Máster en Data Science y Big Data

Memoria del
Trabajo de Fin de Máster

Análisis Estratégico de la Pandemia COVID-19

Un enfoque basado en datos para la Identificación de Patrones y Factores Clave

(VERSIÓN PARA ANALISTAS)

José Manuel Segovia Valdivia

24 de octubre de 2025

CFP Centro de Formación Permanente

ÍNDICE DE CONTENIDOS:

1 OBJETIVOS DEL PROYECTO

1.1 CONTEXTO

La pandemia **COVID-19** fue una situación que azotó al mundo entero. Ningún gobierno estaba totalmente preparado para lo que ocurrió esos años, marcando un antes y un después en los equipos de respuesta sanitaria de cada país.

Esto dio lugar a una cantidad de datos diarios a nivel global masiva sobre la evolución de la pandemia desde el año **2020** hasta **2024**, datos que se analizarán en esta memoria.

1.2 PROBLEMÁTICA

Al tratar una base de datos de esta magnitud, se encontraron varios problemas. Principalmente se centran en tres áreas:

- **Heterogeneidad en la Frecuencia de Reporte**: Los países reportaron sus datos con distintas frecuencias (diaria, semanal, etc.), lo que dificulta la comparación directa y la construcción de modelos temporales consistentes.
- **Calidad y Carencia de Datos**: Se observaron vacíos significativos en los datos de ciertas variables para múltiples países, generando sesgos importantes en los resultados si no se tratan con detenimiento.
- **Alta Dimensionalidad**: Presenta una gran cantidad de variables aumentando la complejidad del tratamiento, que, unido al punto anterior, hacen necesarias técnicas de reducción de dimensionalidad o de identificación de patrones.

1.3 OBJETIVOS

Para abordar esta problemática, se establecen los siguientes objetivos:

- Realizar un **Análisis Exploratorio de Datos (EDA)** para visualizar la evolución de la pandemia a nivel global, continental y nacional, e identificar correlaciones iniciales entre variables.

- **Segmentar los países** en grupos homogéneos mediante un análisis de **clustering** (K-Means) para identificar distintos perfiles de comportamiento y características frente a la pandemia.
- Implementar un sistema **de detección de anomalías** en las series temporales de casos y muertes para identificar y visualizar picos o caídas de datos inusuales.
- Evaluar los **factores asociados a la mortalidad** a través de un modelo de regresión OLS, estimando el impacto de variables como la vacunación, las restricciones y la demografía.
- Construir un **Índice de Eficiencia Sanitaria** para crear un ranking de países basado en su desempeño en la gestión de la mortalidad, controlando por sus características socioeconómicas.

2 PRINCIPALES LOGROS OBTENIDOS

Se ha desarrollado un **pipeline de datos robusto**, procesando y transformando una base de datos de registros diarios masivos en un conjunto de datos semanales, limpios y consistentes. Gracias a esto, se ha combatido la falta de datos faltantes y frecuencias de reporte dispares, obteniendo una fuente fiable para el análisis posterior.

Se han identificado perfiles de países mediante clustering no supervisado, segmentando **235** países en **2** grupos distintos (clústeres) basados en sus indicadores pandémicos y socioeconómicos. Este análisis ha revelado "arquetipos" de respuesta a la pandemia, permitiendo agrupar países con índices de respuesta similar.

Se ha creado un **Índice de Eficiencia Sanitaria** que permite crear un ranking de países, que evalúa la gestión de la mortalidad de cada nación en relación con lo que se esperaría dadas sus características (PIB, demografía, etc.). Adicionalmente, se ha cuantificado el impacto de factores clave, descubriendo una relación significativa entre la vacunación y la edad media: las vacunas tienen un impacto notablemente superior en poblaciones más envejecidas.

Se ha implementado un sistema de detección de anomalías temporales, diseñado para identificar y visualizar automáticamente semanas con picos o caídas de datos anómalos en los reportes de casos y muertes.

3 METODOLOGÍA Y PROCESAMIENTO DE DATOS

Para abordar los objetivos de forma estructurada, el proyecto se ha dividido en varias fases, comenzando por la metodología de procesamiento de datos que se detalla en este capítulo. Esta etapa inicial, correspondiente al script **01_data_preparation.py**, es fundamental, pues abarca desde la selección de las fuentes hasta la creación de los conjuntos de datos limpios y consistentes que servirán de base para todo el análisis posterior.

3.1 FUENTES DE DATOS

Se ha recurrido a dos fuentes de datos principales:

- **Base de datos de COVID-19 de Our World in Data (OWID)** [1]: Este conjunto de datos es reconocido mundialmente, proporcionando una base de datos que incluye no solo indicadores epidemiológicos (casos, muertes), sino también variables sobre vacunación, así como datos demográficos y socioeconómicos (PIB per cápita, edad media, prevalencia de diabetes, etc.) para un gran número de países y regiones.
- **Archivo GeoJSON de Países** [2]: Para las visualizaciones geoespaciales, como el mapa de clústeres, se utilizó un archivo de formato GeoJSON. Este fichero contiene las geometrías poligonales de las fronteras de los países del mundo, permitiendo la vinculación de los datos analíticos con su correspondiente representación geográfica mediante los códigos de país ISO 3166-1 alfa-3.

3.2 PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL)

Se ha desarrollado un proceso ETL diseñado para mejorar la calidad del análisis, realizando filtrados de datos y creando subconjuntos de datos de uso específico.

3.2.1 Limpieza y Filtrado de Datos

3.2.1.1 *Filtrado temporal*

El conjunto de datos actualmente cuenta con datos desde inicios 2020 hasta agosto de 2025. Sin embargo, para esta memoria se decidió acotarlo al rango impuesto inicialmente, por el simple motivo de que en 2025 apenas hay reportes, aportando valores ínfimos en comparación

al total. Por tanto, se acotó al periodo comprendido entre el **1 de enero de 2020** y el **31 de diciembre de 2024**.

3.2.1.2 *División por regiones*

Para facilitar los análisis, el DataFrame inicial se dividió en tres subconjuntos según el formato de la columna ***code***:

- **Países:** Códigos de 3 letras (ej. ESP). Constituyen la unidad principal de análisis.
- **Regiones Agregadas:** Códigos con el formato OWID_ + 3 letras (ej. OWID_EUR). Utilizados principalmente en el análisis exploratorio para visualizaciones a nivel continental.
- **Descartes:** Sin código prácticamente. Entidades que no encajaban en las categorías anteriores, como agrupaciones especiales (ej. Juegos Olímpicos). Dentro de este grupo se incluyó a la Unión Europea (OWID_EU27), ya que sus datos supondrían una duplicidad con los países europeos individuales y el agregado continental.

3.2.1.3 *Análisis y selección de variables*

Se apreciaron 61 variables únicas en la fuente original:

- **Texto** (2): incluye el país/agrupación y su código asociado según OWID.
- **Categórica** (1): incluye el continente donde se encuentra agrupado cada país.
- **Fecha** (1): la fecha del reporte, incluyendo día mes y año.
- **Numérica** (55): incluye datos relacionados con los casos, muertes, población, sanidad, etc.
- **Vacía** (2): variables que deberían reportar el índice de desarrollo humano en esa zona y la esperanza de vida, sin embargo, se encuentran vacías.

Antes de empezar con el análisis exploratorio, usando el subconjunto de datos de países (valores únicos, no añadidos a continentes y repetidos), se comprobó el porcentaje de valores vacíos (NA) de cada variable (Fig. 1):

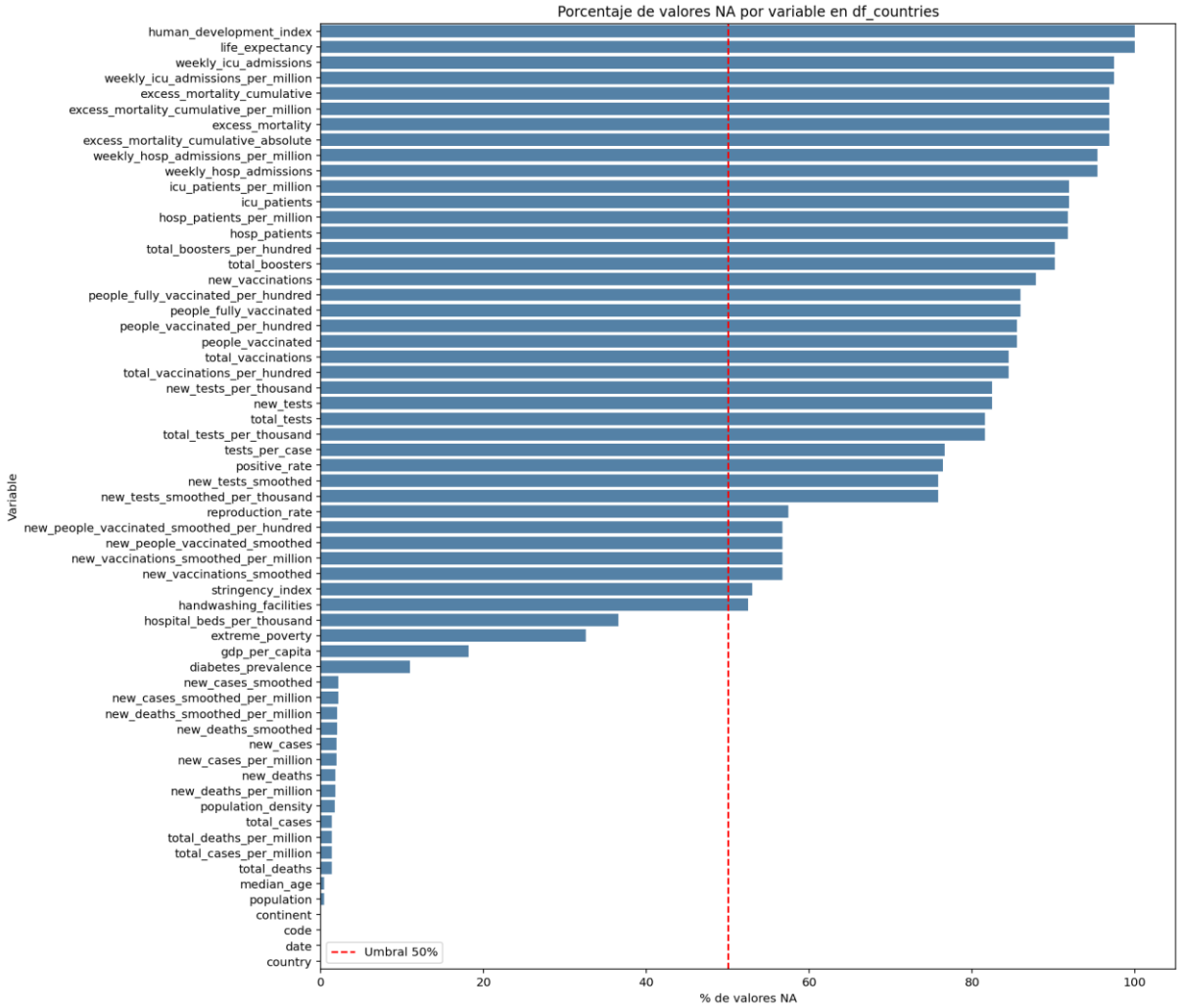

Fig. 1 Porcentaje de valores nulos (NA) por variable en el conjunto de datos de países.

Se estableció un umbral inicial del 50% de valores nulos para la preselección de variables. Como se observa en la Fig. 1, las variables a la izquierda de la línea roja cumplían este criterio. Adicionalmente, y a pesar de superar ligeramente este umbral, se decidió incluir ***stringency_index*** y ***people_vaccinated_per_hundred*** por su alta relevancia para los objetivos del estudio, ya que miden la rigurosidad de las medidas gubernamentales y el avance de la vacunación, respectivamente.

3.2.2 Agregación Temporal

Para el **análisis exploratorio**, se prepararon conjuntos de datos agrupado semanalmente, para evitar los picos que o caídas consecuentes de reportes irregulares, comenzando cada semana

en lunes (**df_countries_graph_weekly, df_regions_graph_weekly**). Para cada tipo de variable se trató de cierta forma:

- **Variables de Flujo** (ej. new_cases): Se sumaron los valores diarios de la semana. Se implementó una regla de calidad: solo se consideraba válida una semana si contenía datos de al menos 5 días, para evitar agregados engañosos a partir de reportes parciales.
- **Variables Acumuladas** (ej. total_cases, people_vaccinated): Se tomó el valor máximo registrado durante la semana.
- **Variables de Índice** (ej. reproduction_rate, stringency_index): Se calculó la media aritmética de los valores de la semana, aplicando también el filtro de un mínimo de 5 días de datos.

