomicPolicyDebt<-
dataSet3Topics[dataSet3Topics$Topic=="Economic Policy &
Debt",]
EconomicPolicyDebt
dim(EconomicPolicyDebt) #[1] 346   2
rownames(EconomicPolicyDebt) #30:77, 85:195, 305:308,
541:737


#--creacion del ID corto para los indicadores, que se usa
#para reemplazar los nombres de indicadores que son muy largos
#ID corto de indicador es 2 letras del topico + la pos entre 1(1a en
la fila de
#dataSet3Topics) y la ultima fila.
#no necesariamente el orden establecido en dataSet3Topics es el
mismo como
#aparecen los indicadores en columnas en el dataSet2 (traspuesta
del dataSet1)
dataSet3Topics$IDVble<-c(rep("c",dim(dataSet3Topics)[1]))
dim(dataSet3Topics)
library(tidyverse) #https://es.r4ds.hadley.nz/cadenas-de-
caracteres.html
for (i in 1:dim(EconomicPolicyDebt)[1])

dataSet3Topics$IDVble[as.numeric(rownames(EconomicPolicyDe
bt)[i])]<-str_c("Ec",
```

```
rownames(EconomicPolicyDebt)[i])
dataSet3Topics$IDVble

#---repetimos para cada topico (separacion) e indicadores (ID)
Education<-dataSet3Topics[dataSet3Topics$Topic=="Education",]
dim(Education) #[1] 147   2
rownames(Education) #766:912
for (i in 1:dim(Education)[1])
  dataSet3Topics$IDVble[as.numeric(rownames(Education)[i])]<-
str_c("Ed",
                                        rownames(Education)[i])

Environment<-
dataSet3Topics[dataSet3Topics$Topic=="Environment",]
dim(Environment) #[1] 140   2
rownames(Environment) #1:29, 196:296, 668, 679:680, 689,
697:698, 1339:1344
for (i in 1:dim(Environment)[1])

dataSet3Topics$IDVble[as.numeric(rownames(Environment)[i])]<-
str_c("En",

rownames(Environment)[i])

FinancialSector<-
dataSet3Topics[dataSet3Topics$Topic=="Financial Sector",]
dim(FinancialSector) #[1] 55  2
rownames(FinancialSector) #78:84, 297:304, 309:339, 669:672,
733:734, 765,
#1106:1107
for (i in 1:dim(FinancialSector)[1])

dataSet3Topics$IDVble[as.numeric(rownames(FinancialSector)[i])]
<-str_c("Fi",

rownames(FinancialSector)[i])

Gender<-dataSet3Topics[dataSet3Topics$Topic=="Gender",]
dim(Gender) #[1] 15  2
rownames(Gender) #913:925, 1266:1267
for (i in 1:dim(Gender)[1])
  dataSet3Topics$IDVble[as.numeric(rownames(Gender)[i])]<-
str_c("Ge",
                                        rownames(Gender)[i])

Health<-dataSet3Topics[dataSet3Topics$Topic=="Health",]
dim(Health) #[1] 249   2
rownames(Health) #926:1081, 1243:1338, 1345
for (i in 1:dim(Health)[1])
  dataSet3Topics$IDVble[as.numeric(rownames(Health)[i])]<-
str_c("He",
                                        rownames(Health)[i])

Infrastructure<-
dataSet3Topics[dataSet3Topics$Topic=="Infrastructure",]
dim(Infrastructure) #[1] 39  2
rownames(Infrastructure) #59:60, 280:281, 340, 466:475, 507:523,
1326:1327,
#1386, 1409, 1415, 1435:1436
for (i in 1:dim(Infrastructure)[1])

dataSet3Topics$IDVble[as.numeric(rownames(Infrastructure)[i])]<-
str_c("In",

rownames(Infrastructure)[i])

Poverty<-dataSet3Topics[dataSet3Topics$Topic=="Poverty",]
dim(Poverty) #[1] 28  2
rownames(Poverty) #1082:1105, 1108:1111
for (i in 1:dim(Poverty)[1])
  dataSet3Topics$IDVble[as.numeric(rownames(Poverty)[i])]<-
str_c("Po",
                                        rownames(Poverty)[i])

PrivateSectorTrade<-
dataSet3Topics[dataSet3Topics$Topic=="Private Sector &
Trade",]
dim(PrivateSectorTrade) #[1] 167   2
rownames(PrivateSectorTrade) #396:465, 524:532, 1346:1434,
1437:1438
for (i in 1:dim(PrivateSectorTrade)[1])

dataSet3Topics$IDVble[as.numeric(rownames(PrivateSectorTrade
)[i])]<-str_c("Pr",

rownames(PrivateSectorTrade)[i])

PublicSector<-dataSet3Topics[dataSet3Topics$Topic=="Public
Sector",]
dim(PublicSector) #[1] 101   2
rownames(PublicSector) #341:395, 476:506, 533:540, 1439:1445

for (i in 1:dim(PublicSector)[1])

dataSet3Topics$IDVble[as.numeric(rownames(PublicSector)[i])]<-
str_c("Pu",

rownames(PublicSector)[i])

SocialProtectionLabor<-
dataSet3Topics[dataSet3Topics$Topic=="Social Protection &
Labor",]
dim(SocialProtectionLabor) #[1] 158   2
rownames(SocialProtectionLabor) #738:764, 1112:1242
for (i in 1:dim(SocialProtectionLabor)[1])

dataSet3Topics$IDVble[as.numeric(rownames(SocialProtectionLa
bor)[i])]<-str_c("So",

rownames(SocialProtectionLabor)[i])

dataSet3Topics$IDVble
#los nombres cortos permiten mejorar la visualizacion en las
graficas


#---totalizando solo para verificar que no haya escapado algun
topico
totalVbles<-dim(EconomicPolicyDebt)[1] + dim(Education)[1]+
dim(Environment)[1]+

dim(FinancialSector)[1]+dim(Gender)[1]+dim(Health)[1]+dim(Infras
tructure)[1]+

dim(Poverty)[1]+dim(PrivateSectorTrade)[1]+dim(PublicSector)[1]+
  dim(SocialProtectionLabor)[1]
totalVbles #[1] 1445 o 1442


#Archivo de salida: Grabo el archivo de uso en R una vez y evito
cargar desde
#WDIEXCELSeries y volver a procesar lo de los topicos por
indicador (factor
#levels x 11), que ocupa espacio memoria
saveRDS(object=dataSet3Topics, file= "dataSet3Topics")


#Cantidad de Indicadores por Topico General
#--formas de contar por valor en una columna, i.e. Topic: solo son
guias.
data.frame(table(dataSet3Topics$Topic))
library(plyr)
plyr::count(dataSet3Topics$Topic)
ddply(dataSet3Topics,.(Topic),nrow)

#--conteo de las cantidad de indicadores por Topico General
conteo<-data.frame(table(dataSet3Topics$Topic))
conteo
conteo$porcentage<-conteo$Freq/sum(conteo$Freq)
```

```
conteo<- conteo[order(conteo$porcentage,decreasing=TRUE), ]
conteo$procentageAcum[1]<-conteo$porcentage[1]
for (i in 2:dim(conteo)[1])
  conteo$procentageAcum[i] <- conteo$procentageAcum[i-1] +
conteo$porcentage[i]
conteo
#                    Var1 Freq porcentage procentageAcum
#1    Economic Policy & Debt  346 0.23944637      0.2394464
#6                    Health  249 0.17231834      0.4117647
#9    Private Sector & Trade  167 0.11557093      0.5273356
#11 Social Protection & Labor  158 0.10934256      0.6366782
#2                 Education  147 0.10173010      0.7384083
#3               Environment  140 0.09688581      0.8352941
#10            Public Sector  101 0.06989619      0.9051903
#4           Financial Sector   55 0.03806228      0.9432526
#7             Infrastructure   39 0.02698962      0.9702422
#8                    Poverty   28 0.01937716      0.9896194
#5                    Gender   15 0.01038062      1.0000000

#SE SELECCIONA el topico 1o: Economic Policy & Debt, con 346
Indicadores, 24%
#del total (1445 o 1442). El otro caso posible es el 2o, Health, con
249, 17,2%
#de los indicadores. Se podrian usar ambos topicos al tiempo para
el analisis
#porque suman 41,1%, y con Private Sector & Trade sube al
52.7%, y si a estos
#se le suman los de interes relacionadas con Poverty, sube a un
54,67%
#PERO esto demandaria mucho tiempo de calculo en el PC, se
requiere reducir
#mas el cubo de datos de alguna manera, pero ya tenemos 4
topicos de interes.


#--% NAs por Topico General
#Archivo de Entrada
indicatorsListNAs2porc<-readRDS("indicatorsListNAs2porc")

#---revisar por Topico General los %NAs de cada indicador
#resultado de Abril 27 de 2022:
Topic<-Health
indicatorsListNAs2porc[c(as.numeric(rownames(Topic)))]
#i.e. Health, no ordenados:
#SH.DTH.2024          SH.DTH.0509          SH.DTH.IMRT
SH.MMR.DTHS
#0.546992481          0.546992481          0.266432210
0.745694882
#SH.DTH.NMRT       SH.UHC.NOP1.TO       SH.UHC.NOP2.TO
SH.UHC.OOPC.10.TO
#0.363570216          0.996361872          0.996361872
0.996361872
#... hasta 249 indicadores que tiene Health

#---ordenar de menor NAs a mayor NAs...
zz<-indicatorsListNAs2porc[c(as.numeric(rownames(Topic)))]
vi<-order(zz, decreasing = TRUE)
vi
vii<-order(zz) #default: decreasing FALSE, mayor a menor
vii
length(zz)
zz[vii]
#i.e. Health ---> hay vbles con % <= 5%
#SP.POP.TOTL          SP.POP.GROW          EN.POP.DNST
AG.SRF.TOTL.K2
#0.006366723          0.022556391          0.034683483
0.083798205
#SP.POP.0004.FE.5Y    SP.POP.0004.MA.5Y
SP.POP.0014.TO.ZS    SP.POP.0014.FE.ZS
#0.090831919          0.090831919          0.090831919
0.090831919
#... y hasta 249 indicadores de Health...

#---Esto se puede repetir para los demas topicos, vectores,
cambiando Topic
#EconomicPolicyDebt
#Education
#Environment
#FinancialSector
#Gender
#Health
#Infrastructure
#Poverty
#PrivateSectorTrade
#PublicSector
#SocialProtectionLabor

#---Calculo de %NAs global por Topico General, aqui solo
miramos 4 de interes
Topic<-EconomicPolicyDebt
resumeSet2vbleTopic<-
summary(dataSet2[,c(as.numeric(rownames(Topic)))])
qtyVblsTopic<-dim(Topic)[1]
indicatorsListNAs2Topic<-resumeSet2vbleTopic[7,1:qtyVblsTopic]
namesIndicatorsListNAs2Topic<-names(indicatorsListNAs2Topic)
indicatorsListNAs2Topic<-substr(indicatorsListNAs2Topic,9,14)
indicatorsListNAs2Topic<-as.numeric(indicatorsListNAs2Topic)
names(indicatorsListNAs2Topic)<-namesIndicatorsListNAs2Topic
viii<-order(indicatorsListNAs2Topic, decreasing = TRUE)
ix<-order(indicatorsListNAs2Topic)
indicatorsListNAs2Topic[ix]
sum(indicatorsListNAs2Topic[ix]) #si se quiere ver orden
descendente [viii]
porcentajeNAsTopic<-
(sum(indicatorsListNAs2Topic[ix]))/(qtyVblsTopic*
                                    (dim(dataSet2)[1]))
porcentajeNAsTopic
#[1] 0.7049507 de Abril 27, 0.6973283 de Sep 1

Topic<-Health
resumeSet2vbleTopic<-
summary(dataSet2[,c(as.numeric(rownames(Topic)))])
qtyVblsTopic<-dim(Topic)[1]
indicatorsListNAs2Topic<-resumeSet2vbleTopic[7,1:qtyVblsTopic]
namesIndicatorsListNAs2Topic<-names(indicatorsListNAs2Topic)
indicatorsListNAs2Topic<-substr(indicatorsListNAs2Topic,9,14)
indicatorsListNAs2Topic<-as.numeric(indicatorsListNAs2Topic)
names(indicatorsListNAs2Topic)<-namesIndicatorsListNAs2Topic
viii<-order(indicatorsListNAs2Topic, decreasing = TRUE)
ix<-order(indicatorsListNAs2Topic)
indicatorsListNAs2Topic[ix]
sum(indicatorsListNAs2Topic[ix]) #si se quiere ver orden
descendente [viii]
porcentajeNAsTopic<-
(sum(indicatorsListNAs2Topic[ix]))/(qtyVblsTopic*
                                    (dim(dataSet2)[1]))
porcentajeNAsTopic
#[1] 0.590369 de Abril 27, 0.5781535 de Sep 1