3.3 CREACIÓN DE DATASETS PARA ANÁLISIS

A partir del DataFrame de países ya filtrado (**df_countries**), el paso final de la preparación de datos consistió en la generación de dos conjuntos de datos especializados, diseñados para los distintos análisis del proyecto:

- **weekly_model_ready.csv**: Este es el dataset principal para los modelos de regresión. Para asegurar su robustez, su construcción partió de un **proceso de imputación de datos exhaustivo** sobre la base diaria. Se aplicaron dos técnicas:
    - **Propagación hacia adelante (forward fill):** Se utilizó para rellenar huecos en variables de serie temporal que cambian de forma intermitente, como ***people_vaccinated_per_hundred*** y ***stringency_index***.
    - **Imputación jerárquica:** Se aplicó a variables estáticas como ***gdp_per_capita*** y ***median_age***, rellenando los valores nulos primero con la mediana de su continente y, si persistían, con la mediana global. Posteriormente, este DataFrame diario e imputado fue agregado a una granularidad semanal.
- **snapshot_2024_12_31_for_clustering.csv**: Un dataset de corte transversal o "foto fija", diseñado para el análisis de clustering. Deriva de la misma base de datos diaria e imputada que el anterior, pero en este caso se extrae una única fila por país, correspondiente al último registro disponible de 2024. Esto permite segmentar a los

países en función de sus características estructurales y su estado acumulado al final del periodo de estudio.

4 ANÁLISIS EXPLORATORIO DE DATOS

Se ha procedido a realizar un análisis exploratorio inicial de los datos, partiendo desde la correlación entre variables, pasando a la evolución de la pandemia global, continental y a nivel de país, finalizando con la relación entre distintos indicadores que tuvieron impacto en la pandemia.

4.1 ANÁLISIS DE CORRELACIÓN ENTRE VARIABLES

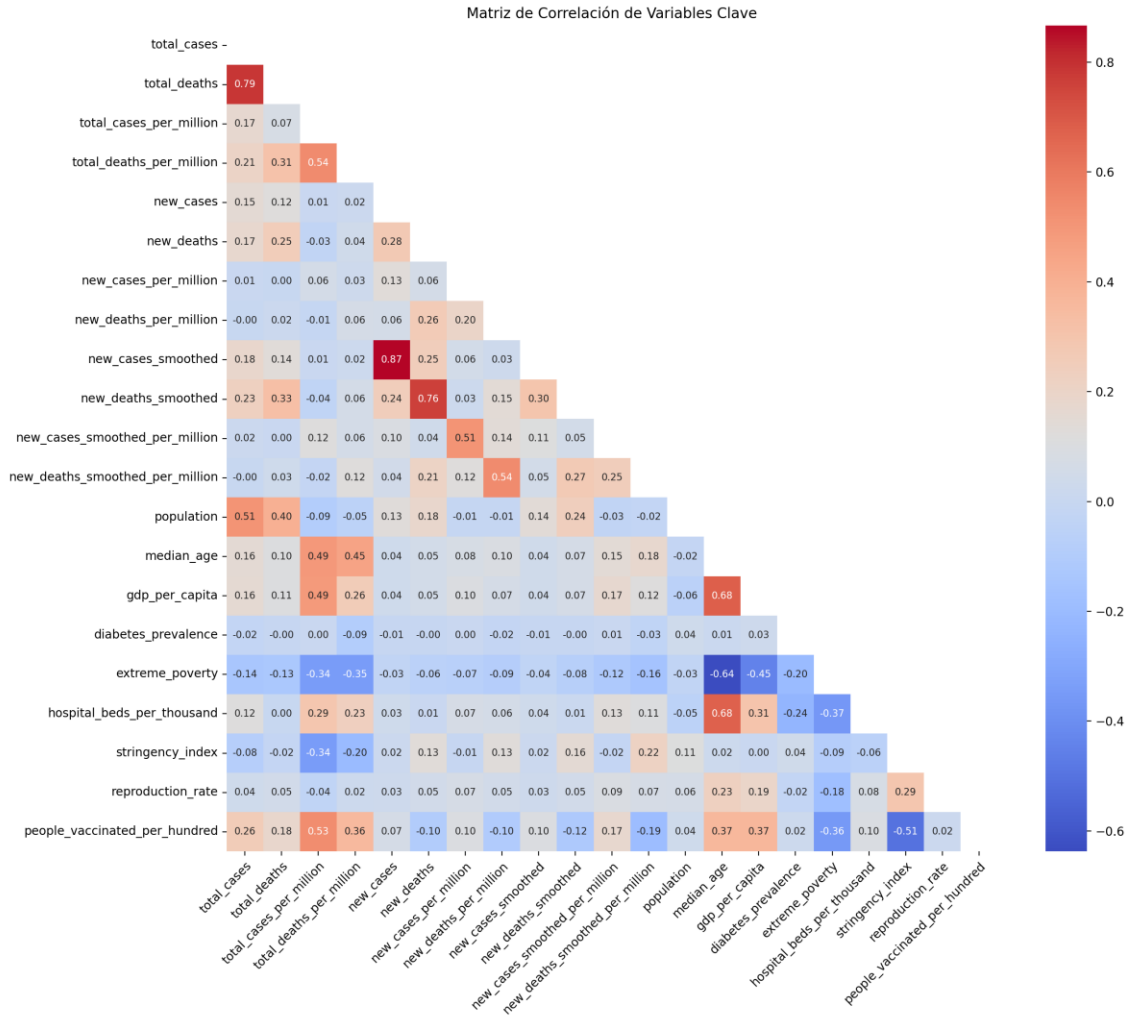

Fig. 2 Matriz de correlación de Pearson entre las variables del estudio.

Antes de comenzar, se generó una matriz de correlación para visualizar las relaciones lineales estáticas entre las variables seleccionadas para el estudio. La Fig. 2 muestra un heatmap de los coeficientes de correlación de Pearson, donde los tonos rojos indican una correlación positiva (cercana a +1), los tonos azules una correlación negativa (cercana a -1) y los colores claros una correlación débil (cercana a 0).

El análisis de la matriz con los valores exactos revela varios puntos clave iniciales:

- **Agrupación de Indicadores de Desarrollo:** Se confirma la existencia de un clúster de variables asociadas al desarrollo de un país. La **edad media** (***median_age***) presenta una fuerte correlación positiva con el **PIB per cápita** (**0.68**) y con las **camas de hospital por millar de habitantes** (***hospital_beds_per_thousand***, **0.68**). Esto indica que los países con poblaciones más envejecidas tienden a ser más ricos y a tener una mayor infraestructura sanitaria.
- **Relación con la Pobreza:** Como es de esperar, la **pobreza extrema** (***extreme_poverty***) es el contrapunto de los indicadores de desarrollo, mostrando una fuerte correlación negativa con la **edad media** (**-0.64**) y una correlación negativa moderada con el **PIB per cápita** (**-0.45**).
- **Hallazgo relevante sobre la Estrategia Pandémica:** Uno de los resultados más interesantes es la fuerte correlación negativa entre el **índice de restricciones** (***stringency_index***) y el **porcentaje de población vacunada** (***people_vaccinated_per_hundred***, **-0.51**). Este valor sugiere una dinámica clara en la gestión de la pandemia: a medida que la vacunación avanzaba en los países, las medidas de restricción tendían a relajarse.
- **Correlaciones con el Impacto del Virus:** Las variables de impacto por población, como ***total_cases_per_million***, se correlacionan positivamente con el nivel de desarrollo (***gdp_per_capita*** **0.49**; ***median_age*** **0.49**), lo que podría indicar una mayor capacidad de testeo y reporte en los países más ricos. Por otro lado, la prevalencia de **diabetes** no muestra una correlación lineal significativa con otras variables macroeconómicas o sanitarias en este análisis agregado.

4.2 EVOLUCIÓN DE LA PANDEMIA A NIVEL AGREGADO

En esta sección, se presenta una visión global y continental de la evolución de la pandemia COVID-19. Se comienza con un análisis del recuento total mundial para identificar las grandes olas y tendencias, para luego pasar a una desagregación por continentes que revele patrones geográficos.

4.2.1 Evolución Mundial de Casos y Muertes

Para entender el panorama global de la pandemia, se ha generado un gráfico de la evolución semanal de los nuevos casos y nuevas muertes a nivel mundial. Para ambas series, se han usado las variables Ambas series han sido suavizadas y las muertes se han multiplicado por 100 para facilitar su visualización conjunta:

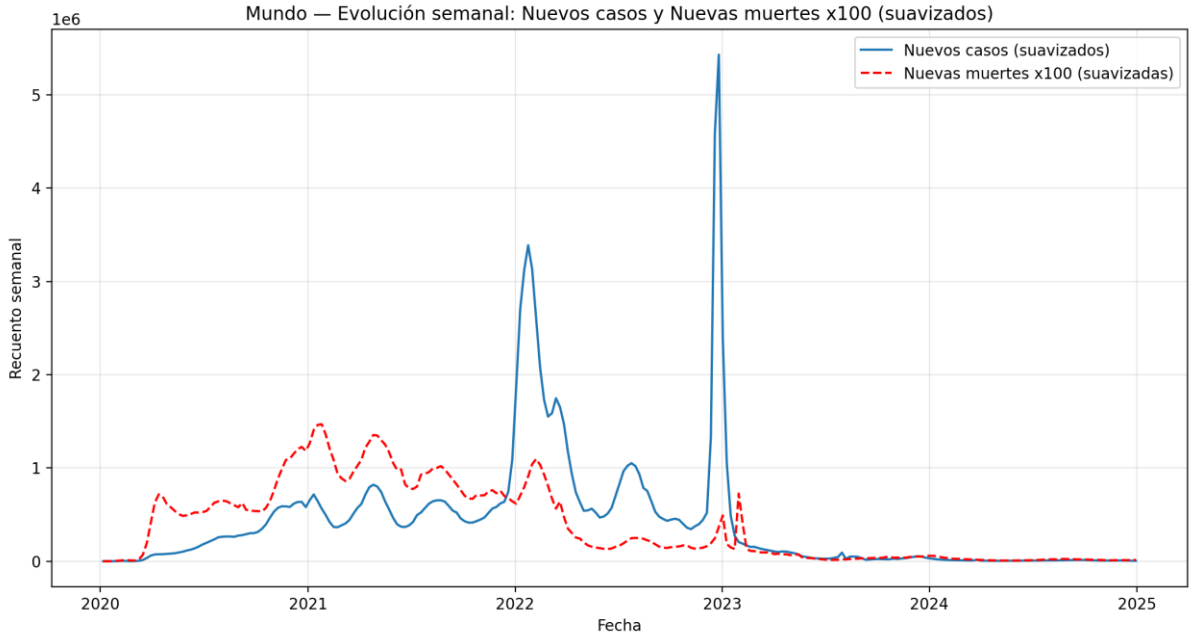

Fig. 3 Mundo - Evolución semanal: Nuevos casos y Nuevas muertes x100 (suavizados)

La Fig. 3 revela una serie de patrones clave en el curso global de la pandemia:

- **Múltiples Olas con Intensidad Variable:** Se observa claramente un patrón de múltiples picos, indicando sucesivas olas de infección a nivel mundial. Las dos olas más prominentes en términos de nuevos casos suavizados se registraron a principios

de 2022 y finales de 2022/principios de 2023. El pico de 2023 es notablemente el más alto en número de casos, superando los 5 millones de casos semanales suavizados.

- **Disociación Casos-Muertes en Olas:**
    - Durante 2020 y parte de 2021, la curva de nuevas muertes (línea roja discontinua) tiende a seguir de cerca la de nuevos casos, incluso con picos de mortalidad comparativamente altos (la línea roja se acerca o supera la línea azul en valor absoluto, recordando que las muertes están multiplicadas por 100). Esto indica una mayor letalidad inicial de la población.
    - Sin embargo, en las grandes olas de 2022 y el pico de 2023, la disociación es evidente. Aunque el número de casos se dispara a niveles récord, el aumento proporcional de muertes es significativamente menor en comparación con las olas previas. La línea roja se mantiene notablemente por debajo de la línea azul en estos picos. Esto puede atribuirse a la combinación de factores como la inmunidad natural adquirida, el avance de la vacunación y la aparición de variantes menos letales, que redujeron la gravedad de la enfermedad.
- **Tendencia General a la Estabilización y Declive**: A partir de 2023, la pandemia entra en una fase de declive sostenido. Ambas curvas, casos y muertes, muestran una clara tendencia descendente y se estabilizan en niveles muy bajos hacia finales de 2023 y durante 2024. Los picos y valles se vuelven menos pronunciados, indicando un control de la propagación y una reducción del impacto en la mortalidad a nivel global.
- **Impacto de la Vacunación y Nuevas Variantes**: La fuerte disociación entre casos y muertes a partir de 2022 es un indicativo del éxito de las campañas de vacunación masiva y, posiblemente, de la predominancia de variantes como Ómicron, que, aunque altamente transmisibles, mostraron una menor severidad.

4.2.2 Análisis comparativo por Continentes

Tras analizar la tendencia mundial, se procedió a desagregar los datos por continentes para observar las dinámicas regionales. El siguiente gráfico muestra la evolución semanal de nuevos casos y muertes (x100, suavizados) en seis continentes principales (Fig. 4).

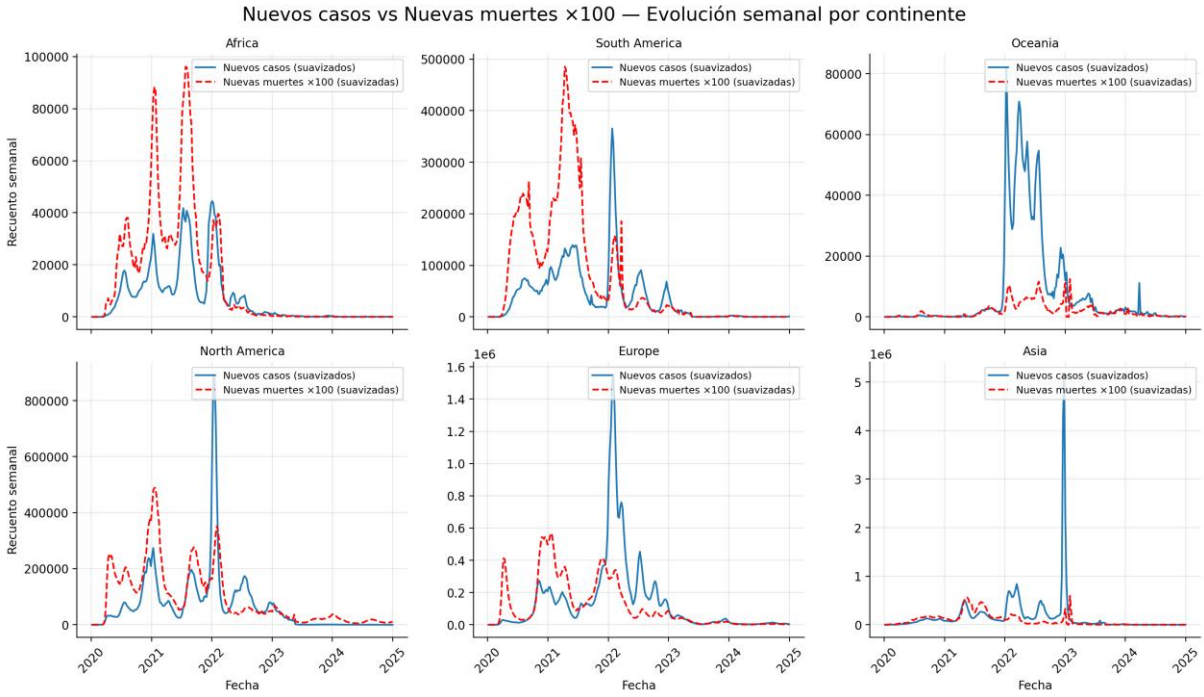

Fig. 4 Continentes - Evolución semanal: Nuevos casos y Nuevas muertes x100 (suavizados)

Este revela patrones de propagación asincrónicos y diferencias notables en el impacto de la pandemia:

- **Asincronía de las Olas:** A diferencia de la visión agregada mundial, es evidente que las olas no ocurrieron simultáneamente en todos los continentes. Aunque, parece ser que para todos excepto Asia, tuvieron los mayores picos a finales de 2021 e inicios de 2022. Asia, por su parte, muestra un comportamiento único con un pico masivo de casos a finales de 2022, mucho después que el resto del mundo.
- **Diferencias de Escala Drásticas:** La magnitud de los contagios varía enormemente. **Asia** y **Europa** son, con diferencia, los continentes que registraron los mayores picos de casos semanales. En el extremo opuesto, **África** y **Oceanía** mantuvieron un recuento de casos confirmados considerablemente más bajo durante la mayor parte del periodo.

- **Impacto Variable de la Mortalidad:** La relación entre casos (línea azul) y muertes (línea roja) también difiere notablemente:
    - En **Sudamérica** y **África**, se observa que, durante 2021, los picos de mortalidad fueron particularmente altos en relación con el número de casos.
    - En **Europa** y **Norteamérica**, la disociación entre casos y muertes es muy visible a partir de 2022, coincidiendo con las campañas de vacunación masiva y la llegada de variantes como Ómicron.
    - En **Oceanía** y **Asia**, parece ser que los picos de muertes fueron bastante más leves, estando repartidos temporalmente.

4.3 COMPARATIVA DE LA EVOLUCIÓN EN PAÍSES SELECCIONADOS

Tras examinar las tendencias a nivel continental, en esta sección se realiza un análisis más detallado sobre la evolución de la pandemia en una selección de ocho países representativos, comparados por pares. Este enfoque permite observar las particularidades de las curvas epidemiológicas nacionales, normalizadas por millón de habitantes.

4.3.1 Caso de Estudio: España vs. Alemania

Se inicia la comparativa con dos de las principales economías europeas, España y Alemania, cuyas trayectorias, aunque similares en el contexto europeo, presentan matices distintivos (Fig. 5).

- **Asincronía en picos de Muertes:** Se puede observar que España experimentó su mayor pico de muertes en el primer trimestre de 2020, debido a la falta de vacunación y medidas estrictas, puesto que hasta marzo de 2020 no comenzó el famoso confinamiento. Por su contraparte, Alemania tuvo su mayor pico a finales de 2020, siendo este menor que el de España.
- **Sincronía en picos de Casos:** Ambos países experimentaron las grandes olas de casos de forma simultánea, especialmente el gran pico de la variante Ómicron a principios de 2022.

- **Desacople conjunto de Casos frente a Muertes:** En ambos países se observa claramente cómo la curva de muertes se desacopla de la de casos a finales de 2021. A pesar de picos masivos de contagios, la mortalidad proporcional fue mucho menor, evidenciando el impacto de la vacunación y la menor letalidad de las nuevas variantes.

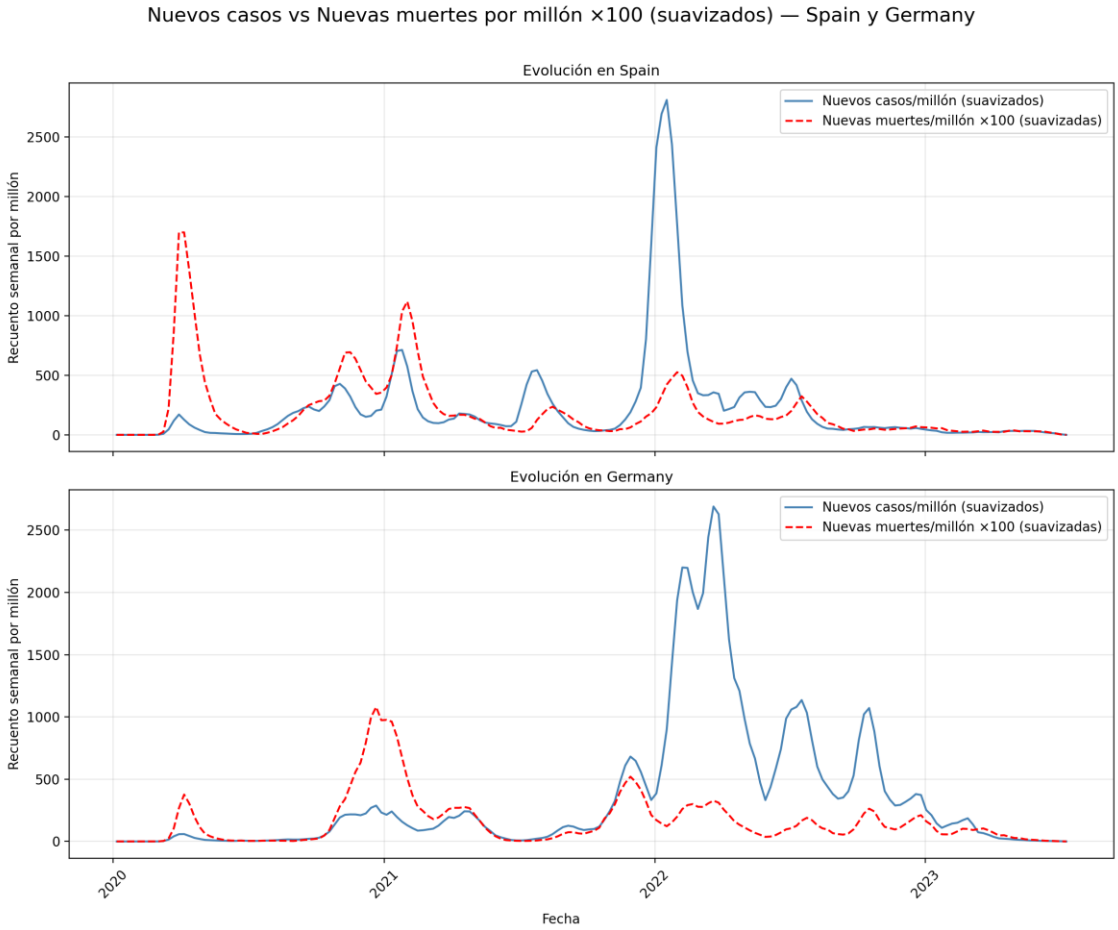

Fig. 5 España vs. Alemania: Nuevos casos y Nuevas muertes x100 (suavizados)

4.3.2 Caso de Estudio: Estados Unidos vs. Brasil

Al comparar Estados Unidos, una potencia económica de Norteamérica, con Brasil, la mayor economía de Sudamérica, se observan dinámicas de la pandemia bastante diferentes (Fig. 6):

- **Picos de Mortalidad Altos y Asincrónicos:** A diferencia del caso europeo, las olas de mortalidad no fueron simultáneas. **Brasil** sufrió un pico de muertes por millón extremadamente alto a mediados de 2021, muy superior al de Estados Unidos en ese momento. Por su parte, **Estados Unidos** experimentó sus picos de mortalidad más severos a finales de 2020 y durante la ola de Ómicron a principios de 2022.
- **Magnitud de Casos en Estados Unidos:** La curva de contagios en Estados Unidos es de una escala mucho mayor que la de Brasil, destacando el pico de la variante Ómicron a principios de 2022, que fue uno de los más altos registrados a nivel mundial en términos de casos por millón de habitantes.