Topic<-PrivateSectorTrade
resumeSet2vbleTopic<-
summary(dataSet2[,c(as.numeric(rownames(Topic)))])
qtyVblsTopic<-dim(Topic)[1]
indicatorsListNAs2Topic<-resumeSet2vbleTopic[7,1:qtyVblsTopic]
namesIndicatorsListNAs2Topic<-names(indicatorsListNAs2Topic)
indicatorsListNAs2Topic<-substr(indicatorsListNAs2Topic,9,14)
indicatorsListNAs2Topic<-as.numeric(indicatorsListNAs2Topic)
names(indicatorsListNAs2Topic)<-namesIndicatorsListNAs2Topic
viii<-order(indicatorsListNAs2Topic, decreasing = TRUE)
ix<-order(indicatorsListNAs2Topic)
indicatorsListNAs2Topic[ix]
sum(indicatorsListNAs2Topic[ix]) #si se quiere ver orden
descendente [viii]
porcentajeNAsTopic<-
(sum(indicatorsListNAs2Topic[ix]))/(qtyVblsTopic*
                                    (dim(dataSet2)[1]))
porcentajeNAsTopic
#[1] 0.665924 de Abril 27, 0.6674635 de Sep 1

Topic<-Poverty
```

```
resumeSet2vbleTopic<-
summary(dataSet2[,c(as.numeric(rownames(Topic)))])
qtyVblsTopic<-dim(Topic)[1]
indicatorsListNAs2Topic<-resumeSet2vbleTopic[7,1:qtyVblsTopic]
namesIndicatorsListNAs2Topic<-names(indicatorsListNAs2Topic)
indicatorsListNAs2Topic<-substr(indicatorsListNAs2Topic,9,14)
indicatorsListNAs2Topic<-as.numeric(indicatorsListNAs2Topic)
names(indicatorsListNAs2Topic)<-namesIndicatorsListNAs2Topic
viii<-order(indicatorsListNAs2Topic, decreasing = TRUE)
ix<-order(indicatorsListNAs2Topic)
indicatorsListNAs2Topic[ix]
sum(indicatorsListNAs2Topic[ix]) #si se quiere ver orden
descendente [viii]
porcentajeNAsTopic<-
(sum(indicatorsListNAs2Topic[ix]))/(qtyVblsTopic*
                                    (dim(dataSet2)[1]))
porcentajeNAsTopic
#[1] 0.8472528 de Abril 27, 0.8240359 de Sep 1

#Resumiendo, %NAs en dataSet2 por TOPICO, resultado de Abril
27 de 2022:
#EconomicPolicyDebt    [1] 0.7049507
#Education             [1] 0.6642371
#Environment           [1] 0.7589654
#FinancialSector       [1] 0.7761747
#Gender                [1] 0.6266068
#Health                [1] 0.590369
#Infrastructure        [1] 0.6184646
#Poverty               [1] 0.8472528
#PrivateSectorTrade    [1] 0.665924
#PublicSector          [1] 0.6463356
#SocialProtectionLabor [1] 0.6086668
#Queda que el TOPICO con menos NAs es el de Health 59%, le
sigue
#SocialProtectionLabor 60.8%, Infraestructura con 61.84%,
Gender 62.6%,
#PublicSector 64.63%, Education 66.4% y PrivateSectorTrade
66.6%, el resto >70%
#Se propone aqui entonces tomar los topicos donde la
combinacion %Cantidad de
#Indicadores y %NAs se mas alta, y a partir de estos crear el
dataSet para
#aplicar las tecnicas solo a los topicos generales seleccionados.
#Resta de todas maneras aun restringir a que paises se
aplicarian, y a que Years.



#------------------------------------------------------------------------
#Formar nuevo dataSet con lista de indicadores segun los topicos
seleccionados
#------------------------------------------------------------------------

#formar la lista que se va usar con los indicadoress de interes
segun TOPICO:

#Archivo de Entrada
setwd("~/R")
dataSet3Topics<-readRDS(file="dataSet3Topics")

#se escogen los de interes segun lo expuesto en apartado
anterior, resultado
#de la combinacion por topico de su %cantidad indicadores por
%NAs
TopicSel=c("Economic Policy & Debt","Health","Private Sector &
Trade","Poverty")

#Se forma la lista con los indicadores de los topicos generales
seleccionados
#[1] para el nombre completo del indicador, o [4] para ID indicador
corto
TopicSel<-dataSet3Topics[dataSet3Topics$Topic %in%
TopicSel,][4]
TopicSel
str(TopicSel) #dataFrame
dim(TopicSel) #[1] 790 1

#Se renombran columnas con ID indicadores nombres cortos:
dataSet2<-readRDS(file="dataSet2")
dataSet2a<-dataSet2
colnames(dataSet2a)<-
c(dataSet3Topics$IDVble,"Country","Year")

#se crea dataSet reducido con los indicadores de los topicos
generales
#seleccionados de interes
dataSet4IndXTopSel<-dataSet2a[,names(dataSet2a) %in%
TopicSel$IDVble]
#cuando se pone !colnames(dataSet2) %in% Topic$`Series
Code`, o sea la
#admiracion antecediendo, quiere indicar que es negacion. se
puede usar
#colnames o names...
dim(dataSet4IndXTopSel) #[1] 16492   790
names(dataSet4IndXTopSel)
#hay que agregar "Country" y "Year"
dataSet4IndXTopSel$Country<-dataSet2a$Country
dataSet4IndXTopSel$Year<-dataSet2a$Year
dim(dataSet4IndXTopSel) #[1] 16492   792
#fix(dataSet4IndXTopSel)


#Archivo de salida
saveRDS(object=dataSet4IndXTopSel, file=
"dataSet4IndXTopSel")

#El dataSet2 ahora llamado dataSet4IndXTopSel, pero solo con
los indicadores
#de los topicos generales escogidos. Resta ahora filtrar Year y
que Paises
#(LA, +World, + Europa, u otro mas... para comparar)



#------------------------------------------------------------------------
#Filtrar los paises de interes
#------------------------------------------------------------------------

#Archivo de Entrada
setwd("~/R")
library(readxl)
WDIEXCELDataCountries <- read_excel("WDIEXCEL.xlsx", sheet
= "Country")

#se deja solo Country Code, Short Name y el campo Region
Countries<-WDIEXCELDataCountries[,c(1,2,8)]

#Encontramos las economias que NO son paises y no pertenecen
a ninguna region
#geografica
#encuentro las posiciones, se crea lista identificando donde hay
NAs, no hay Region
ListCountriesNA<-which(is.na(Countries$Region))
#uso la lista para crear objeto con solo esos numeros de filas
identificados
CountriesNA<-Countries[ListCountriesNA,]
str(CountriesNA) #48 obs

#se factoriza para ver cuantos niveles: son 48 niveles de 48
observaciones
#son las economias que agrupan paises, sus indicadores son un
promedio de ellos
CountriesNA$`Short Name`<-as.factor(CountriesNA$`Short
Name`)
levels(CountriesNA$`Short Name`)
#ejemplos:
#[3] "Arab World"
#[14] "European Union"
#[15] "Fragile and conflict affected situations"
```

```
#[16] "Heavily indebted poor countries (HIPC)"
#[24] "Latin America & Caribbean"
#[28] "Low & middle income"
#[29] "Low income"
#[31] "Middle East & North Africa"
#[35] "North America"
#[36] "OECD members"
#[48] "World"
#fix(CountriesNA)

#Se factoriza ahora las regiones que se registran, un pais esta
geograficamente
#siempre clasificado en alguna region:
Countries$Region<-as.factor(Countries$Region)
str(Countries) #265 obs
Countries$Region
levels(Countries$Region)
#[1] "East Asia & Pacific"       "Europe & Central Asia"
#[3] "Latin America & Caribbean"  "Middle East & North Africa"
#[5] "North America"              "South Asia"
#[7] "Sub-Saharan Africa"


#crear vectores con las economias de interes
#Interesan todos los paises de la Region LatinoAmerica y el
Caribe,
#pero se puede agregar otra Region o cambiar de las 7 posibles
ListLA<-which(Countries$Region=="Latin America & Caribbean")
ListLA
#Interesa ademas las Economias con los siguientes nombres:
ListWorld<-which(Countries$`Short Name`=="World")
ListEU<-which(Countries$`Short Name`=="European Union")
ListLAC<-which(Countries$`Short Name`=="Latin America &
Caribbean")
ListMENA<-which(Countries$`Short Name`=="Middle East & North
Africa")
ListOECD<-which(Countries$`Short Name`=="OECD members")
ListLDC<-which(Countries$`Short Name`=="Least developed
countries: UN classification")
ListNA<-which(Countries$`Short Name`=="North America")
CountriesSel<-Countries[c(ListLA,ListWorld, ListEU, ListLAC,
ListMENA,
                          ListOECD,ListLDC,ListNA),]
CountriesSel
length(CountriesSel)
dim(CountriesSel) #[1] 49  3
str(CountriesSel)

#Archivo de salida:
saveRDS(object=CountriesSel, file= "CountriesSel")




#-------------------------------------------------------------------------
#Filtrar los Year de interes
#-------------------------------------------------------------------------

#Archivos de Entrada
setwd("~/R")
dataSet4IndXTopSel<-readRDS(file="dataSet4IndXTopSel")
CountriesSel<-readRDS(file="CountriesSel")

#ordenar ascendente por Year
dataSet4 <-
dataSet4IndXTopSel[order(dataSet4IndXTopSel$Year),]

#se aplica FILTRO de grupos de paises seleccionados...
library(dplyr)
dataSet5CYNA <- dataSet4 %>% dplyr::filter(dataSet4$Country
%in%
                                 CountriesSel$`Country Code`)

#Seleccion de los Year, se escogen los que tienen menores %NAs
como se habia
#concluido al principio desde dataSet1: 2015 a 2019, y sumado a
2020, 2021 tiene
#un muy alto %NAs.
dataSet5CYNA<-dataSet5CYNA %>%
dplyr::filter(dataSet5CYNA$Year
                                          %in% c("2015","2016",
                                          "2017","2018",
                                          "2019","2020"))
dim(dataSet5CYNA) #[1] 294 792

#dataSet5CYNA es la matriz con sus indicadores numericos y
NAs, y que
#mantiene sus factores. Esto interesa para otros analisis posibles
de 2 vias
#seleccionando solo un Year a la vez, i.e.permanova, bootstrap

#Archivo de salida: Matriz numerica (indicadores), CON factores y
con NAs
saveRDS(object=dataSet5CYNA, file= "dataSet5CYNA")


#Se eliminan los 2 factores Country y Year en dataSet5
#en la herramienta en R para el STATIS-dual la matriz de entrada
debe ser
#numerica y completa (sin NAs)
a<-dim(dataSet5CYNA)[2] #[1] 288 790 ojo!!! cada vez hay que
cambiar...
a
names(dataSet5CYNA[a-1])
names(dataSet5CYNA[a])
dataSet5<-dataSet5CYNA[-c((a-1),a)]
#eliminando Country y Year en dataSet5
dim(dataSet5) #[1] 294 790

#Ajustar los niveles de los factores, a pesar que disminuyan
cantidad de
#valores en el df, los levels se siguen manteniendo y hay que
eliminarlos
dataSet5CYNA$Country<-
factor(as.character(dataSet5CYNA$Country))
dataSet5CYNA$Year<-factor(as.character(dataSet5CYNA$Year))

#Archivo de salida: Matriz numerica (indicadores), SIN factores
con NAs
saveRDS(object=dataSet5, file= "dataSet5")



#-------------------------------------------------------------------------
#Eliminar indicadores por altos %NAs
#-------------------------------------------------------------------------

#Archivo de Entrada:
setwd("~/R")
#dataSet6fNAs<-readRDS(file="dataSet2") #matriz completa
inicial, no hay factores
#dataSet6fNAs<-readRDS(file="dataSet5") #matriz reducida
numerica resultado de la
#previa selecion pais y years
dataSet6fNAs<-readRDS(file="dataSet5CYNA") #dataSet5
incluyendo Pais y Year

#Eliminacion de indicadores que superen un %NAs
is.na(dataSet6fNAs) #True/False si estan NAs por cada vlr el
dataSet
colMeans(is.na(dataSet6fNAs))
porcentage <- .10 #se fija .0, .1, .6, .9... a criterio particular
columnas_a_borrar <-
which(colMeans(is.na(dataSet6fNAs))>porcentage)
dataSet6fNAs[,columnas_a_borrar] <- NULL
dim(dataSet6fNAs)
```