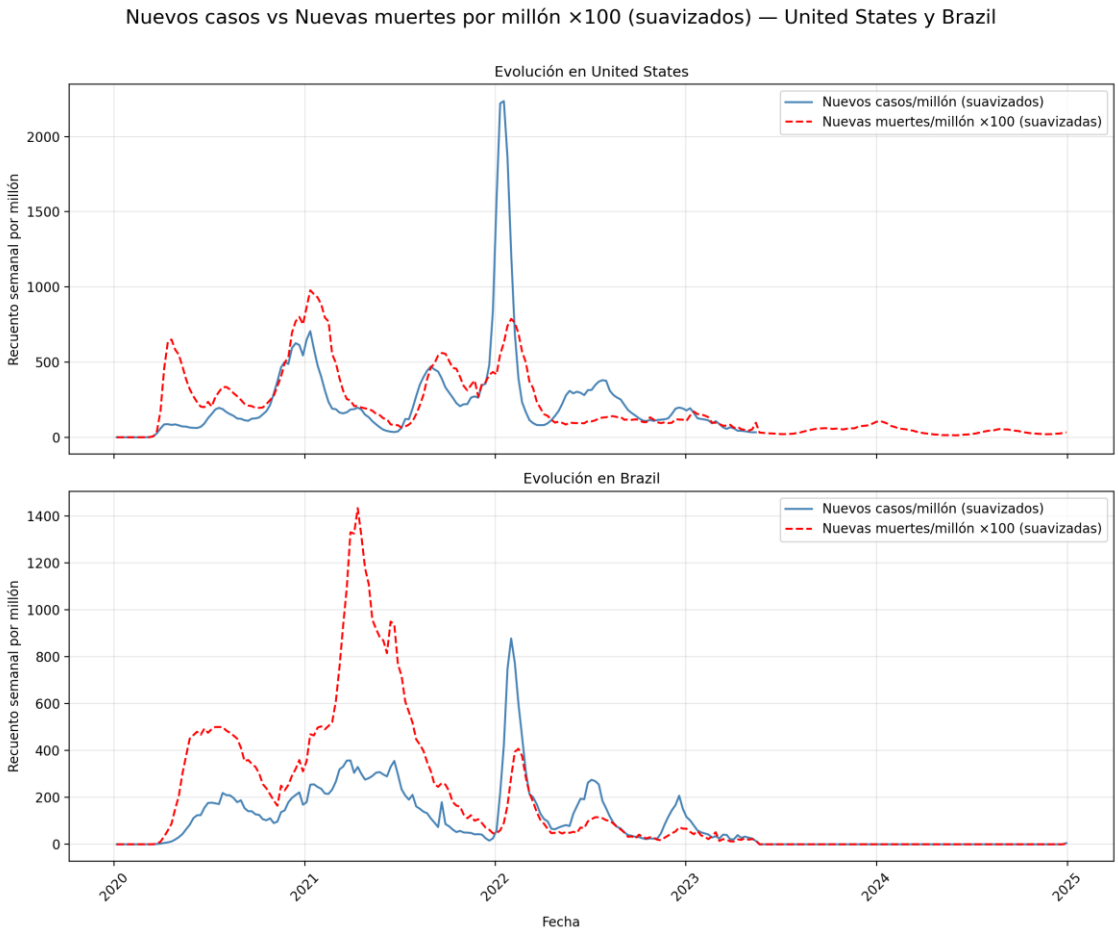

Fig. 6 Estados Unidos vs. Brasil: Nuevos casos y Nuevas muertes x100 (suavizados)

4.3.3 Caso de Estudio: India vs. China

La comparación entre las dos naciones más pobladas del mundo, India y China descubre que fueron completamente opuestas, influenciadas por sus distintas políticas de gestión (Fig. 7):

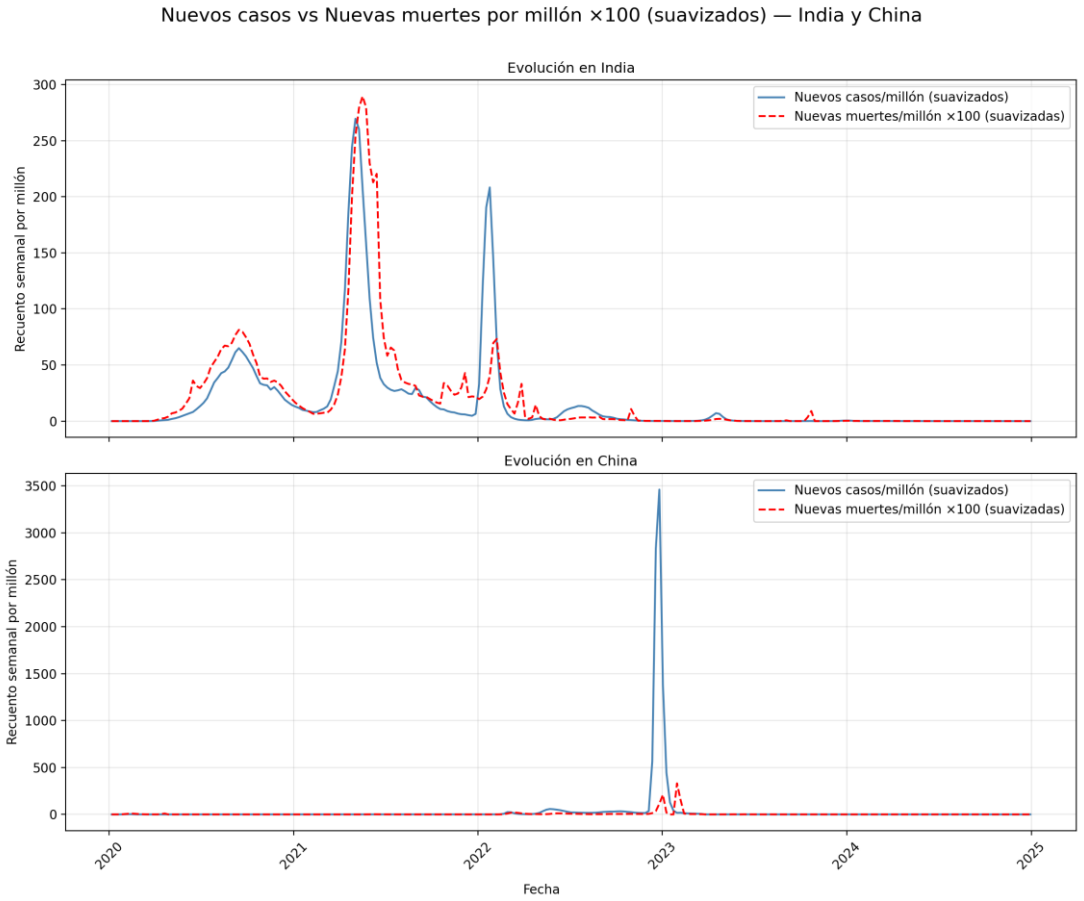

Fig. 7 India vs. China: Nuevos casos y Nuevas muertes x100 (suavizados)

- **Política "COVID Cero" vs. Olas Naturales: China** mantuvo sus cifras de casos y muertes por millón en niveles prácticamente nulos durante casi tres años, resultado de su estricta política "COVID Cero". Por el contrario, **India** experimentó olas de contagio masivas, destacando la devastadora ola de la variante Delta a mediados de 2021, que provocó su pico de mortalidad más elevado.
- **Explosión de Casos Post-Relajación en China:** El gráfico muestra un pico de casos en China a finales de 2022 de una magnitud sin precedentes, que eclipsa a cualquier otra ola vista en los países analizados. Este pico coincide con el abrupto fin de su política "COVID Cero", que liberó la transmisión en una población con baja inmunidad.

- **Datos de Mortalidad en China:** A pesar del gigantesco pico de casos, la mortalidad reportada oficialmente por China se mantuvo extremadamente baja. Esta notable discrepancia entre casos y muertes ha sido debatida a nivel internacional y sugiere un posible subreporte en las cifras de mortalidad o diferencias en la metodología de conteo durante esa ola masiva.

4.3.4 Caso de Estudio: Sur África vs. Australia

La comparación entre Sudáfrica y Australia es particularmente interesante al contrastar un país de renta media-alta del hemisferio sur con una economía desarrollada y geográficamente aislada (Fig. 8):

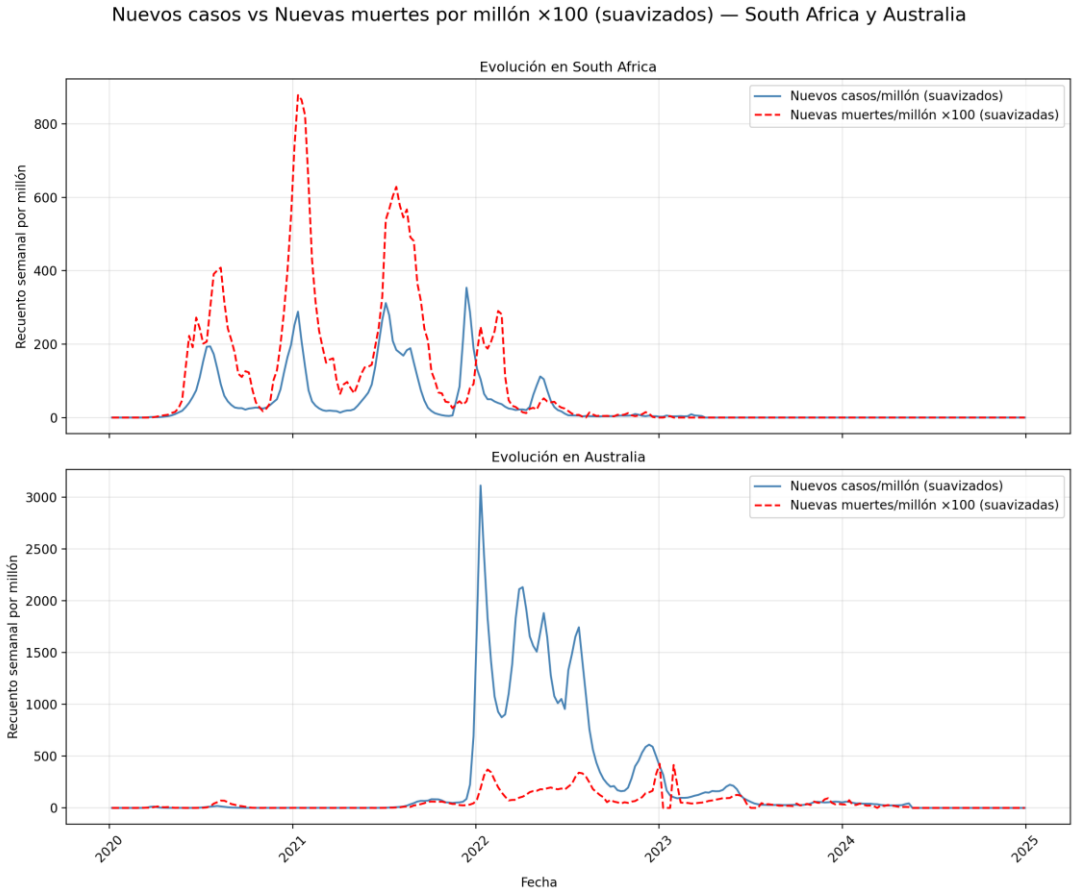

Fig. 8 Sur África vs. Australia: Nuevos casos y Nuevas muertes x100 (suavizados)

- **Gestión inicial opuesta: Australia** es un claro ejemplo de una estrategia de supresión exitosa durante los dos primeros años de la pandemia. Gracias a estrictos controles fronterizos y confinamientos, mantuvo sus casos y muertes en niveles casi nulos hasta

finales de 2021. En contraste, **Sudáfrica** experimentó varias olas significativas de contagio y mortalidad desde 2020 hasta mitad de 2022, no siendo estas tan graves como fueron para otros países.

- **Impacto de la Variante Ómicron:** Ambos países vieron sus mayores picos de casos con la llegada de la variante Ómicron a principios de 2022. Para **Australia**, esto marcó el fin de su política de "COVID Cero" y su primera ola masiva a nivel nacional. En **Sudáfrica**, la ola de Ómicron (variante que además se identificó por primera vez en este país) también generó un pico de casos muy elevado, pero su impacto en la mortalidad fue visiblemente menor que en sus olas anteriores (Beta y Delta).
- **Mortalidad Contenida en Australia:** A pesar de la explosión de casos en 2022, la curva de mortalidad en Australia se mantuvo en niveles muy bajos en comparación con Sudáfrica y otros países analizados. Esto demuestra la alta eficacia de su campaña de vacunación, que alcanzó una gran cobertura justo antes de la apertura del país y la llegada de Ómicron.

4.4 RELACIÓN ENTRE INDICADORES SOCIOECONÓMICOS Y EL IMPACTO DE LA PANDEMIA

Una vez analizada la evolución temporal de la pandemia, esta sección se centra en explorar las relaciones entre las características estructurales de los países y el impacto acumulado de la COVID-19. Para ello, se utiliza el dataset de "foto fija" que refleja la situación de cada país a finales de 2024.