```
#Abril 27 de 2022:
#con dataSet2: 90%:[1]16492 1105 (mas Indicadores), 60%:16492
563, 10%: 16492 19
#(menos vbles)
#con dataSet5: 60%: 480 896, 30%: 480 431, 20%:480 242,
15%:480 76
#Septiembre 1 de 2022: 10% [1] 294  50

#Ajustar los niveles de los factores, a pesar que disminuyan
cantidad de
#valores en el df, los levels se siguen manteniendo y hay que
eliminarlos
dataSet6fNAs$Country<-
factor(as.character(dataSet6fNAs$Country))
dataSet6fNAs$Year<-factor(as.character(dataSet6fNAs$Year))

#Archivo de salida
saveRDS(object=dataSet6fNAs, file= "dataSet6fNAs")




#-------------------------------------------------------------------------
#Imputar datos perdidos
#-------------------------------------------------------------------------

#Archivo de Entrada:
setwd("~/R")
dataSet6fNAs<-readRDS(file="dataSet6fNAs")


#Normalizar/estandarizar si o no?, es necesario?
#dataSet6<-scale(dataSet6)


#Otra forma de Visualizacion por consola de un df
library(kableExtra)
kable(dataSet6fNAs[1:4,1:7],"markdown")


#Contar el total de NAs
dim(dataSet6fNAs) #recordar que esto ya es un df reducido en
indicadores
sum(is.na(dataSet6fNAs)) #NAs total
colSums(is.na(dataSet6fNAs)) #NAs por columna

base1 <- na.omit(dataSet6fNAs) #Esto Omite pero las FILAS con
NAs, y ya no
#esta considerando los mismos paises o economias para todos
los Year.
#la idea es mantener todo el tiempo las mismas economias/paises
para todos los
#Year... este na.omit no funcionaria para nuestros intereses
dim(dataSet6fNAs) #[1] 294  50
dim(base1) #[1] 208  50

#Visualizacion de proporcion de datos faltantes por columna
indicador
library(VIM)
aggr(dataSet6fNAs,numbers=T,sortVar=T, combined=F)
aggr(dataSet6fNAs,col=c('navyblue','green'),numbers=T,sortVars=
T,
    labels=names(dataSet6fNAs),cex.axis=.7,gap=3,
    ylab=c("Histogram of missing data","Pattern"))
#cuando hay muchas columnas se pierde esta visibilizacion:
#not enough vertical space to display frequencies (too many
combinations)

#comparar par indicadores,i.e. Pr421 con He962
marginplot(dataSet6fNAs[,c("Pr421","He962")])
marginplot(dataSet6fNAs[,c("Pr421","He965")])
marginplot(dataSet6fNAs[,c("Pr421","Ec696")])

#--Imputacion por la MEDIA: Tecnica 1, mice
columns <- c(colnames(dataSet6fNAs))
Lc<-length(columns)
columns <- columns[-c((Lc-1),Lc)]
#crear un vector con los nombres de columnas de solo los
indicadores que se van
#a imputar, para el caso serian las que quedaron en
dataSet6fNAs, menos las dos
#ultimas Year y Country que son factores no numericas.
columns

library(mice)
summary(dataSet6fNAs)
md.pattern(dataSet6fNAs) #ver el patron de los NAs: i.e.208
muestras totales
#9 de las cuales en la 1a columna/indicador Ec33 tiene 9 NAs,
siguen 9 en Ec35..
#metodos: mean=media, pmm=predictive mean, cart=classification
& Regression trees
imputed_data <- mice(dataSet6fNAs[,names(dataSet6fNAs) %in%
columns],m=1,
                     maxit=1,method="mean",seed=2022,print=F)
#scatter plots entre indicadores a partir de sus vlrs imputados
xyplot(imputed_data,Pr421~He965+He962+Ec696,pch=18,cex=1)
xyplot(imputed_data,Pr421~He965+He962,pch=18,cex=1)
xyplot(imputed_data,Pr421~He962,pch=18,cex=1)

#otros plots...
#plot(imputed_data) #revisar funcionamiento
#stripplot(imputed_data,pch=20,cex=1.2) #revisar funcionamiento
#densityplot(imputed_data) ##revisar funcionamiento
#Error in density.default(x = c(NA_real_, NA_real_, NA_real_,
NA_real_,  :
#need at least 2 points to select a bandwidth automatically

complete.data <- mice::complete(imputed_data)
sum(is.na(complete.data)) #permanecen NAs... por que?.
#metodos usados: mean, pmm y cart. metodo "norm.predict" y
"norm.nob" dieron
#error al aplicar
#Error in solve.default(xtx + diag(pen)) :
#system is computationally singular: reciprocal condition number =
4.93296e-31
#revisar que parametro de mice hay que cambiar para completar
NAs al 100%
colSums(is.na(complete.data))
#dataSet7<-complete.data #Aun NAs... no sirve para usar en hxta
Statis-dual

#dataSet7<-na.omit(complete.data) Esto Omite pero las FILAS
con NAs, y ya no
#esta considerando los mismos paises o economias para todos
los Year.
#la idea es mantener todo el tiempo las mismas economias/paises
para todos los
#Year... este na.omit no funcionaria para nuestros intereses ya
que desordena
#la cantidad de filas para el statis y causa otro error, pasa i.e. de
288 a 269...

#mice tambien permite el metodo de IMPUTACION MULTIPLE
#que resuelve los errores estándar "demasiado pequeños". PERO
NO FUNCIONA...
#revisar funcionamiento
base<-dataSet6fNAs
imputed_data <- mice(base[,names(base) %in%
columns],seed=2022,print=F,m=30)
#Error in solve.default(xtx + diag(pen)) :
#  system is computationally singular: reciprocal condition number
= 9.50446e-31
complete.data<- mice::complete(imputed_data)


#--Imputacion por la MEDIA: Tecnica 2, por MULTIPLES
COLUMNAS
```

```
#crear vectores con posiciones de NAs por cada indicador de
interes, aqui 4
Ec696_miss_ind<-is.na(dataSet6$Ec696)
Ec696_miss_ind
dataSet6$Ec696
He965_miss_ind<-is.na(dataSet6$He965)
He962_miss_ind<-is.na(dataSet6$He962)
Pr421_miss_ind<-is.na(dataSet6$Pr421)

#se crea df con esos vectores de vbles de datos perdidos
dataTestingImput<-
data.frame(dataSet6$Ec696,dataSet6$He965,dataSet6$He962,
                        dataSet6$Pr421)
head(dataTestingImput)

#Imputacion de la MEDIA por cada vector anterior, para comparar
la densidad
#dataTestingImput$dataSet6.Ec696[is.na(dataTestingImput$dataS
et6.Ec696)]<-mean(dataTestingImput$dataSet6.Ec696,na.rm=T)
#dataTestingImput$dataSet6.He965[is.na(dataTestingImput$data
Set6.He965)]<-mean(dataTestingImput$dataSet6.He965,na.rm=T)
#dataTestingImput$dataSet6.He962[is.na(dataTestingImput$data
Set6.He962)]<-mean(dataTestingImput$dataSet6.He962,na.rm=T)
#dataTestingImput$dataSet6.Pr421[is.na(dataTestingImput$dataS
et6.Pr421)]<-mean(dataTestingImput$dataSet6.Pr421,na.rm=T)
#na.rm=T indica que dentro de mean() que los vlrs no deberian ser
usados para
#calcular la media (na.rm=F seria imposible y lleva a un error)

#imputacion de la MEDIA sobre todo el dataTestinImput a la vez y
no uno a uno
#como se hizo antes
for(i in 1:ncol(dataTestingImput)) {
  dataTestingImput[,i][is.na(dataTestingImput[,i])]<-
mean(dataTestingImput[,i],
                                        na.rm=T)
}
head(dataTestingImput)

#grafica de la evaluacion de vlrs imputados
#en teoria (autores en la web) imputar con la MEDIA deberia NO
es la mejor
#opcion, en general esto hace que cambie mucho la densidad por
cada indicador
#pero para nuestro caso, WDI con 42 indicadores, revisando 4
indicadores, no
#hubo mayor variacion... pero esto NO ES LO NORMAL...

#densidades de los datos observados
#plot(density(dataTestingImput$dataSet6.Ec696[Ec696_miss_ind=
=FALSE]),
#     xlim = c(- 4, 4),
#     ylim = c(0, 0.9),
#     lwd = 2,
#    main = "Density Pre and Post Mean Imputation",
#     xlab = "Ec696")
#no sale el grafico de densidad... son los limites de xlim ylim...
quitarlos
par(mfrow=c(1,1))
plot(density(dataTestingImput$dataSet6.Pr421[Pr421_miss_ind==
FALSE]),
     lwd = 2,
     main = "Density Pre and Post Mean Imputation",
     xlab = "Pr421", col="brown")

#grafica: densidad de los datos observados e imputados
points(density(dataTestingImput$dataSet6.Pr421),
       lwd = 2,
       type = "l",
       col = "gray")
legend("topright",
       c("Before Imputation", "After Imputation"),
       lty = 1,
       lwd = 2,
       col = c("brown", "gray"))

#a continucion las otras graficas de densidad:
#plot(density(dataTestingImput$dataSet6.Ec696[Ec696_miss_ind=
=FALSE]),
#     lwd = 2,
#     main = "Density Pre and Post Mean Imputation",
#     xlab = "Ec696", col="black")
#points(density(dataTestingImput$dataSet6.Ec696),
#       lwd = 2,
#       type = "l",
#       col = "red")
#legend("topleft",
#       c("Before Imputation", "After Imputation"),
#       lty = 1,
#       lwd = 2,
#       col = c("black", "red"))

#plot(density(dataTestingImput$dataSet6.He965[He965_miss_ind
==FALSE]),
#     lwd = 2,
#     main = "Density Pre and Post Mean Imputation",
#     xlab = "He965",col="blue")
#points(density(dataTestingImput$dataSet6.He965),
#       lwd = 2,
#       type = "l",
#       col = "yellow")
#legend("topleft",
#       c("Before Imputation", "After Imputation"),
#       lty = 1,
#       lwd = 2,
#       col = c("blue", "yellow"))

#plot(density(dataTestingImput$dataSet6.He962[He962_miss_ind
==FALSE]),
#     lwd = 2,
#     main = "Density Pre and Post Mean Imputation",
#     xlab = "He962", col="green")
#points(density(dataTestingImput$dataSet6.He962),
#       lwd = 2,
#       type = "l",
#       col = "orange")
#legend("topleft",
#       c("Before Imputation", "After Imputation"),
#       lty = 1,
#       lwd = 2,
#       col = c("green", "orange"))

#plot resumen de los 4:
par(mfrow=c(2,2))
plot(density(dataTestingImput$dataSet6.Pr421,na.rm =
T),col=2,main="Pr421")
lines(density(complete.data$Pr421),col=3)
plot(density(dataTestingImput$dataSet6.He962,na.rm =
T),col=2,main="He962")
lines(density(complete.data$He962),col=3)
plot(density(dataTestingImput$dataSet6.Ec696,na.rm =
T),col=2,main="Ec696")
lines(density(complete.data$Ec696),col=3)
plot(density(dataTestingImput$dataSet6.He965,na.rm =
T),col=2,main="He965")
lines(density(complete.data$He965),col=3)
par(mfrow=c(1,1))

#Estadistica descriptiva para Ec696, He965, He962 y Pr421
#pre-imputacion
round(summary(dataTestingImput$dataSet6.Ec696[Ec696_miss_i
nd == FALSE]), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#57.60   70.69   78.54   76.57   82.88   88.10
#pos-imputacion:
round(summary(dataTestingImput$dataSet6.Ec696), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#57.60   71.19   77.94   76.57   81.57   88.10

#media es igual antes y despues de la imputacion
```

```
#el problema se ve comparando el 1er y 3er cuartil de cada
indicador pre y
#pos imputacion
#1er cuartil pre y pos: 70.64 vs. 71.19.
#3er cuartil pre y pos: 82.88 vs. 81.57.
#Ambos cuartiles se desplazan en teoria hacia cero, después de
sustituir los
#datos faltantes por la media, cuartiles están muy sesgados.
#Hay que tener presente que como los datos no son random, las
medias estimadas
#si estan sesgadas.

#pre-imputacion:
#round(summary(dataTestingImput$dataSet6.He965[He965_miss
_ind == FALSE]), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#12.56   56.76   93.85   77.57   98.98  100.00
#pos-imputacion:
#round(summary(dataTestingImput$dataSet6.He965), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#12.56   64.25   91.49   77.57   98.84  100.00
#pre-imputacion:
#round(summary(dataTestingImput$dataSet6.He962[He962_miss
_ind == FALSE]), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#45.48   85.50   96.75   89.41   99.93  100.00
#pos-imputacion:
#round(summary(dataTestingImput$dataSet6.He962), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#45.48   87.11   96.49   89.41   99.90  100.00
#pre-imputacion:
#round(summary(dataTestingImput$dataSet6.Pr421[Pr421_miss_i
nd == FALSE]), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#0.84    1.48    1.87    2.50    3.05    7.17
#pos-imputacion:
#round(summary(dataTestingImput$dataSet6.Pr421), 2)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#0.84    1.54    1.97    2.50    2.98    7.17


#AHORA SI!!! imputacion de la MEDIA sobre todo el dataSet de
entrada
base<-dataSet6fNAs
for(i in 1:ncol(base)) {
  base[,i][is.na(base[,i])]<-mean(base[,i],na.rm=TRUE)
}
head(base)
sum(is.na(base)) #[1] 0
colSums(is.na(base))
rownames(base)
names(base)


#grafico resumen de densidades de 4 indicadores pre y pos
imputacion
par(mfrow=c(2,2))
plot(density(dataSet6fNAs$Pr421,na.rm = T),col=1,main="Pr421")
lines(density(base$Pr421),col=2)
plot(density(dataSet6fNAs$He962,na.rm =
T),col=3,main="He962")
lines(density(base$He962),col=4)
plot(density(dataSet6fNAs$Ec696,na.rm = T),col=5,main="Ec696")
lines(density(base$Ec696),col=6)
plot(density(dataSet6fNAs$He965,na.rm =
T),col=7,main="He965")
lines(density(base$He965),col=8)
par(mfrow=c(1,1))

#Ajustar los niveles de los factores, a pesar que disminuyan
cantidad de
#valores en el df, los levels se siguen manteniendo y hay que
eliminarlos
base$Country<-factor(as.character(base$Country))
base$Year<-factor(as.character(base$Year))

#Archivo de salida: Este es el archivo de entrada en STATIS-dual
finalmente
dataSet7CYMEAN<-base
saveRDS(object=dataSet7CYMEAN, file= "dataSet7CYMEAN")


#--OTROS 1: imputacion LOCF (Last Observation Carried
Forward, datos longitudinales)
library(tidyr)
imputar <- tidyr::fill(dataSet6fNAs, all_of(columns))
dim(imputar)
sum(is.na(imputar))
sum(is.na(dataSet6fNAs))

#grafico resumen de densidades de 4 indicadores pre y pos
imputacion
par(mfrow=c(2,2))
plot(density(dataSet6fNAs$Pr421,na.rm = T),col=1,main="Pr421")
lines(density(imputar$Pr421),col=2)
plot(density(dataSet6fNAs$He962,na.rm =
T),col=3,main="He962")
lines(density(imputar$He962),col=4)
plot(density(dataSet6fNAs$Ec696,na.rm = T),col=5,main="Ec696")
lines(density(imputar$Ec696),col=6)
plot(density(dataSet6fNAs$He965,na.rm =
T),col=7,main="He965")
lines(density(imputar$He965),col=8)
par(mfrow=c(1,1))

#Archivo de salida: Este es el archivo de entrada en STATIS-dual
finalmente
dataSet7CYLOCF<-imputar
saveRDS(object=dataSet7CYLOCF, file= "dataSet7CYLOCF")


#--OTROS 2: imputacion ALEATORIA
#se define la funcion aleatoria
rand.imput <-function(x){
  missing <- (is.na(x)) #vector booleano
  n.missing <- sum(missing)#Numero de NA's
  x.obs <- x[!missing]#Datos no NA
  imputed <- x
  imputed[missing] <- sample(x.obs,n.missing,replace = T)
  #Se extrae una muestra aleatoria conocida y se remplazan estos
en los NA
  return(imputed)}

#hay que aplicar la funcion a cada indicador, generar el codigo
para un FOR...
#y aplicar a todos los indicadores
complete.dataPr421 <- rand.imput(dataSet6fNAs$Pr421)
sum(is.na(complete.dataPr421)) #[1] 0
complete.dataHe962 <- rand.imput(dataSet6fNAs$He962)
complete.dataEc696 <- rand.imput(dataSet6fNAs$Ec696)
complete.dataHe965 <- rand.imput(dataSet6fNAs$He965)

#grafico resumen de densidades de 4 indicadores pre y pos
imputacion
par(mfrow=c(2,2))
plot(density(complete.dataPr421,na.rm = T),col=1,main="Pr421")
lines(density(dataSet6fNAs$Pr421,na.rm=T),col=2)
plot(density(complete.dataHe962,na.rm = T),col=3,main="He962")
lines(density(dataSet6fNAs$He962,na.rm=T),col=4)
plot(density(complete.dataEc696,na.rm = T),col=5,main="Ec696")
lines(density(dataSet6fNAs$Ec696,na.rm=T),col=6)
plot(density(complete.dataHe965,na.rm = T),col=7,main="He965")
lines(density(dataSet6fNAs$He965,na.rm=T),col=8)
par(mfrow=c(1,1))
```

```
#########################################################
#########################
#Analisis de 3 vias
#########################################################
#########################

#-------------------------------------------------------------------------
# Sparse STATIS dual y comparacion con STATIS-dual (escuela
SALMANTINA)
#-------------------------------------------------------------------------

#install.packages("SparseSTATISdual")
#package 'SparseSTATISdual' is not available for this version of R

#RECORDAR correr previamente los 6 SCRIPTS no
empaquetados de la libreria:
#ElasticNetStatis
#plot_Compromise
#plot_Interstructure
#plot_Intrastructure
#STATIS
#STATISdual

#Llamando directamente Scripts desde github no funciona: Error in
source
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/Elastic_net_Statis_d.R")
#Error in
source("https://github.com/CCRM07/SparseSTATISdual/blob/main
/Elastic_net_Statis_d.R") :
#
https://github.com/CCRM07/SparseSTATISdual/blob/main/Elastic_
net_Statis_d.R:8:1: unexpected '<'
#7:
#  8: <
#  ^
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/plot_Compromise_d.R")
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/plot_Interstructure_d.R")
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/plot_Intrastructure_d.R")
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/Statis.R")
#source("https://github.com/CCRM07/SparseSTATISdual/blob/mai
n/Statis-dual.R")

#Carga de las source de los Scripts desarrollados por Rodriguez-
Martinez(2020)
setwd("~/R")
source("4 7 03 ElasticNetStatis.R")
source("4 7 04 plot_Compromise.R")
source("4 7 05 plot_Interstructure.R")
source("4 7 06 plot_Intrastructure.R")
source("4 7 08 STATIS.R")
source("4 7 09 STATISdual.R")

#Archivo de Entrada
setwd("~/R")
#BASES<-readRDS(file="dataSet5")
#sum(is.na(dataSet5))
#BASES<-readRDS(file="dataSet7CYLOCF")
#sum(is.na(dataSet7CYLOCF))
BASES<-readRDS(file="dataSet7CYMEAN")
sum(is.na(dataSet7CYMEAN))


#verificar la base a usar
names(BASES) #bases debe ser totalmente numerica y
COMPLETA (sin NAs)
Yrmin<-
as.integer(as.character(BASES$Year[BiocGenerics::which.min(BA
SES$Year)]))
Yrmax<-
as.integer(as.character(BASES$Year[BiocGenerics::which.max(B
ASES$Year)]))
Ctrmax<-dim(BASES)[1]/(Yrmax-Yrmin+1)
BASES$Country<-NULL
BASES$Year<-NULL
sum(is.na(BASES))

#Nombrar las matrices (k-tablas) que corresponde a cada Year
mnames<-c()
for (i in Yrmin:Yrmax) #se selecciono pasos atras los 6 Years:
2015 a 2020
  mnames<-c(mnames,i)
mnames
length(mnames)
mnames<-as.character(mnames)
mnames

#Etiquetas para Trayectorias
vnamest<-c(names(BASES))
vnamest

#Este paso es para separar las matrices:
#Indicar la ubicacion de la ultima fila, de cada matriz en la base
datos
M<-c()
for (i in 1:length(mnames)) #ya no es 62 anos, ni 10 anos, ahora
solo 6
  M<-c(M,i*Ctrmax) #ya no son 266 paises, ahora solo 49: LA +
otros clasif/SubReg
#ojo!!! puede que al eliminar se reduzcan los paises, y ya no sean
los 48...
#y no irian en mulitplos de 48 (i.e. 6x48 =288, y solo aparecen 269
filas...
#REVISAR como se forma dataSet7 cuando se eliminan los NAs.
M
length(M)
#se tiene un solo archivo, donde se tienen todas
#las matrices, una debajo de la otra, hay que indicarle donde
termina cada
#una, i.e. la matriz 1 termina en la fila 10, y asi sucesivamente

# Para verificar la cantidad de matrices
LM<-length(M) #cuantas matrices hay
LM

#Separacion de las matrices
M2<-c(0,M)
M2

Y<-vector("list",LM)
for (i in 1:LM){Y[[i]]<-BASES[(M2[i]+1):(M2[i+1]),]}
Y
str(Y) #Lista de 6

#Asignarle nombre a las matrices
names(Y) <- c(mnames)
names(Y)
str(Y)


#STATIS-dual (escuela SALMANTINA)
#recordar que es dual porque la estructura de datos de entrada ya
esta dada
#para dar enfasis a las variables (indicadores)

sts <- Statis(Y, transform = 'scale')
#class(sts) #list
#str(sts)
names(sts)
#[1] "X"              "tab_names"        "X_Transformed"
"Scalar_Products"
#[5] "Cosine"         "Interstructure"   "Weight"
"Compromise"
```

```
#[9] "Quality"          "Scores"          "Projection_Matrix"
"Coordinates"

sts$Interstructure
#          Dim1        Dim2         Dim3        Dim4        Dim5
Dim6
#2015 -0.9941951  0.052030213  0.0903552813 -0.016621886
0.009713972  0.018281898
#2016 -0.9965125  0.072589700  0.0002930972 -0.006005808
0.001326593 -0.040690052
#2017 -0.9948711  0.074707589 -0.0493894651  0.016445358 -
0.039915839  0.018635071
#2018 -0.9972297 -0.005396044 -0.0433663599  0.025692659
0.053691630  0.008955402
#2019 -0.9942062 -0.081296860 -0.0314556108 -0.062649018 -
0.004968022  0.002409273
#2020 -0.9918378 -0.113105390  0.0338088369  0.043166136 -
0.020035622 -0.007554693

sts$Compromise
#           Ec33       Ec35       Ec39       Ec42       Ec123
Pr397
#Ec33   48.0466771  -0.2159830   2.3321563   2.58746744  -
4.99164608   0.1768127
#Ec35   -0.2159830  48.0466771  -3.4067271  -2.80017701 -
26.52558169  10.2687123
#Ec39    2.3321563  -3.4067271  48.0466771  -4.09172240
12.95360912  -3.1278486
#...

stsScoresDF193ind<-data.frame(sts$Scores)
stsScoresDF193ind
#write.csv(stsScoresDF193ind, "stsDFScores42indV2.csv") #V2
porque ya hay unos
#resultados guardados usando la base de Abril 27 de 2022.
Cualquier nuevo csv
#sera llamado V2
#write.csv(stsScoresDF193ind, "stsDFScores193indV2.csv") #42 o
193 indicadores
#tener presente antes de salvar y estar claro que en el csv se
calcula la
#la varianza y la acumulada

sts$Quality #pero este quality no es el coseno^2...
#ver adelante en Ade4 el resultado con statis1$cos2

sts$Weight #pesos de cada tabla (momento / Year)... todos los
Year (tablas)
#aparecen con pesos iguales. Esta info de pesos sin la del cos2
no da para
#lograr la typological value (grafico) que tiene en la hxta Ade4
statis1
#[1] 0.1664551 0.1665346 0.1676704 0.1672639 0.1668055
0.1669810

sts$Scores #puntuaciones de c/indicador en las dimensiones (48,
tot indicadores)

sts$Cosine #esto equivale a los coef.correlacion vectorial
(Ade4:statis1$RV)
#      2015  2016  2017  2018  2019  2020
#2015 1.0000 0.9961 0.9774 0.9765 0.9713 0.9720
#2016 0.9961 1.0000 0.9807 0.9751 0.9709 0.9733
#2017 0.9774 0.9807 1.0000 0.9951 0.9914 0.9915
#2018 0.9765 0.9751 0.9951 1.0000 0.9860 0.9890
#2019 0.9713 0.9709 0.9914 0.9860 1.0000 0.9859
#2020 0.9720 0.9733 0.9915 0.9890 0.9859 1.0000
#ver adelante en Ade4 el resultado con statis1$RV, y comparar de
ser necesario
#ver adelante en Ade4 el resultado con statis1$RV.eig

sts$X #matriz original
sts$X_Transformed #matriz transformada

##################################################
#########################
####          Sparse STATIS-dual  (escuela SALMANTINA)
###
##################################################
#########################