4.4.1 Casos totales vs Muertes totales por millón

El primer paso es analizar la relación fundamental entre el total de casos detectados y el total de muertes, ambos normalizados por población para permitir una comparación justa entre países de diferentes tamaños.

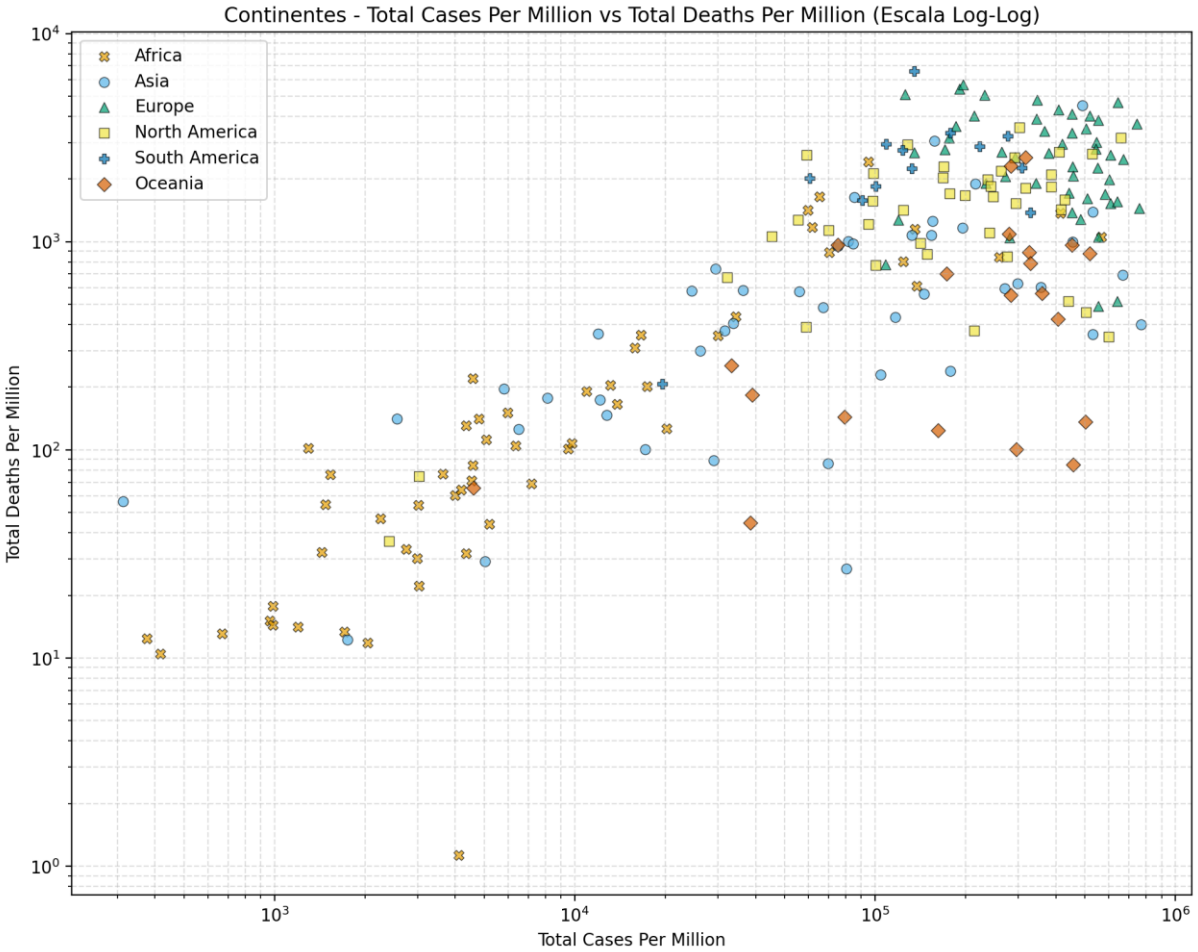

Fig. 9 Continentes - Casos totales vs. Muertes totales por millón en escala logarítmica

El gráfico de dispersión (Fig. 9), representado en una escala logarítmica, revela varias conclusiones:

- **Fuerte Correlación Positiva:** Como era de esperar, existe una **clara y fuerte tendencia positiva**. Los países con un mayor número de casos totales por millón de habitantes también tienden a registrar un mayor número de muertes por millón.
- **Segmentación por Continentes:** Se observa una evidente agrupación geográfica. Los países de **Europa (verde) y Norteamérica (amarillo)** se sitúan predominantemente en la parte superior derecha, indicando un alto número tanto de casos como de muertes reportadas por millón. Por el contrario, los países de **África (rojo)** se concentran en la parte inferior izquierda, reflejando un impacto reportado mucho menor en ambas métricas. Asia y Sudamérica muestran una dispersión más amplia.

- **Variabilidad en la Letalidad Aparente:** A pesar de la tendencia general, el hallazgo más significativo es la **dispersión vertical** de los puntos. Para un mismo nivel de casos por millón (una línea vertical en el gráfico), existe una amplia gama de resultados de mortalidad. Por ejemplo, para casos de $> 10^5$ en el X, se observa que varios países de **Sudamérica (azul oscuro)** se sitúan en la parte superior de la banda, lo que sugiere una mayor tasa de letalidad aparente (más muertes por cada mil casos) en comparación con los de **Oceanía (naranja)**, que tienden a estar en la parte inferior.

Esta variabilidad indica que, aunque la relación entre casos y muertes es directa, el resultado final no fue uniforme. Estuvo fuertemente influenciado por factores específicos de cada país o región, como la demografía, el sistema sanitario o la letalidad de las variantes predominantes.

4.4.2 Edad media vs. Muertes por millón

Se ha comparado la mortalidad acumulada por COVID-19 con la estructura demográfica de los países mediante un gráfico boxplot en la Fig. 10. En el eje horizontal, los países se agrupan en cuatro cuartiles según su edad media (de las poblaciones más jóvenes a la izquierda a las más envejecidas a la derecha). El eje vertical muestra las muertes por millón en escala logarítmica, y cada punto de color es un país (Fig. 10).

La conclusión principal es que existe una **fuerte y clara correlación positiva**: a mayor edad media de la población, mayor es la tasa de muertes por millón. Esto se evidencia en cómo las cajas (que representan la distribución de los datos) suben consistentemente de izquierda a derecha.

La distribución continental confirma esta tendencia. Los países de **África** (poblaciones más jóvenes) se agrupan en el primer cuartil con la mortalidad más baja, mientras que los países de **Europa** y **Norteamérica** (poblaciones más envejecidas) dominan los cuartiles de la derecha, donde se registran las tasas de mortalidad más altas.

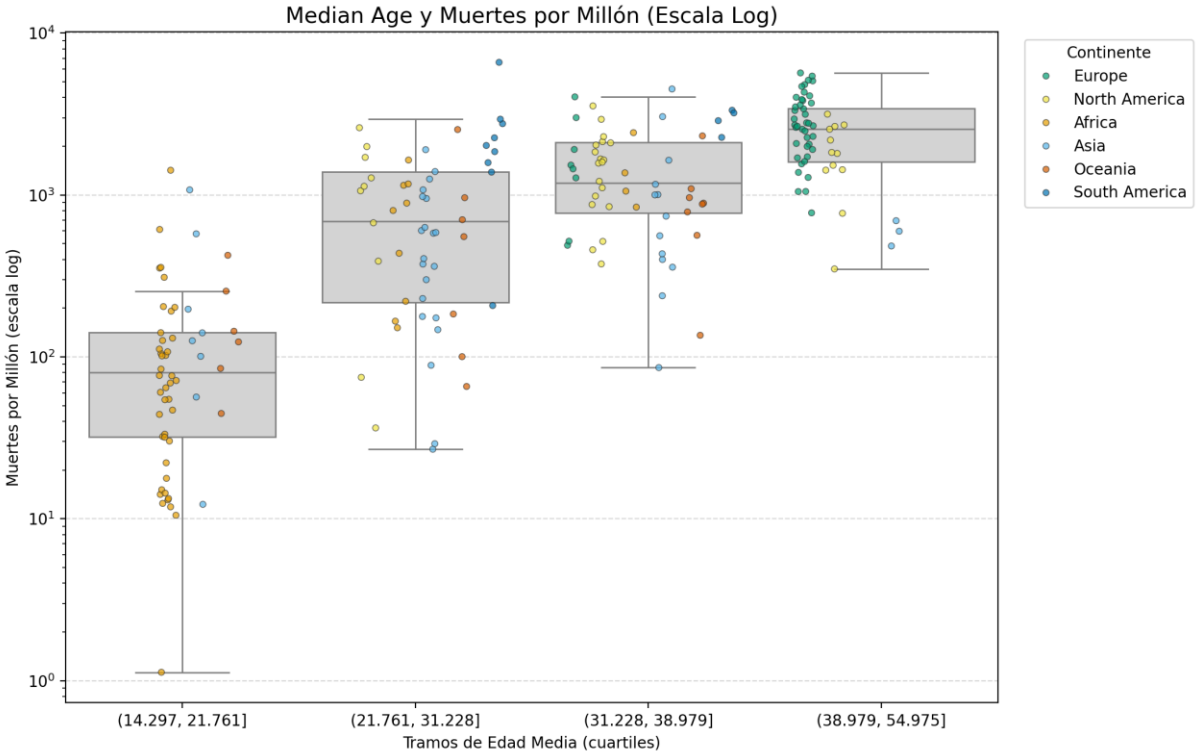

Fig. 10 Edad media vs. Muertes por millón, boxplot en escala logarítmica

4.4.3 Pobreza extrema vs. Muertes por millón

Una comparativa interesante puede ser explorar la relación entre el nivel de **pobreza extrema** de un país y su mortalidad acumulada por COVID-19. Este gráfico de violín (Fig. 11) agrupa los países en cuatro cuartiles en el eje horizontal, desde los de menor pobreza extrema a la izquierda hasta los de mayor pobreza a la derecha. La forma de cada "violín" muestra dónde se concentra la mayoría de los países de ese grupo (Fig. 11).

Se observa una **correlación negativa entre la pobreza extrema y la mortalidad reportada**. A medida que aumenta el nivel de pobreza (de izquierda a derecha), la mediana de muertes por millón desciende. Los países con menor pobreza (Q1) presentan la mayor mediana y dispersión de muertes.

Este resultado puede parecer contraintuitivo. Sin embargo, se explica por la fuerte correlación que ya vimos entre la pobreza, la edad media y el desarrollo. Los países con menor pobreza extrema suelen ser naciones más desarrolladas y con poblaciones más envejecidas, que fueron demográficamente más vulnerables al virus.

Además, es probable que los países con mayor pobreza extrema tuvieran una **menor capacidad de reporte**, lo que podría resultar en un subregistro significativo tanto de casos como de muertes.