EN_sts <- Elastic_net_Statis(Y, Transform.Data = 'scale',
                             Alpha = 0.001,
                             Lambda = 0.001)
#class(EN_sts)
#str(EN_sts)
names(EN_sts)

EN_sts$Quality

EN_sts$Weight

EN_sts$Scores #se notan los "0", cargas nulas de muchos
indicadores en las
#dimensiones (48, tot indicadores)
#dimensiones mas alla de la 16 las puntuaciones de los
indicadores es "0"


EN_stsScoresDF42ind<-data.frame(EN_sts$Scores)
EN_stsScoresDF42ind
#write.csv(EN_stsScoresDF42ind, "EN_stsDFScores42indV2.csv")
#V2 porque ya hay unos
#resultados guardados usando la base de Abril 27 de 2022.
Cualquier nuevo csv
#sera llamado V2
#write.csv(EN_stsScoresDF42ind,
"EN_stsDFScores193indV2.csv") 42 o 193 indicadores
#tener presente antes de salvar y estar claro que en el csv se
calcula la
#la varianza y la acumulada

#Archivos .CSV: leer desde EXCEL y calcular la VARIANZA, tanto
para STATIS dual
#como para el Sparse STATIS dual, a partir de los scores
recordando que
#varianza sale de: 1) scores para cada Componente Principal CP
que se elevan al
#cuadrado, a_ij^2, 2)luego se suman por c/CP, llamado esto el SS
loadings, luego
#se totaliza todos los SS loadings sumandolos, y 3)luego se
calculan los %

#De los archivos .CSV, ademas revisados y comparados en
EXCEL para los dos
#escenarios que se han planteado (42 indicadores y 193
indicadores, resultantes
#de la WDI descargada de Abril 27 de 2022 despues de aplicar el
%NAs para
#eliminacion de indicadores que lo superasen, al 10% dio 42
indicadores y al
#20% 193).

#De la comparacion hecha desde los .CSV, revisado los
indicadores con mayores
#cargas en cada dimension, se tiene que:
#para el caso de 42 ind se logran estos:
EN_stsIndMayorCarga42ind<-
c("Ec123","Ec562","Ec565","Ec600","Ec696","He1052",

"He962","Pr1358","Pr1418","Pr403","Pr421","Pr429",
                    "Pr448","Pr465","Pr530")

#para el caso de 193 ind se logran estos:
EN_stsIndMayorCarga193ind<-
c("Ec123","Ec696","He999","He1000","He1001","He1003",
```

```
"He1005","He1007","He1008","He1009","He1021",
                    "He1022","He1023","He1024","He1290")

#Identificados los indicadores de mayor carga, interesantes para este estudio,
#se busca a que indicadores corresponden. Se averigua desde WDI nombres /
#contenido de algunos que se ven en el biplot resultante, para revisar y concluir

#formar la lista que se va usar con las vbles de interes segun TOPICO:
#Archivo de Entrada:
setwd("~/R")
dataSet3Topics<-readRDS(file="dataSet3Topics")

IndInteres42<-dataSet3Topics[dataSet3Topics$IDVble %in% EN_stsIndMayorCarga42ind,]
IndInteres42$`Series Code`
IndInteres42$Topic
#buscar uno a uno la explicacion de cada indicador en WDI sheet
IndInteres193<-dataSet3Topics[dataSet3Topics$IDVble %in% EN_stsIndMayorCarga193ind,]
IndInteres193$`Series Code`
IndInteres193$Topic

#Comparacion grafica de STATIS-dual y Sparse STATIS-dual:
#Graficar interestructura
Plot_Interstructure(sts, color = "red") + ggtitle(label = "Inter-estructura")
Plot_Interstructure(EN_sts, color = "red") + ggtitle(label = "Inter-estructura Sparse")

#Graficar compromiso
plot_Compromise(sts) + ggtitle(label = "Compromiso")
plot_Compromise(EN_sts) + ggtitle(label = "Compromiso-Sparse")

#Graficar intraestructura
Plot_Intrastructure(sts) + ggtitle(label = "Intra-estructura")
Plot_Intrastructure(EN_sts) + ggtitle(label = "Intra-estructura Sparse")

# Graficar Trayectorias (no hay source para esta hxta, pendiente codigo)
Plot_Trayectorias(sts) + ggtitle(label = "Trayectorias")
#Error in Plot_Trayectorias(sts) : could not find function "Plot_Trayectorias"
Plot_Trayectorias(EN_sts) + ggtitle(label = "Trayectorias Sparse")
#Error in Plot_Trayectorias(EN_sts) : could not find function "Plot_Trayectorias"




#------------------------------------------------------------------------------
# STATIS-dual desde Ade4 (escuela francesa)
#------------------------------------------------------------------------------

#Archivo de Entrada
setwd("~/R")
#datosAde4<-readRDS(file="dataSet5")
#datosAde4<-readRDS(file="dataSet7CYLOCF")
datosAde4<-readRDS(file="dataSet7CYMEAN")
sum(is.na(datosAde4))

#quitar factores que trae, esto lo que hace es eliminar los levels de mas
#que se arrastran, cada vez que se ejecuta el codigo para generar los Archivos
#de entrada
datosAde4$Country<-as.character(datosAde4$Country)
datosAde4$Year<-as.character(datosAde4$Year)

#defino las k-tablas de entrada a la hxta Ade4, y la dejo solo numerica
datosKtab<-datosAde4
datosKtab$Country<-NULL
datosKtab$Year<-NULL
datosAde4$Country<-factor(datosAde4$Country)
levels(datosAde4$Country)
datosAde4$Year<-factor(datosAde4$Year)
levels(datosAde4$Year) #ahora los levels son los que se requieren de acuerdo
#a la seleccion que se haga previamente de Year y de Country


library(ade4)

#objeto aplicando componentes principales
plot(withinpca(datosKtab, datosAde4$Year, scann = FALSE))
plot(withinpca(datosKtab, datosAde4$Country, scann = FALSE))
withinpca(datosKtab, datosAde4$Country, scann = FALSE)
withinpca(datosKtab, datosAde4$Year, scann = FALSE)

#caso factor YEAR sin scaling
kta1 <- ktab.within(withinpca(datosKtab, datosAde4$Year, scann = FALSE))
#NOTA: hay casos donde no se procede a estandarizar los datos (tienen una misma
#unidad escala de medición). Si se quisiera es agregando scaling="total" o
#"partial". Es elección del usuario si hace un preprocesamiento de datos o no.

#esto es solo informacion para tener presente:
withinpca(datosKtab, datosAde4$Year, scaling="total", scann = FALSE)
#el argumento de scaling viene dentro del comando Ade4::withinpca
#This functions implements the 'Bouroche' standardization. In a first step,
#the original variables are standardized (centred and normed). Then, a second
#transformation is applied according to the value of the scaling argument.
#For "partial", variables are standardized in each sub-table (corresponding to
#each level of the factor). Hence, variables have null mean and unit variance
#in each sub-table. For "total", variables are centred in each sub-table and
#then normed globally. Hence, variables have a null mean in each sub-table and
#a global variance equal to one.

#caso factor YEAR y con scaling partial
kta1 <- ktab.within(withinpca(datosKtab, datosAde4$Year, scaling="partial",
                              scann = FALSE))
#si estandarizara partial, por cada sub-tabla de Year, NO cambian las
#posiciones de las vbles indicadores en el compromiso...
#estandarizar o no, no tiene mucho efecto, o si?...


#caso factor COUNTRY sin scaling
#kta1 <- ktab.within(withinpca(datosKtab, datosAde4$Country, scann = FALSE))
#caso factor COUNTRY y con scaling partial
#kta1 <- ktab.within(withinpca(datosKtab, datosAde4$Country, scaling="partial",
#                              scann = FALSE))


kta1

statis1 <- statis(kta1, scann = FALSE) #Realiza el análisis STATIS
#class(statis1)
```

```
#str(statis1)
statis1

#Grafico de lntEREstructura
plot(statis1)

#Valores
statis1$tab.names
statis1$cos2
statis1$RV
statis1$RV.eig
str(statis1)

#Calculo de la Varianza
k=dim(statis1$RV.coo)
k=k[1]
k
eigt=0
for (i in 1:k)
  eigt=statis1$RV.eig[i]+eigt
eigt
iner=c(rep(0),k)
for (i in 1:k)
  iner[i]=statis1$RV.eig[i]/eigt*100
iner
for (i in 2:k)
  iner[i]=iner[i]+iner[i-1]
iner
statis1$C.eig
str(statis1)
k=statis1$C.rank
k
eigt=0
for (i in 1:k)
  eigt=statis1$C.eig[i]+eigt
eigt
iner=c(rep(0),k)
for (i in 1:k)
  iner[i]=statis1$C.eig[i]/eigt*100
iner
for (i in 2:k)
  iner[i]=iner[i]+iner[i-1]
iner

#Grafico de la lntRAEstructura
kplot(statis1)



#-------------------------------------------------------------------------
#COMPLEMENTO 1: PTA Partial Triadic Analysis, X-STATIS,
desde Ade4
#-------------------------------------------------------------------------

#Archivo de Entrada
setwd("~/R")
#datosAde4<-readRDS(file="dataSet5")
#datosAde4<-readRDS(file="dataSet7CYLOCF")
datosAde4<-readRDS(file="dataSet7CYMEAN")
sum(is.na(datosAde4))

datosAde4$Country<-as.character(datosAde4$Country)
datosAde4$Year<-as.character(datosAde4$Year)

#defino las k-tablas de entrada a la hxta Ade4, y la dejo solo
numerica
datosKtab<-datosAde4
datosKtab$Country<-NULL
datosKtab$Year<-NULL
datosAde4$Country<-factor(datosAde4$Country)
levels(datosAde4$Country)
datosAde4$Year<-factor(datosAde4$Year)
levels(datosAde4$Year)

library(ade4)
#objeto aplicando componentes principales
plot(withinpca(datosKtab, datosAde4$Year, scann = FALSE))
plot(withinpca(datosKtab, datosAde4$Country, scann = FALSE))
withinpca(datosKtab, datosAde4$Country, scann = FALSE)
withinpca(datosKtab, datosAde4$Year, scann = FALSE)

wt1<-withinpca(datosKtab, datosAde4$Year, scaling="partial",
scann = FALSE)

#install.packages("taRifx") #no se puede instalar
#library(taRifx)
#help(taRifx) #este contiene la funcion remove.factors(df)...