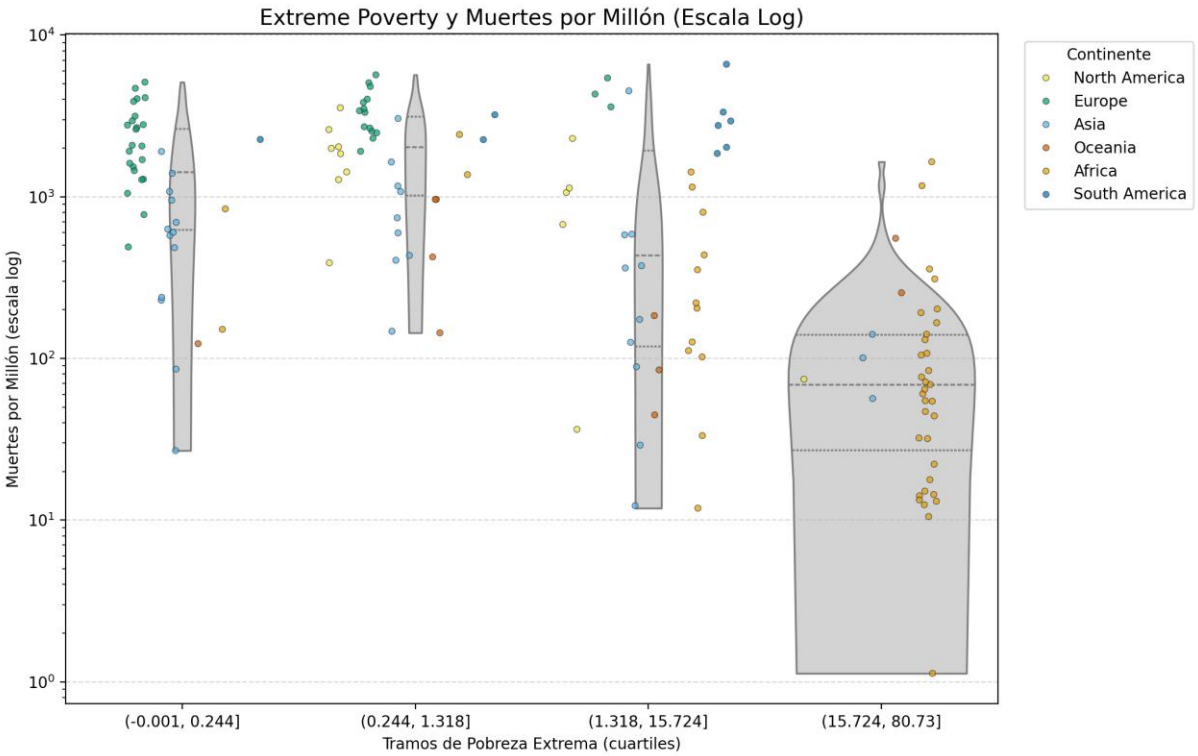

Fig. 11 Pobreza extrema vs. Muerte por millón, violinplot en escala logarítmica

4.4.4 Capacidad hospitalaria vs. Muertes por millón

Para acabar este primer análisis exploratorio, se ha comprobado la capacidad la relación entre la capacidad hospitalaria de un país (medida en camas por cada mil habitantes) y la mortalidad por COVID-19. Al igual que los anteriores, agrupa a los países en cuatro cuartiles en el eje horizontal, desde los que tienen menor número de camas a la izquierda hasta los que tienen mayor capacidad a la derecha.

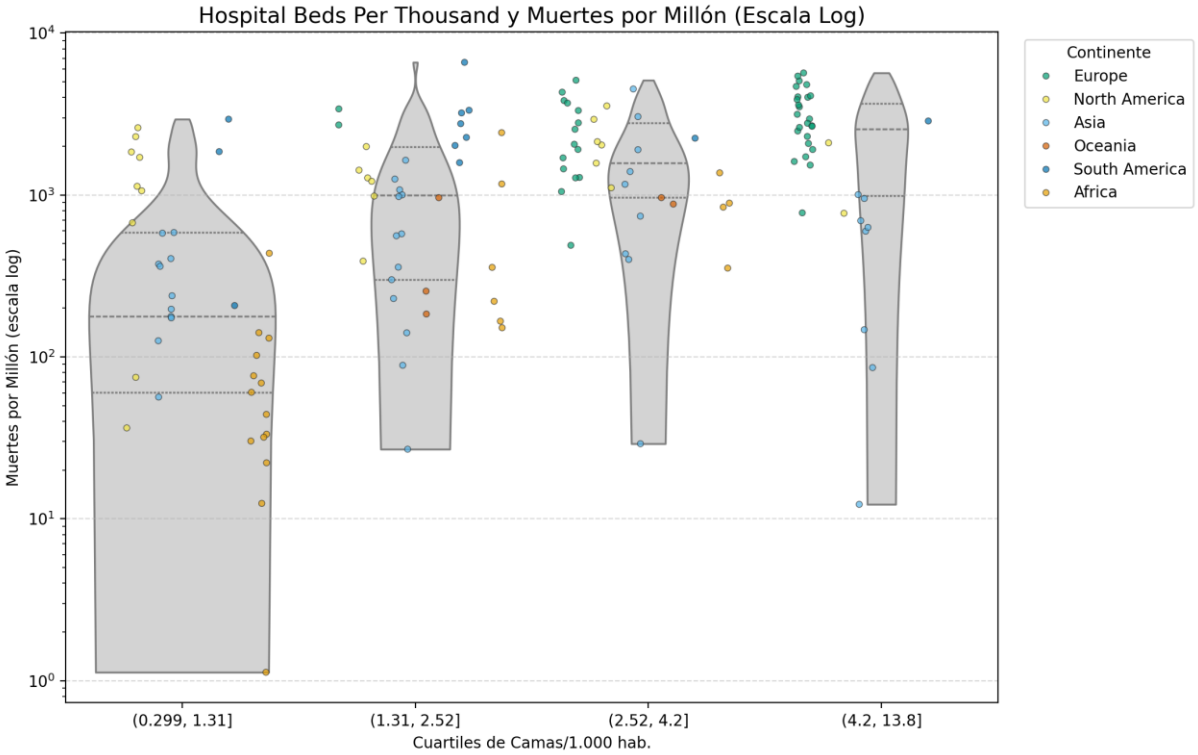

Fig. 12 Capacidad hospitalaria vs. Muertes por millón, violinplot en escala logarítmica

A primera vista, la Fig. 12 muestra una correlación positiva inesperada: los países con más camas por cada mil habitantes (Q4, a la derecha) son los que presentan una mediana de muertes por millón más alta.

Precisamente, son los países más ricos y envejecidos (principalmente de Europa y Norteamérica) los que tienen sistemas sanitarios más extensos y, a la vez, fueron demográficamente los más vulnerables a la mortalidad del virus.

5 MODELADO Y RESULTADOS

Una vez visto un análisis inicial de los datos de la COVID-19, se pasó a uno más avanzado aplicando modelos de ciencia de datos, comenzando por un análisis de clustering.

5.1 ANÁLISIS DE CLUSTERING PARA PERFILADO DE PAÍSES (K-MEANS)

Con el objetivo de segmentar los países en grupos homogéneos según su perfil pandémico y estructural, se aplicó el algoritmo de clustering no supervisado **K-Means**. Este análisis se realizó sobre el dataset de "foto fija" (snapshot) de finales de 2024, que, como se detalló en el capítulo 3, ya había sido sometido a un proceso de limpieza e imputación inicial.

5.1.1 Metodología

Para garantizar la máxima robustez del modelo, se siguieron los siguientes pasos metodológicos:

- **Validación Final de Datos:** Como **paso de seguridad** previo al modelado, se realizó una última verificación para asegurar la ausencia total de valores nulos. Cualquier ***NA*** remanente que pudiera haber persistido fue imputado utilizando la mediana de su columna, garantizando así que el 100% de los 235 países entraran en el análisis.
- **Escalado de Variables:** Dado que el algoritmo K-Means es sensible a las diferentes escalas de los datos (ej. comparar el PIB con la edad media), todas las variables fueron **estandarizadas (escalado Z-score)**. Este proceso las transforma para que todas tengan una media de 0 y una desviación estándar de 1, asegurando que cada variable contribuya de forma equitativa al cálculo de las distancias.
- **Determinación del Número Óptimo de Clústeres (k):** Para determinar el número óptimo de clústeres, se evaluaron dos métricas complementarias para un rango de k de 2 a 8, como se muestra en la Fig. 13:
    1. **Método del Codo (Inertia):** La primera métrica, la Inertia, mide la suma de las distancias al cuadrado de cada punto a su centroide, sirviendo como un indicador de la cohesión interna de los clústeres. Aunque se busca minimizar la Inertia, se observa en el gráfico izquierdo que la curva decrece suavemente sin un "codo" claramente definido, lo que hace su interpretación bastante subjetiva.

2. **Método de la Silueta (Silhouette Score):** La segunda métrica es la Puntuación de Silueta, un indicador robusto que evalúa tanto la cohesión (qué tan similares son los puntos dentro de un clúster) como la separación (qué tan distintos son los clústeres entre sí). El objetivo es maximizar esta puntuación, que varía de -1 a 1.

Como se aprecia en el gráfico derecho, la Puntuación de Silueta, alcanza un máximo claro en $k = 2$, siendo este el número óptimo elegido de clústeres para la segmentación de los países.

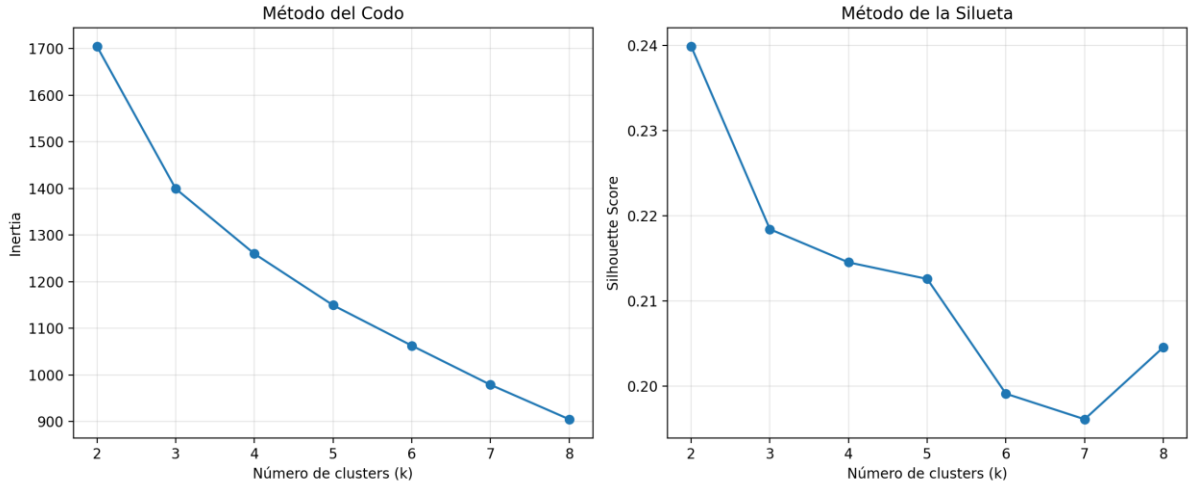

Fig. 13 Cálculo de k: Método del Codo & Método de la Silueta

5.1.2 Resultados: Interpretación de los Perfiles de Clústeres (análisis de centroides con heatmap, PCA y radar)

Una vez determinado el número de grupos, se procedió a analizar las características de cada clúster para definir sus perfiles.

Para comenzar, el Análisis de Componentes Principales (PCA) de la Fig. 14 confirma que ambos clústeres están **claramente definidos y separados** en el espacio, dando coherencia a las agrupaciones.

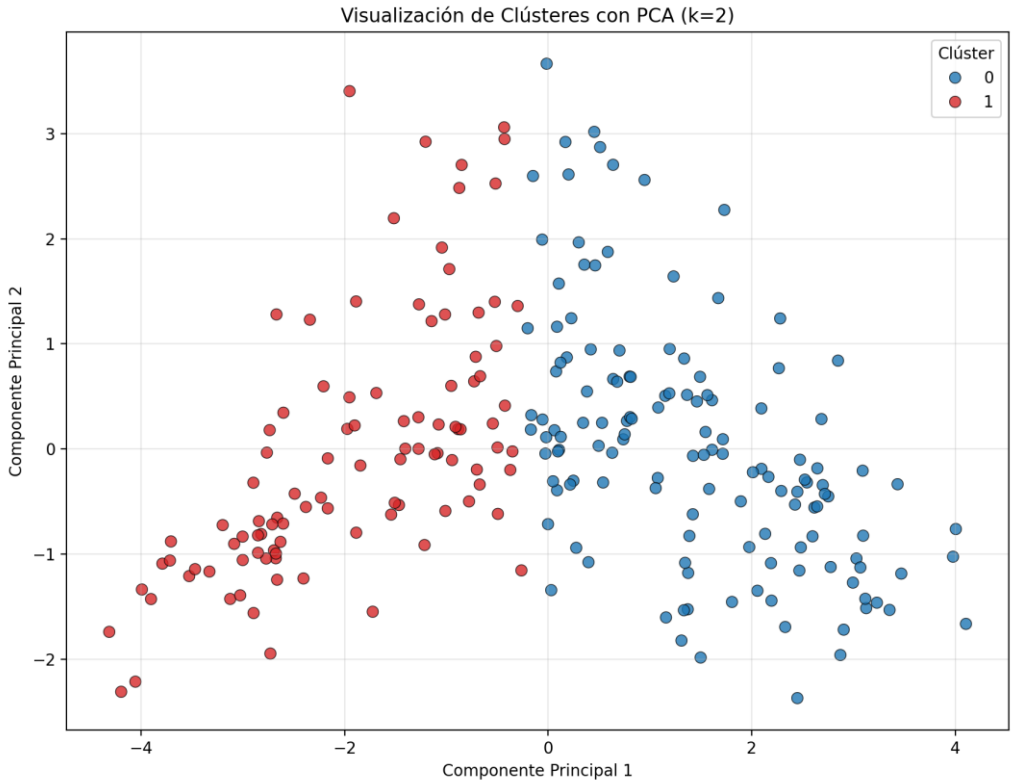

Fig. 14 Clústeres - Análisis de Componentes Principales (PCA)

Para estudiar qué propiedades posee cada clúster, se utilizaron los centroides estandarizados, que representan el "valor promedio" de cada variable para los países de ese grupo.