#Creo el vector con los nombres de paises / economias
ctr<-length(levels(datosAde4$Country))
vectornames<-as.character(datosAde4$Country[1:ctr]) #quito que
sea factor
vectornames

#Creo las k-tablas, las Economias repetidas por Year
yr<-length(levels(datosAde4$Year))
kta1 <- ktab.within(wt1,colnames=rep(vectornames,yr))
names(kta1)
#str(kta1)
#$ 2015:'data.frame':    42 obs. of  49 variables...
#notar que LAS VBLES (indicadores) QUEDAN como
filas/individuos despues del
#ktab.within, por eso hay que trasponer

#hacer la traspuesta
kta2<-t(kta1) #por?: para que vbles queden en las columnas
EFECTIVAMENTE

#aplicar PTA
pta1<-pta(kta2,scann=FALSE)
pta1

#valores
pta1$RV
pta1$cos2
pta1$tabw
pta1$RV.eig
pta1$eig

#Grafico de la lntEREstructura
plot(pta1)

#Grafico de la lntRAEstructura
kplot(pta1)



#-------------------------------------------------------------------------
#COMPLEMENTO 2A: MFA Analisis factorial multiple (factoextra
& FactoMineR)
#-------------------------------------------------------------------------
#Este tecnica implica cambiar la estructura del dataSet de entrada
#que las filas sean las economias, y las columnas sean por Years,
y los indicadores
#por cada Year. Es decir las matrices, de cada Year, estan
concatenadas de
#manera horizontal

#Archivo de Entrada
setwd("~/R")
dataMFA<-readRDS(file="dataSet7CYMEAN")

#Quitar los factores que sobran
dataMFA$Country<-as.character(dataMFA$Country)  #quitar
elfactor que trae dataSet7
```

```
#y que cuenta con todos los factores que traia desde dataSet2... y
dataSet5...
#o sea a pesar que no esten en el dataSet como dato, siguen
apareciendo como
#levels...i.e. 266... y en el caso de reduccion de paises quedan
solo 49
dataMFA$Year<-as.character(dataMFA$Year) #quitar elfactor que
trae dataSet7
#en el caso de reduccion de years quedan solo 6... y siguen
apareciendo muchos
#fix(dataSet7) #debe venir ordenado por Year


qtyEcon<-length(levels(as.factor(dataMFA$Country)))
qtyInd<-dim(dataMFA)[2]-2
qtyYr<-dim(dataMFA)[1]/qtyEcon
qtyEcon
qtyYr


#crear df vacio como sigue no es necesario en este caso, si se
define un dataSet8
#de inicio
#dataSet8<-data.frame(matrix(vector(), qtyEcon, qtyInd*qtyYr),
#                     stringsAsFactors=F)

#prueba inicial de llenado... como inicializacion del dataSet8
i=1
dataSet8<-dataMFA[((i-1)*qtyEcon+1):i*qtyEcon,1:qtyInd]
dataSet8

#llenado
for (i in 1:qtyYr)
 {
 dataSet8[1:qtyEcon,((i-1)*qtyInd+1):(i*qtyInd)]<-
   dataMFA[((i-1)*qtyEcon+1):(i*qtyEcon),1:qtyInd]
 }
rownames(dataSet8)<-dataMFA$Country[1:qtyEcon]

head(dataSet8)
#dataSet8 ya es solo valores cuantitativos, no hay factores
#           Ec33     Ec35      Ec39      Ec42     Ec123    Pr397
Pr403
#WLD2015  2.264968 12.84498 1.43554025 0.37513410
35.299753 23.89138 2.197840e+11
#         Ec33.1   Ec35.1     Ec39.1     Ec42.1   Ec123.1  Pr397.1
Pr403.1
#WLD2015  2.468407 12.65004 1.33114093 0.500977159
34.820768 23.68584 2.068615e+11
#         Ec33.5   Ec35.5     Ec39.5     Ec42.5   Ec123.5  Pr397.5
Pr403.5
#WLD2015  2.630479 12.64428 1.59071069 2.25714280
33.092414 22.62033 2.082863e+11
#...
#al verse repetidos los nombres de columnas, R agrega .1, .2... a
los nombres
#sucesivamente para diferenciar las variables


#Archivo de Salida: Este es el archivo de entrada para MFA a
continucion
saveRDS(object=dataSet8, file= "dataSet8")


library(factoextra) #para poder presentar los graficos
library(FactoMineR) #tienen MFA
#library(Ad4) #tambien tiene MFA

#hay que tener el vector de qtyInd en cada k-tabla, aqui seria
todos con i.e.42
c(rep(qtyInd,qtyYr)) #[1] 42 42 42 42 42 42...
#si fueran diferentes hay que definirlo como un i.e.
c(40,32,45,78,55,21)

#hay que tener el vector para estandarizacion por cada k-tabla,
aqui seria
#todos como "s" = scale (estandarizar reduciendo varianza a 1 de
cada columna
#o cada variable), si se pone "c" no lo hace por columna... esto es
para vbles
#cuantitativas, si fuera cualitativa la vble, se pone "n", y cuando
sean tablas
#de contingencia o sea la vble contiene un conteo / poner poner
"f"
c(rep("s",qtyYr))

res.mfa<-MFA(dataSet8, group=c(rep(qtyInd,qtyYr)),
type=c(rep("s",qtyYr)),
          name.group = levels(as.factor(dataMFA$Year)), graph =
F)
res.mfa
res.mfa$eig #valores propios del ACP global
#         eigenvalue percentage of variance cumulative percentage
of variance
#comp 1  5.9938322698          31.028645796
31.02865
#comp 2  4.7965885771          24.830799610
55.85945
#comp 3  1.7350790583           8.982092108
64.84154
#comp 4  1.3544620144           7.011728089
71.85327
#en los 2 primeros componente hay una % vza acumulado de
55%, el AFM esta
#explicando esta evalaucion de los paises en un 55%, este
reteniendo un 55% de
#la varianza

#resultados por separado por cada k-tabla, o sea los Years
res.mfa$separate.analyses
res.mfa$separate.analyses$`2020`$eig
res.mfa$separate.analyses$`2020`$var
res.mfa$separate.analyses$`2020`$ind
res.mfa$separate.analyses$`2020`$svd
res.mfa$separate.analyses$`2020`$call

res.mfa$group
res.mfa$group$Lg #calcula el coeficiente de la relacion entre k-
tablas
#de esa similaridad, no hay un limite superior, pueden ser vlrs
mayores que 1
#2015     2016     2017     2018     2019     2020      MFA
#2015 1.769187 1.780899 1.808655 1.803134 1.786327 1.787406
1.791109
#2016 1.780899 1.814365 1.840308 1.830809 1.813502 1.818553
1.818275
#2017 1.808655 1.840308 1.898020 1.888782 1.867447 1.878580
1.865550
#2018 1.803134 1.830809 1.888782 1.914118 1.876207 1.900053
1.870774
#2019 1.786327 1.813502 1.867447 1.876207 1.878807 1.879049
1.852127
#2020 1.787406 1.818553 1.878580 1.900053 1.879049 1.949150
1.870722
#MFA  1.791109 1.818275 1.865550 1.870774 1.852127 1.870722
1.846658

res.mfa$group$RV #este coef.corr.vectorial... entre k-tablas, vlrs
acotados a 1

res.mfa$group$coord #coordenadas de cada k-tabla en el plano
factorial
#como se proyectan

res.mfa$group$contrib #como contribuyen a los componenetes
cada k-tabla, cargas

res.mfa$group$cos2 #calidad de representacion de cada k-tabla
```

```
res.mfa$group$dist2 #distancia de c/k-tabla hacia el origen

res.mfa$group$correlation #correlacion c/k-tabla con cada
componenete/eje

res.mfa$inertia.ratio #razon de inercia por c/componente en el
global, no el
#separado

res.mfa$ind
res.mfa$ind$coord #coordenada c/economia individuo en el plano
factorial global
res.mfa$ind$contrib #cuanto contribuye cada individuo/economia a
cada dimension
#i.e. ejej1 : BRA  2.103514057, CHL 12.721616145, CRI
71.835133552
#eje2: COL 5.526232328, BOL 6.329969843, SUR 7.064063203,
LDC 7.082818828,
#PRI 11.552527797, SXM 12.120650576
#eje3: GUY 21.235724088, LDC 22.043422294, SUR
22.233190935

res.mfa$ind$cos2 #calidad de representacion de
individuos/economias

res.mfa$ind$within.inertia #inercia interna, dentro del analisis, para
c/individuo
#para cada dimension

res.mfa$ind$coord.partiel #coor.parciales, se ve mejor en un
grafico, como
#proyecta sobre la misma dimension del plano bifactorial las k-
tablas (Years)
#y para cada uno de las economias/paises. se hace un grafico en
forma de estrella
#que algunos los llaman TRAYECTORIAS!!!

res.mfa$summary.quanti #analisis para las variables/indicadores
cuantitativas
#group variable      moyenne   ecart.type      minimum
maximum
#1      1     Ec33  1.084300e+02 6.610800e+02  0.000000e+00
4.683590e+03
#2      1     Ec35  1.338000e+01 5.040000e+00  2.350000e+00
2.569000e+01
#promedio moyenne, desv.std ecart.type, min y max

res.mfa$summary.quali #caso de cuantitativas, pero no hay aqui...

res.mfa$quanti.var #se abren opciones para las variables, algo
como se hizo
#arriba para los individuos, esta las $coord coordenadas para
proyectar sobre
#el plano bifactorial, $contrib cuanto contribuye cada vble en cada
dimension,
#$cos2 calidad de representacion por cada vble por cada Year,
$cor correlaciones
#con los ejes, $

res.mfa$partial.axes #coordenadas y correlaciones entre las
dimensiones o
#los componenetes de cada k-tabla
res.mfa$partial.axes$coord #coordenadas sobre el plano global
bifactorial
#group variable      moyenne   ecart.type      minimum
maximum
#1      1     Ec33  1.084300e+02 6.610800e+02  0.000000e+00
4.683590e+03
#2      1     Ec35  1.338000e+01 5.040000e+00  2.350000e+00
2.569000e+01

res.mfa$call #resumen del MFA AFM... de lo que se esta
trabajando
res.mfa$call$row.w #pesos por fila o sea por individuo, o
economia

res.mfa$call$col.w #pesos por columna, se puede seccionar para
cada k-tabla o
#Year, se observa que tienen todas el mismo peso al final, pero se
mantendria
#igual para cada k-tabla, pero diferentes entre k-tablas, aqui dio
igual
#porque al final son la misma cantidad de vbles en cada k-tabla,
42

res.mfa$call$X #data original que se trabaja

res.mfa$call$XTDC #data estandarizada, la final

res.mfa$call$nature.group #tipo de grupo: cuantitativo
res.mfa$call$nature.var #tipo vble: cuantitativa

res.mfa$global.pca$eig #info de vlrs propios de cada una de las
dimensiones
#o ejes del AFM MFA, con su % vza retenida, y la acumulada
#eigenvalue percentage of variance cumulative percentage of
variance
#comp 1  5.9938322698          31.028645796
31.02865
#comp 2  4.7965885771          24.830799610
55.85945
#comp 3  1.7350790583           8.982092108
64.84154
#con 2 ejes se retiene el 55% de la vza acumulada


#ver de manera grafica
fviz_eig(res.mfa) #scree plot, diagrama del codo, 2 primeras
componentes
#entre las 2 primeras retienen mas del 50%

fviz_mfa_group(res.mfa) #grupos de vbles, o sea por k-tablas
#se parece a la INTERestructura del STATIS-dual, donde se ubica
cada k-tabla
#el MFA AFM presenta un resultado similar en este caso, pero no
es la generalidad
#para el caso AFM MFA este analisis retiene un 55%

fviz_mfa_var(res.mfa) #proyectar las vbles indicadores de cada k-
tabla
#ver si se forman grupos diferentes: de hecho para el caso de 42
indicadores
#se forman 3 agrupaciones en el grafico
#esto es que cada agrupacion de variables/indicadores tienen
comportamientos
#diferentes, o que calificacion o puntuacion diferentes
#muy denso, hay que disminuir y dejar 2 k-tablas (Years, i.e. 2020
y 2019)

#help("fviz_mfa_var")

#select.ind, select.var, select.axes
#a selection of individuals/partial individuals/ variables/groups/axes
to be
#drawn. Allowed values are NULL or a list containing the
arguments name, cos2
#or contrib: name is a character vector containing
individuals/variables to be
#drawn. cos2 if cos2 is in [0, 1], ex: 0.6, then individuals/variables
with a
#cos2>0.6 are drawn. if cos2 > 1, ex: 5, then the top 5
individuals/variables
#with the highest cos2 are drawn. contrib if contrib > 1, ex: 5, then
the top 5
#individuals/variables with the highest cos2 are drawn

fviz_mfa_var(res.mfa,select.var=list(name="Pr421.5",cos2=0.6,con
trib=5))
fviz_mfa_var(res.mfa,select.var=list(name=NULL,cos2=0.6,contrib
=100))
```

```
fviz_mfa_var(res.mfa,select.var=list(name=NULL,cos2=NULL,cont
rib=1))

fviz_mfa_var(res.mfa,geom=c("point","text")) #eliminando los
vectores en el graf
#se ven solo puntos y textos
#de este grafico: las vbles indicadores opuestos entre ellos en un
eje, quiere
#decir que se diferencian entre ellas
#ejemplo, las de eje2 abajo, se alcanza a ver Ec696.5, indicando
que tiene las
#puntuaciones mas altas por Year (k-tabla), pero tambien vemos
hacia arriba
#en el 2o eje el He959.5 con altas puntuaciones, pero en sentido
contrario
#esto significa que para, este ejemplo, ambas indicadores son
contrarios y
#pueden ser indicativos de alguna relacion en el mismo Year, son
.5 = 2020
#vbles/indicadores cerca al origen, son puntuaciones bajas

fviz_mfa_var(res.mfa,select.var=list(name=c("Ec696.5","Pr421.5","
He965.5",
                                    "He962.5","Ec696.4","Pr421.4",
                                    "He965.4","He962.4","Ec696.3",
                                    "Pr421.3","He965.3","He962.3",
                                    "Ec696.2","Pr421.2","He965.2",
                                    "He962.2","Ec696.1","Pr421.1",
                                    "He965.1","He962.1","Ec696",
                                    "Pr421","He965","He962"),
                             cos2=NULL,contrib=NULL),
        geom=c("point","text"))

fviz_mfa_var(res.mfa,select.var=list(name=c("Ec696.5", "Pr421.5",
"He965.5",
                                    "He962.5"),
                             cos2=NULL,contrib=NULL),
        geom=c("point","text"))

fviz_mfa_ind(res.mfa) #se ven los individuos, se pueden identificar
agrupamientos
#se identifican varios grupos, y lejos de todos ellos BRA, CHI y
CRI, lo que indica
#que se diferencian del resto
#se ve tambien en eje2 abajo LDC y SUR, y arriba SXM y PRI,
estos dos, lo que
#significa que presentaron marcadas caracteristicas opuestas
dado que estan
#sentido contrario, alguno(s) indicadores que fueron altos para
unos fueron
#bajos para los otros
#esta tambien un grupo eje2 arriba BOL, COL, TTO, GUY y otros
que no se logran
#identificar, que estan cercanos a SXM-PRI