Las Fig. 15 y Fig. 16 muestran una clara diferencia entre los dos grupos. Los valores positivos (azul) indican que el clúster está por encima de la media en esa variable, mientras que los negativos (rojo) indican que está por debajo. A partir de este análisis, se definen dos perfiles:

- **Clúster 0 (Azul): Perfil de Países Desarrollados.** Este grupo se caracteriza por tener valores significativamente por encima de la media en indicadores de desarrollo: mayor **PIB per cápita**, poblaciones más **envejecidas**, más **camas de hospital** y una **mayor cobertura de vacunación**. Consecuentemente, también son los países que han reportado un mayor impacto acumulado en términos de casos y muertes por millón.
- **Clúster 1 (Rojo): Perfil de Países en Desarrollo.** Este clúster es la imagen inversa del anterior. Agrupa a los países con indicadores de desarrollo por debajo de la media, como un **menor PIB per cápita** y **edad media**. A su vez, presenta el nivel más alto de **pobreza extrema**. En cuanto al impacto de la pandemia, reportaron un total de casos y muertes por millón considerablemente inferior al promedio.

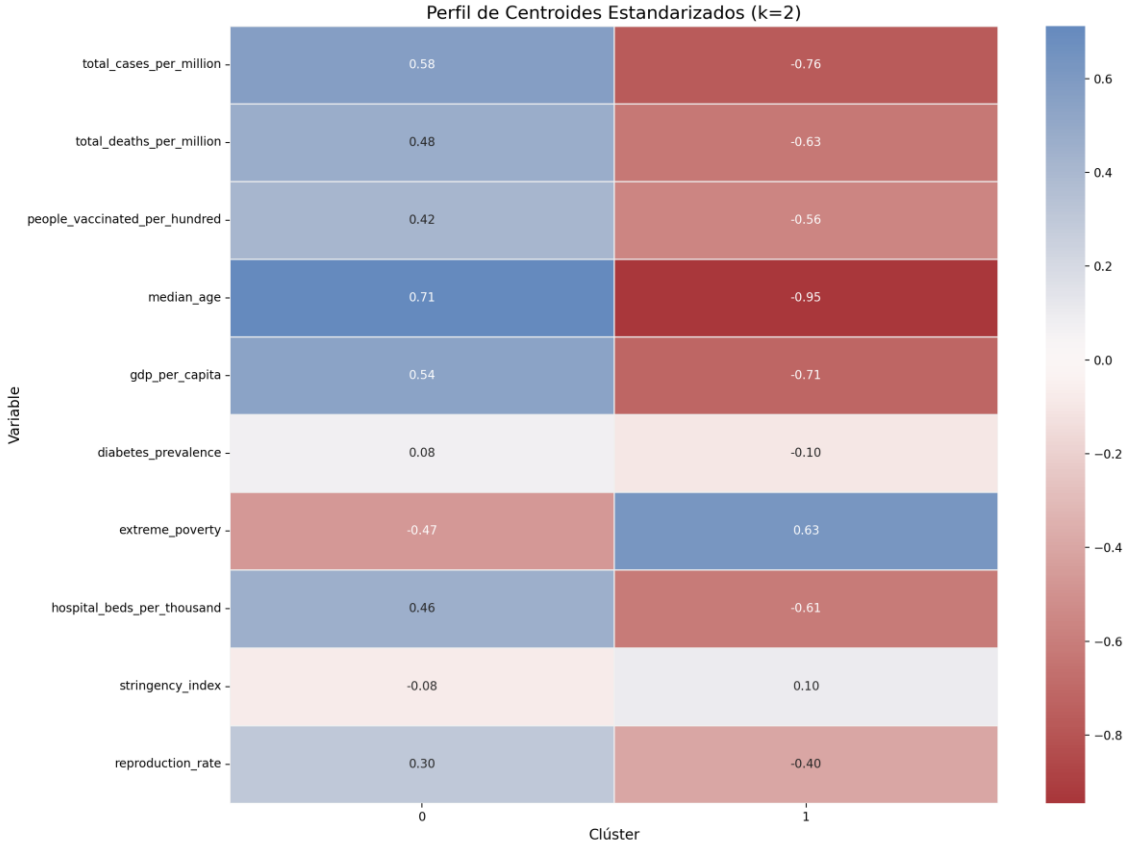

Fig. 15 Clústeres - Centroides estandarizados de cada variable

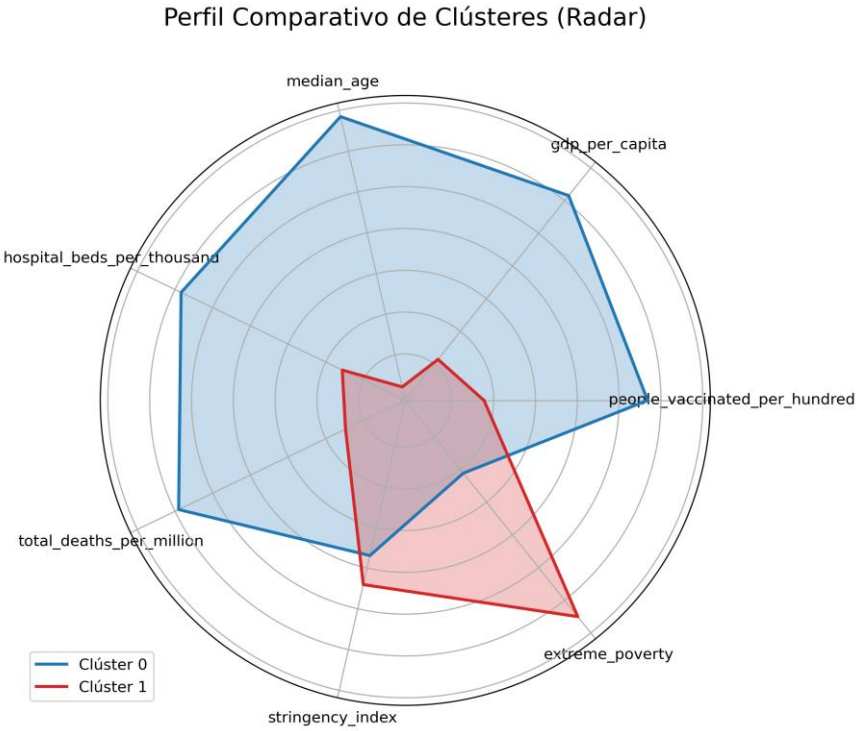

Fig. 16 Clústeres – Radar de cada variable

5.1.3 Visualización Geográfica de los Clústeres

Para completar el análisis de perfiles, se ha proyectado la asignación de cada país a su clúster correspondiente en un mapa mundial (Fig. 17). Esta visualización permite identificar patrones geográficos.

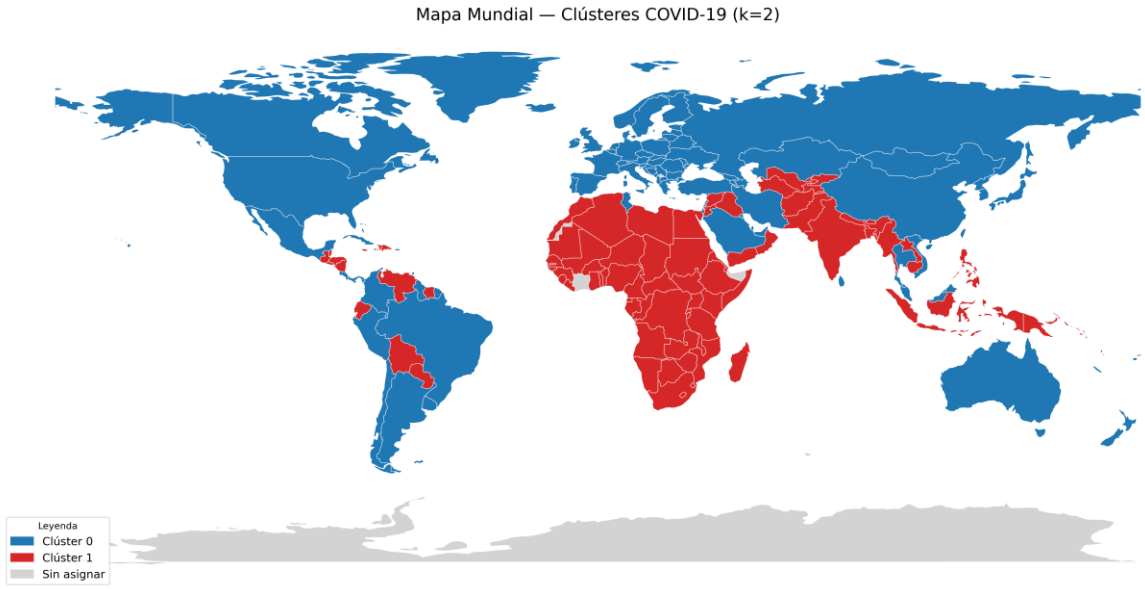

Fig. 17 Clústeres – Mapa Mundial con clústeres asignados por país

El mapa confirma visualmente las conclusiones extraídas del análisis de centroides:

- **Clara División Geográfica:** La distribución de los clústeres no es aleatoria, sino que sigue una marcada división geográfica y económica. El **Clúster 0 (azul)**, que corresponde a los países con indicadores de mayor desarrollo, predomina claramente en **Norteamérica, Europa y Oceanía**.
- **Concentración del Clúster 1:** Por otro lado, el **Clúster 1 (rojo)**, que agrupa a los países con menor PIB, poblaciones más jóvenes y mayor pobreza extrema, se concentra de forma casi total en los continentes de **África y Asia**.
- **Región Mixta: América Latina** presenta una región mixta, con países pertenecientes a ambos clústeres, reflejando la heterogeneidad socioeconómica del continente.

5.2 DETECCIÓN DE ANOMALÍAS EN SERIES TEMPORALES

Finalizado el apartado de clustering, se pasa a un apartado bastante interesante, donde se detalla un método implementado para identificar y visualizar anomalías en los reportes, como picos o caídas inusuales. Estos eventos pueden indicar tanto problemas en la calidad o frecuencia del reporte de datos como eventos reales extremos.

5.2.1 Metodología: Cálculo de z-score sobre residuos de media móvil

Para detectar estas desviaciones de forma sistemática, se implementó un método estadístico basado en el análisis de residuos. El proceso, aplicado a las series semanales de nuevos casos y muertes por millón de cada país, consistió en los siguientes pasos:

1. **Establecimiento de una Línea Base:** Se calculó una **media móvil de 8 semanas** para cada serie temporal. Esta media móvil actúa como una línea de tendencia suavizada que representa el comportamiento "esperado" de los datos en un momento dado.
2. **Cálculo de Residuos:** Para cada semana, se calculó el **residuo**, que es la diferencia entre el valor real observado y el valor de la media móvil. Un residuo grande indica que el dato de esa semana se desvía significativamente de su tendencia reciente.
3. **Normalización con Z-score:** Finalmente, se estandarizaron estos residuos calculando su **puntuación Z (z-score)**. Esta métrica indica cuántas desviaciones estándar por encima o por debajo de la media se encuentra un residuo.
4. **Definición de Anomalía:** Se definió una **anomalía** como cualquier semana cuyo z-score tuviera un **valor absoluto mayor a 3**. Este evento es estadísticamente muy improbable, convirtiéndose así en un evento anómalo.