#las vbles/indicadores serian los mencionados como ejemplos
He959.5 y Ec696.5
#SXM-PRI tiene vlrs indicadores altos en He959 en 2020, y bajos
en Ec696 2020
#y es contrario a lo que sucedio para LDC y SUR
#las altas puntuaciones es sinonimo de lo caracteristico del
individuo/economia
#o agrupaciones de ellos (como pasa con LDC)

fviz_mfa_ind(res.mfa,select.ind=list(name=c("LDC","SUR","SXM","
PRI","CHL","BRA",

"CRI","WLD","COL","EUU","LCN","MEA",
                                    "OED"),cos2=NULL,contrib=NULL))

fviz_mfa_ind(res.mfa,select.ind=list(name=c("LDC","SUR","SXM","
PRI","CHL","BRA",
                                    "CRI","WLD","COL","LCN","MEA"),
                             cos2=NULL,contrib=NULL))

fviz_mfa_ind_starplot(res.mfa,select.ind=list(name=c("LDC","SUR"
,"SXM","PRI","CHL","BRA",

"CRI","WLD","COL","LCN","MEA"),
                                    cos2=NULL,contrib=NULL))
#Warning message:In fviz_mfa_ind_starplot(res.mfa, select.ind =
list(name =
#c("LDC",: This function is deprecated. It will be removed in the
next version.
#Use fviz_mfa_ind(res.mfa, partial = 'All') instead.
#este es el grafico de trayectorias para cada
#individuo/economia, y cada color representa la k-tabla, y en el
sentido que
#sale del punto ese vector (largo o corto), significa la puntuacion
que tiene
#cada individuo/economia en cada k-tabla/Year
#del grafico que tenemos, resulta que no hay grandes vectores,
todos mas o menos
#cercanos al punto, es decir que a traves de las k-tablas/Years
recibieron
#puntuaciones similares cada vez, y no hubo grandes saltos en
las indicadores
#este es un grafico de coordenadas parciales

fviz_mfa_ind(res.mfa,partial="All",select.ind=list(name=c("LDC","S
XM","CHL",
                                    "BRA","CRI","WLD",
                                    "COL","LCN")))

res.mfa$ind$coord.partiel #coor.parciales, se ve mejor en un
grafico, como
#proyecta sobre la misma dimension del plano bifactorial las k-
tablas (Years)
#y para cada uno de las economias/paises. se hace un grafico en
forma de estrella
#que algunos los llaman TRAYECTORIAS!!!

#estas starplot o trayectorias se compara con el compromiso (De
los individuos)
#de STATIS (que seria la salida de trayectorias) para caso con 42
indicadores,
#con factor Year.
#ver Trayectorias STATIS en Ade4, caso 42 indicadores, Factor
usado: Year.
#En general puede que haya una estructura factorial diferente al
compromiso de
#STATIS, pero de lo visto en el caso de 42 indicadores, en
terminos generales
#se ven muy similares, haciendo los giros adecuados horizontal y
verticalmente
#aunque con una mejor ampliacion se puede llegar a unas
mejores impresiones
#del comportamiento de los individuos/economias
#dependiendo de la tecnica que se aplique se puede encontrar
estructuras diferentes

fviz_mfa_axes(res.mfa) #se analiza las dimensiones, la proyeccion
de los
#componentes ppales de cada ACP que se realiza de manera
individual en c/k-tabla
#el CP1 de c/ktabla esta muy asociado a la 1a
componente/dimension del plano
#bifactorial o plano del analisis factorial multiple, igualmente pasa
con las
#2as dimensiones de c/ktabla estan asociadas a la 2a dimension
del plano
#bifactorial o factorial multiple, las otras dimensiones para cada
ACP independiente
#tienen vectores muy pequenos, la variabilidad es muy pequeno,
no aportan



#------------------------------------------------------------------------
#COMPLEMENTO 2B: MFA Analisis factorial multiple (Ade4)
```

```
#-------------------------------------------------------------------------

#Archivos de Entrada
setwd("~/R")
dataSet7CYMEAN<-readRDS(file="dataSet7CYMEAN")
dataSet8<-readRDS(file="dataSet8")

#Quitar los factores que sobran
dataSet7CYMEAN$Country<-
as.character(dataSet7CYMEAN$Country)
dataSet7CYMEAN$Year<-as.character(dataSet7CYMEAN$Year)

#fijar constantes
qtyEcon<-length(levels(as.factor(dataSet7CYMEAN$Country)))
qtyInd<-dim(dataSet7CYMEAN)[2]-2
qtyYr<-dim(dataSet7CYMEAN)[1]/qtyEcon
qtyEcon
qtyYr


library(ade4)
w1=scalewt(dataSet8,scale=T)
w1=data.frame(w1)

c(rep(qtyInd,qtyYr))
levels(as.factor(dataSet7CYMEAN$Year))

kta1=ktab.data.frame(w1,c(rep(qtyInd,qtyYr)),

tabnames=levels(as.factor(dataSet7CYMEAN$Year)))
afm1=mfa(kta1,scannf=F,nf=2)
#Grafica de la IntEREstructura
plot(afm1)
#Se ve ubicacion de las k-tablas Years, con respecto al
componente 1 y 2
#Tambien se visualiza la proyeccion de las vbles/indicadores por
c/ktabla (seria
#el compromiso), y se forman los agrupamientos, y se hace el
similar analisis
#de bajas y altas puntaciones/ scores, cuales estan opuestas
entre si.
#tambien se visualiza la proyeccion de las filas / Economias en el
plano factorial

#Grafica de la IntRAEstructura
kplot(afm1)
#se muestran las k-tablas que son los Year, y los grupos de
indicadores por c/u
#Y deberia estar la proyeccion por individuos, las trayectorias,
pero no se
#logran ver los vectores porque son muy pequenos, indicativo que
no han tenido
#mayor variacion entre k-tablas, son los puntos negros...




########################################################
##########################
#Analisis de 2 vias, escogiendo solo un Year, o sea 1 sola k-tabla
########################################################
##########################

#-------------------------------------------------------------------------
# STATIS biplot
#-------------------------------------------------------------------------

#Archivo de Entrada
setwd("~/R")
dataSet7CYMEAN<-readRDS(file="dataSet7CYMEAN")

#grabar dataSet7CYMEAN como .CSV para usar en MultBiplot
stand alone como
#entrada tipo EXCEL
write.csv(dataSet7CYMEAN, "dataSet7CYMEAN.csv")

datosXconvert<-dataSet7CYMEAN #trae 2 factores, country y
year, y sin NAs
#str(datosXconvert)

#convierto a factores country y year de ser necesario
#datosXconvert$Year=factor(datosXconvert$Year)
#datosXconvert$Country=factor(datosXconvert$Country)

h<-dim(datosXconvert)[2]
h
library(MultBiplotR)
X=Convert2ThreeWay(datosXconvert[,c(-h+1,-
h)],datosXconvert$Year,
                   RowNames=datosXconvert$Country)
#help(Convert2ThreeWay)
#str(X)
#X
#names(X)
#dim(X$`2020`)
#X$`2020` #los Years / momentos / ocasiones

#Escoger el ultimo Year, la ultima k-tabla
qtyOca<-length(X) #cantidad ocasiones/ momentos / Years
nombreUlt<-names(X)[qtyOca] #nombre del ultimo
nombreUlt #[1] "2020"
concatena<-paste("X$`",nombreUlt,"`")
concatena
#dim(concatena) #no funciona asi...
#Toco MANUALMENTE
qtyInd<-dim(X$`2020`)[2]

#HJ-Biplot inducido por el STATIS
stbip=StatisBiplot(X, SameVar=TRUE)
#str(stbip)
#help(StatisBiplot)
#StatisBiplot #se ve el codigo que confirma la funcion

summary(stbip)
plot(stbip,PlotType="InterStructure")
plot(stbip, VarColorType="ByTable")
plot(stbip, VarColorType="ByTable",WhatVars=1:2)
plot(stbip, VarColorType="ByVar")
plot(stbip, VarColorType="ByVar",WhatVars=1:2)

plot(stbip, PlotRowTraj = F, PlotVars=F,
    LabelTraj='Begining',PlotVarTraj=T,
    VarColorType="ByVar", ShowBox=TRUE)
#plot(stbip, PlotRowTraj = T, PlotVars=F,
#     LabelTraj='Begining',PlotVarTraj=T,
#     VarColorType="ByVar", ShowBox=TRUE) #no se ve muy
bien...

plot(stbip, PlotRowTraj = T, LabelInd=F, PlotVars=F,
LabelTraj='Begining',
    PlotVarTraj=F, VarColorType="ByVar",
RowRandomColors=TRUE, ShowAxis=TRUE)
plot(stbip, PlotRowTraj = F, LabelInd=F, PlotVars=F,
LabelTraj='Begining',
    PlotVarTraj=T, VarColorType="ByVar",
RowRandomColors=TRUE, ShowAxis=TRUE)
#no se ve muy bien... muy denso

#variable 1 = Ec33
#var1=rep(c(1,rep(0,192)), 6) #son 1+192 indicadores, y en 6
ocasiones/tablas
#seria para el caso de 42, algo similar...
var1=rep(c(1,rep(0,qtyInd-1)), qtyOca)
var1
length(var1) #[1] 252, para el caso de 42 ind x 6 ocasiones = 252
plot(stbip, VarColorType="ByVar", WhatVars=var1, CexVar=0.7)
#no se ven reflejados los colores de los indicadores / variables
oc1=c(rep(1,qtyInd), rep(0,5*qtyInd)) #vector solo para ocasion 1a,
#las otras 5 en 0
```

```
#qtyInd pasa de 193 en el caso de ese numero de indicadores, cambia a 42 otro caso
oc1
length(oc1) #[1] 252, para el caso de 42 ind x 6 ocasiones = 252
plot(stbip, VarColorType="ByTable", WhatVars=oc1, CexVar=0.7)
#no se ven reflejados los colores de las tablas
plot(stbip, VarColorType="ByTable", WhatVars=1:5, CexVar=0.7)

#variable ultima (i.e. en 42 ind era "Pr1418") de todas las ocasiones
var42<-rep(c(rep(0,qtyInd-1),1),qtyOca)
var42
length(var42)
plot(stbip, mode="ah",VarColorType="ByTable", WhatVars=var42, CexVar=0.7)

names(dataSet7CYMEAN) #seleccionar la vble de interes para el biplot inducido Statis
#i.e. Ec123 = indicador / variable en posicion 5, acomodar el vector
#variable 1 de todas las ocasiones
#[1] "Ec33"   "Ec35"   "Ec39"   "Ec42"   "Ec123"  "Pr397"
"Pr403"  "Pr405"
#[9] "Pr421"  "Pr429"  "Pr448"  "Pr449"  "Pr455"  "Pr456"
"Pr459"  "Pr460"
#[17] "Pr461"  "Pr465"  "Pr528"  "Pr530"  "Ec560"  "Ec562"
"Ec565"  "Ec594"
#[25] "Ec600"  "Ec629"  "Ec696"  "He927"  "He933"  "He937"
"He959"  "He962"
#[33] "He965"  "He987"  "He1051"  "He1052"  "He1063"
"Pr1352"  "Pr1358"  "Pr1416"
#[41] "Pr1417"  "Pr1418"
#los coincidentes de Sparse STATIS-dual 42 indicadores son:
#Ec696, Pr421, He965, He962, y que tambien se ven en el Sparse HJ-Biplot
#entonces usamos esos (ellos resultaron de aplicar el codigo al WDI descargado
#en Abril 27 de 2022 que tenia 1445 indicadores base) para mantener el analisis
#sobre ellos mismos aqui en el biplot inducido por Statis, estas son las
#posiciones 27, 9, 33, 32 de ese entonces
#pero en el WDI ultimo descargado de Septiembre viene con 1442 indicadores
#y esto cambia las proporciones al calcular los %NA para eliminar indicadores
#y en consecuencia cambia el orden de los indicadores: 28, 10, 39, 38
names(dataSet7CYMEAN)[27] #Ec696
names(dataSet7CYMEAN)[28]
names(dataSet7CYMEAN)[9] #Pr421
names(dataSet7CYMEAN)[10]
names(dataSet7CYMEAN)[33] #He965
names(dataSet7CYMEAN)[39]
names(dataSet7CYMEAN)[32] #He962
names(dataSet7CYMEAN)[38]
pos<-39 #escoger la posicion de 1 hasta la cantidad de variables / indicadores
#para el caso de 1 a 42, seleccionar un numero/posicion
varN<-rep(c(rep(0,pos-1),1,rep(0,qtyInd-pos)),qtyOca)
varN
length(varN)
plot(stbip, mode="ah",VarColorType="ByTable", WhatVars=varN, CexVar=0.7)
plot(stbip,VarColorType="ByTable", WhatVars=varN, CexVar=0.7)
plot(stbip, mode="ah",VarColorType="ByVar", WhatVars=varN, CexVar=0.7) #aqui
#no se ve nada diferencial... los colores...

#class(stbip) #"StatisBiplot"
#help(plot.StatisBiplot)

#dibujar variables entre la 11 a la 20
plot(stbip, mode="ah", WhatVars=11:20, margin=0.2, CexInd=0.5)
#mode ah extiende la linea-flecha

#dibujar variables de la 1 a la 5
plot(stbip, WhatVars=1:5, margin=0.2, CexInd=0.5)

#circulo de correlaciones
plot(stbip, VarColorType="ByVar", PlotType = "Correlations")
#plot(stbip, WhatVars=1,VarColorType="ByVar", PlotType = "Correlations") #nada
#plot(stbip, VarColorType="ByTable", PlotType = "Correlations")

#Grafico de contribuciones
plot(stbip, VarColorType="ByVar", PlotType = "Contributions")
#plot(stbip, VarColorType="ByTable", PlotType = "Contributions")



#------------------------------------------------------------------------------
# Sparse HJ-Biplot
#------------------------------------------------------------------------------

#Archivo de Entrada
dataSet7CYMEAN<-readRDS(file="dataSet7CYMEAN")

library(SparseBiplots)
Data<-dataSet7CYMEAN[dataSet7CYMEAN$Year=="2020",]
#aqui se fija el Year
#names(Data)
#Data$Year
Data$Year<-NULL
rownames(Data)<-Data$Country
rownames(Data)
Data$Country<-NULL
names(Data)
DataHJ <- HJBiplot(Data)
Plot_Biplot(DataHJ)
DataRidge <- Ridge_HJBiplot(Data, Lambda = 0.01)
#DataRidge <- Ridge_HJBiplot(Data) #es lo mismo que aplicar HJBiplot(Data)
Plot_Biplot(DataRidge, axis = c(1,2)) #plano 1-2, es el default sino se
#pone nada, i.e. Plot_Biplot(DataRidge)
Plot_Biplot(DataRidge, axis = c(1,3)) #plano 1-3
#los puntos rojos son los individuos se les puede poner los nombres y
#cambiar de color, argumentos del Plot_Biplot: ind.color, ind.label, etc
Plot_Biplot(DataRidge, axis = c(1,2), ind.label = T)
DataLASSO <- LASSO_HJBiplot(Data, Lambda = 0.1)
Plot_Biplot(DataLASSO, ind.label = T) #He965, He962, Ec696, Ec35
DataElastic <- ElasticNet_HJBiplot(Data, Lambda = 0.01, Alpha = 0.001)
Plot_Biplot(DataElastic, ind.label = T)
DataElastic$n_ceros

#averiguar desde WDI nombres de algunos que se ven en el biplot resultante,
#para revisar y concluir
IndSparseHJBiplot42<-c("He962","He965","Ec35","Ec696", "Pr449")
dataSet3Topics<-readRDS(file="dataSet3Topics")
IndInteresSHJB42<-dataSet3Topics[dataSet3Topics$IDVble %in% IndSparseHJBiplot42,]
IndInteresSHJB42$`Series Code`



#------------------------------------------------------------------------------
# PermANOVA & Bootstrap
#------------------------------------------------------------------------------

#usando la linea de trabajo como si fueran datos omicos, aunque aqui los
```

```
#factores son 2 pero tienen demasiadas categorias, i.e. Country
266 y Year
#62, la matriz de contraste C seria inmanejable... pero se puede
calcular la
#matriz de distancias para el permanova y para el bootstrap, pero
... time!!!

#Archivo de Entrada
setwd("~/R")
basePermanova<-readRDS(file="dataSet5CYNA")
#basePermanova<-readRDS(file="dataSet7CYMEAN")
head(basePermanova)
names(basePermanova)

#dejar solo 1 Year para el analisis, volverlo de dos vias
yr<-c("2020")
library(dplyr)
basePermanova<-basePermanova %>%
dplyr::filter(basePermanova$Year %in% yr)

#Poner como rownames los nombres de las Economias/Paises
rownames(basePermanova)<-basePermanova$Country

#Quitar Country y Year de la basePermanova, para dejarla solo
numerica
b<-dim(basePermanova)[2]
b #[1] 50 si dataSet7 en Sep2, [1] 792 si dataSet5 en Sep2
#la idea al tratarlo como omico es un numero mayor de
indicadores que de individuos
basePermanova<-basePermanova[,c(-b+1,-b)]

#Descriptiva
boxplot(basePermanova) #en estas diferencias aqui es que se
nota que hay
#que normalizar o no... NOTA: si aplico el scale despues no
calcula las
#distancias euclideas en permanova

#Estandarizar (no se aplica)
#basePermanova<-scale(basePermanova)
#si se estandariza, al dar el comando de Permanova
#bdca=BootDistCanonicalAnalysis(DistDatos, group)... hay error,
por no aplicar
#Error in svd(B) : infinite or missing values in 'x'
boxplot(scale(basePermanova)) #si se estandarizara... pero NO es
el caso

#NAs cuantos?
sum(is.na(basePermanova)) #[1] 0 si viene de dataSet7, [1] 21745
si dataSet5
dim(basePermanova)[1]*dim(basePermanova)[2] #Total datos
sum(is.na(basePermanova))/(dim(basePermanova)[1]*dim(basePe
rmanova)[2])
#[1] 0.5617412

#Imputacion de multiples columnas
for(i in 1:ncol(basePermanova)) {
  basePermanova[,i][is.na(basePermanova[,i])]<-
mean(basePermanova[,i],
                                    na.rm=T)
}
head(basePermanova, 1)

#otra vez: NAs cuantos?
sum(is.na(basePermanova)) #[1] 0
dim(basePermanova)[1]*dim(basePermanova)[2] #[1] 38710
sum(is.na(basePermanova))/(dim(basePermanova)[1]*dim(basePe
rmanova)[2])
#[1] 0.2721519 ... Aun persisten los NAs... por que?

#la otra seria aplicar desde aqui el omit por filas
#basePermanova<-na.omit(basePermanova)
#sum(is.na(basePermanova)) #[1] 0
#dim(basePermanova)[1]*dim(basePermanova)[2] #[1] 38710
#sum(is.na(basePermanova))/(dim(basePermanova)[1]*dim(baseP
ermanova)[2])
#pero se puede quedar sin filas... 0

#quitar NAs por col segun %
colMeans(is.na(basePermanova))
porcentaje <- .1
columnas_a_borrar <-
which(colMeans(is.na(basePermanova))>porcentaje)
basePermanova[,columnas_a_borrar] <- NULL

#revision de NAs 3a vez:
sum(is.na(basePermanova)) #[1] 0
dim(basePermanova)[1]*dim(basePermanova)[2] #[1] 38710
sum(is.na(basePermanova))/(dim(basePermanova)[1]*dim(basePe
rmanova)[2]) #[1] 0

dim(basePermanova) #[1] 49 575

#borrar el individuo INX, si estuviese, que no contiene
informacion...
which(rownames(basePermanova)=="INX")
basePermanova<-
basePermanova[!(rownames(basePermanova)=='INX'),]
dim(basePermanova) #[1] 49 575

#Hasta aqui la basePermanova esta sin NAs, matriz completas,
mas variables que
#individuos

#Ls grupos que nos interesan
#WDI tiene por Region, Income o Lending

#Archivo de Entrada
setwd("~/R")
library(readxl)
#WDIEXCEL conTablasDinamicas conGrpRegionIncome.xlsx es
un EXCEL modificado
#de la WDI descargada, que en la hoja Country se le agregaron
Regiones a las
#Economias que no tenian, tambien a los Incomes, y a los
Lendings, se dio
#valor NA donde no aplicara para no dejarlo en blanco
groupbase <- read_excel("WDIEXCEL conTablasDinamicas
conGrpRegionIncome.xlsx",
                        sheet = "Country")

names(groupbase)
head(groupbase)
groupbase<-groupbase[,c(1,8,9,14)]
#[1]Country, [8]Region, [9]Income Group, [14]Lending category
groupbase

Economias<-c(rownames(basePermanova))
Economias
library(dplyr)
groupbase<-groupbase %>% dplyr::filter(groupbase$`Country
Code` %in% Economias)
#fix(groupbase)

#Selecciono el grupo a usar en Permanova...
group<-factor(groupbase$Region)
levels(group)
length(group)

#ahora si el permanova, de la libreria permnova
library(PERMANOVA)
DistDatos <- DistContinuous(basePermanova)
#Error in y[i, ] : subscript out of bounds
DistDatos$D
PERM <- PERMANOVA(DistDatos,group)
summary(PERM)
bdca=BootDistCanonicalAnalysis(DistDatos, group)
#si se estandariza basePermanova... hay error:
#Error in svd(B) : infinite or missing values in 'x'
```

```
plot(bdca, centred = FALSE)
#sale el Plot pero hay un error:
#Error in CoordinatesMeans - x$MeanCoordinates : non-
conformable arrays

plot(PERM, LabelInd = FALSE, PlotInd=TRUE)
#Error in plot.PERMANOVA(PERM, LabelInd = FALSE, PlotInd =
TRUE) :
#  You have to calculate the principal components when run
PERMANOVA

bca <- BootDistCanonicalAnalysis(DistDatos, group, dimens=2,
                          ProcrustesRot=TRUE, PCoA="Weighted")
plot(bca, A1=1, A2=2, confidence=0.95, BootstrapPlot="el",
centred=FALSE,
    PlotReplicates=FALSE)
#sale el plot con los grupos

library(MultBiplotR)
pco1=PrincipalCoordinates(DistDatos)
pco1=AddCluster2Biplot(pco1, ClusterType = "us", Groups =
group)
pco1$TypeData="Continuous"

plot(pco1, PlotClus = TRUE, ClustCenters = TRUE )

bip=PCA.Biplot(basePermanova)
#Error in svd(X, nu = dimension, nv = dimension) :
#infinite or missing values in 'x'
bip=AddCluster2Biplot(bip, ClusterType = "us", Groups = group)

plot(bip, MinQualityVars = 0.8, PlotClus = TRUE, ClustCenters =
TRUE,
    PlotInd = FALSE, LabelInd = FALSE)
plot(bip, MinQualityVars = 0.9, PlotClus = TRUE, ClustCenters =
TRUE,
    PlotInd = FALSE, LabelInd = FALSE)


#----- FIN!!!
```

Estudiante: La ciencia que aprendí en la Cueva de Salamanca, de donde yo soy natural, si se dejara usar sin miedo de la Santa Inquisición, yo sé que cenara y recenara a costa de mis herederos; y aun quizá no estoy muy fuera de usalla, siquiera por esta vez, donde la necesidad me fuerza y me disculpa; pero no sé yo si estas señoras serán tan secretas como yo lo he sido.

Cervantes. Entremés: De la cueva de Salamanca

Advierte hija mía que estas en Salamanca que es llamada en todo el mundo Madre de las Ciencias, Archivo de las Habilidades, Tesorera de los Buenos Ingenios y que de ordinario cursan en ella y habitan diez o doce mil estudiantes, gente moza, antojadiza, arrojada, libre, aficionada, gastadora, discreta, diabólica y de humor.

Cervantes. La Tía Afligida.

Mek A tel yu somtin nou

If A no waahn go nowe bikaa somtaim A no fiil fi go nowe, A no go.

But mi gaan big junivorsiti pan Salamanca bikaa mi waahn stodi muo beta

Laarn hou demya big data tingz staat fonkshian.

If A wehn nuo, A  wuda kom suuna

Dat da da

(San Andres Creole, Archipelago San Andres, Providence and Saint Ketliina)

Edwin Alexander Betancur