5.2.2 Resultados: Visualización de semanas anómalas por país (heatmaps y mapas de outliers)

El método se aplicó a las series de casos y muertes, limitando el análisis hasta enero de 2024 para centrarse en el periodo de mayor actividad pandémica. Los heatmaps y mapas a continuación permiten visualizar qué países y en qué momentos presentaron la mayor cantidad de estas semanas anómalas.

5.2.2.1 *Anomalías en Nuevos Casos por millón*

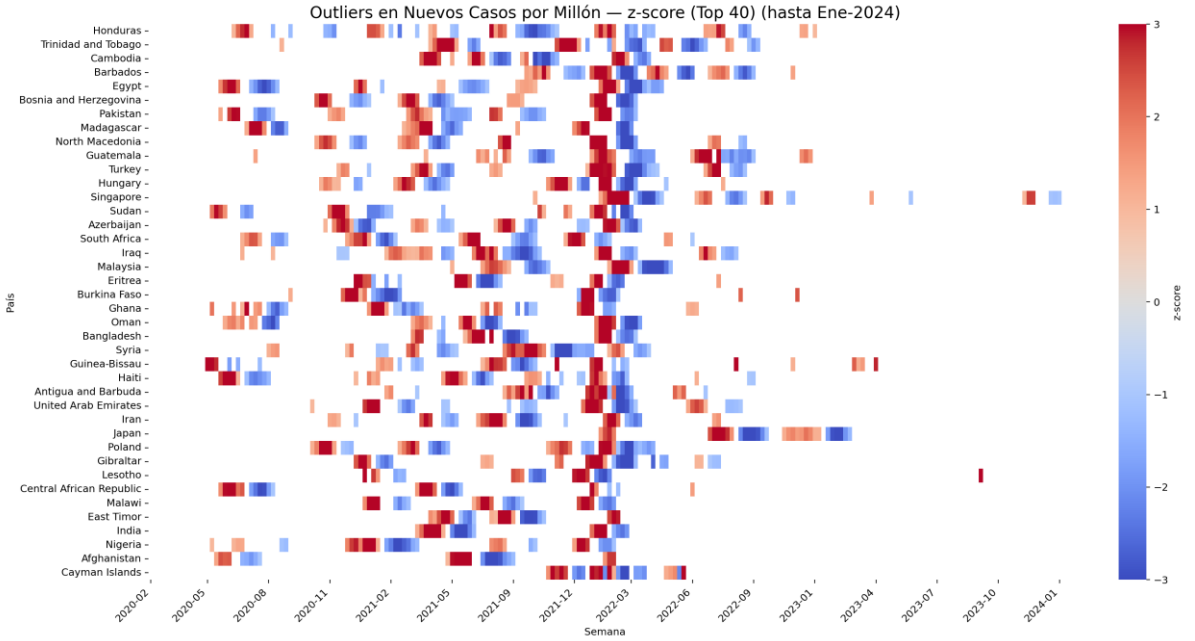

Fig. 18 Outliers en reportes de nuevos casos por millón – Top 40 (hasta Ene-2024)

En la Fig. 18 se observa:

- Una gran concentración de **anomalías positivas** (picos, en rojo intenso) durante el **primer trimestre de 2022**. Esto puede justificarse con la llegada de la **variante Ómicron**, puesto que fue bastante contagiosa, rompiendo récords en todo el mundo.
- Por otro lado, las **anomalías negativas** (caídas, en azul intenso), suelen indicar correcciones de datos o cambios en la frecuencia de reporte, donde tras un volcado de datos se suceden periodos con reportes muy bajos. Por ejemplo, el **segundo trimestre de 2022.**

El mapa de **picos de casos** (Fig. 19) muestra que la proporción de semanas con aumentos anómalos fue especialmente alta en **África** (Níger, Libia) y el sur de **Asia** (India, Nueva Guinea). Esto sugiere que, más allá de las olas globales, estas regiones experimentaron episodios de crecimiento de contagios extremadamente rápidos en comparación con sus tendencias locales.

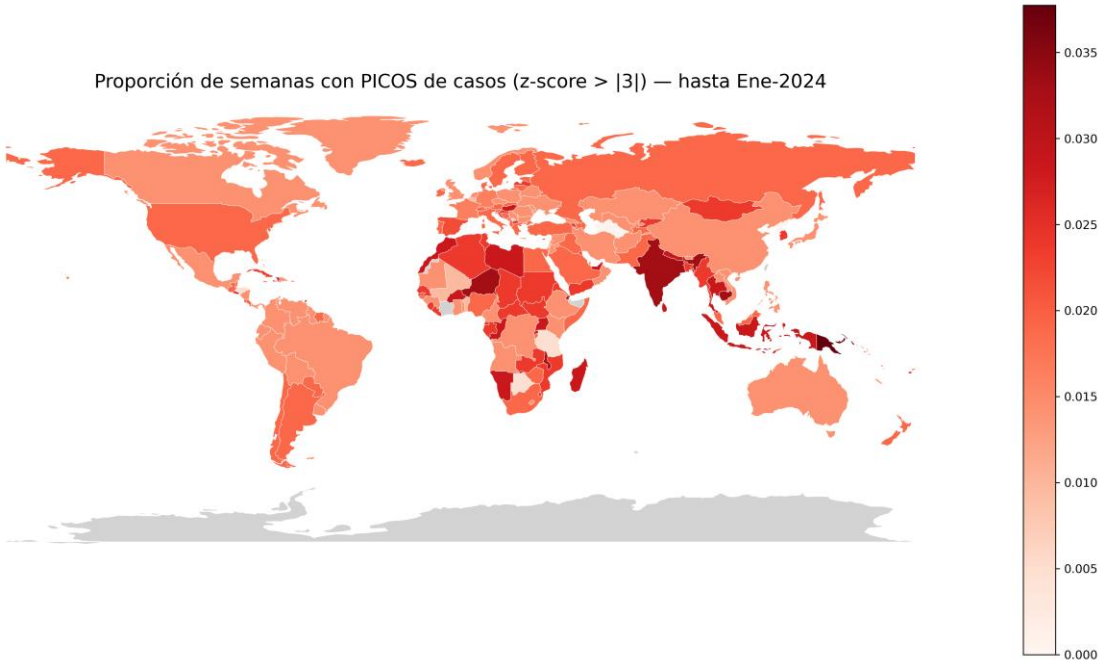

Fig. 19 Outliers - Mapa con Picos de Casos (hasta Ene-2024)

Por otro lado, el mapa de **caídas anómalas de casos** (Fig. 20) muestra una mayor incidencia sobre todo en el **norte de Europa** (Noruega, Suecia). Esto es un indicador de irregularidades en el reporte de datos, como la acumulación de casos en una semana seguida de semanas con reportes muy bajos o nulos.

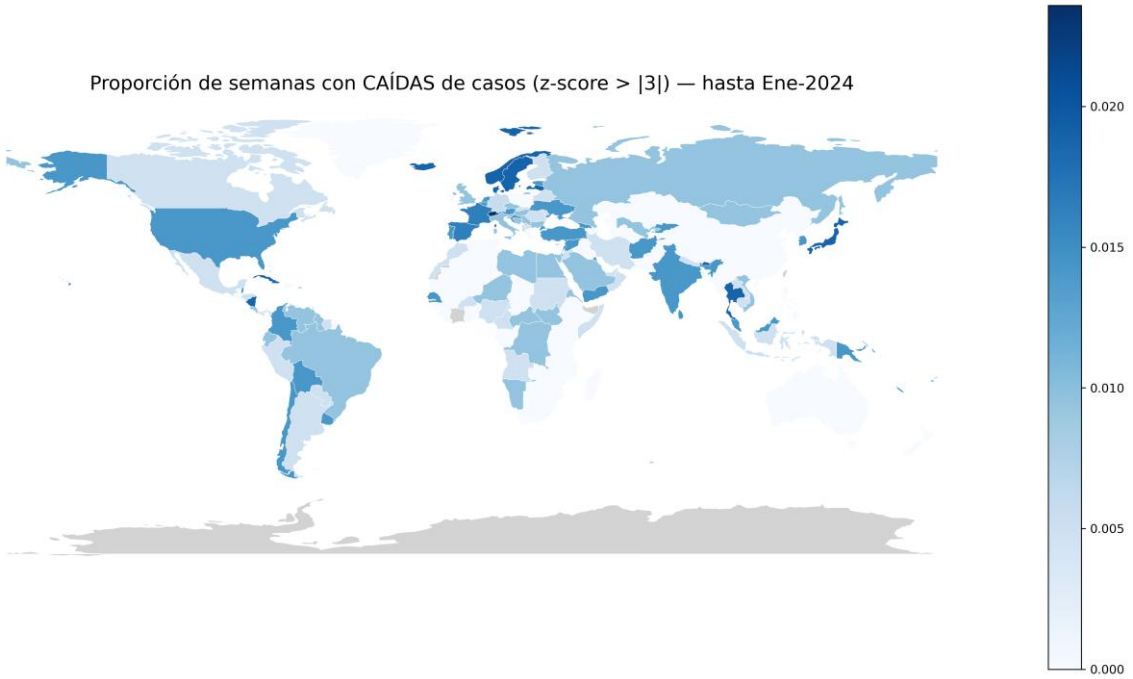

Fig. 20 Outliers - Mapa con Caídas de Casos (hasta Ene-2024)

5.2.2.2 *Anomalías en Nuevas muertes por millón*

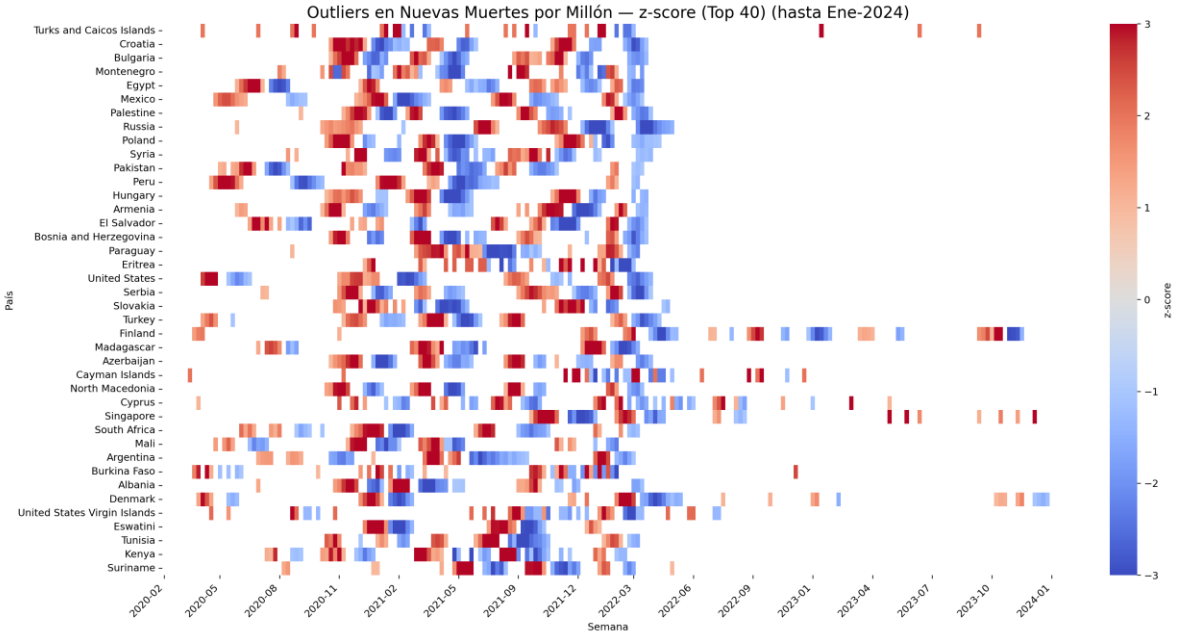

Fig. 21 Outliers en reportes de nuevas muertes por millón – Top 40 (hasta Ene-2024)

El heatmap de **muertes por millón** (Fig. 21) muestra un patrón temporal diferente al de los casos. Las anomalías positivas están repartidas temporalmente. Por otro lado, para las anomalías negativas, vemos además una tendencia en el segundo trimestre de 2022, pudiendo estar reflejado de nuevo por la poca mortalidad de la variable Ómicron.

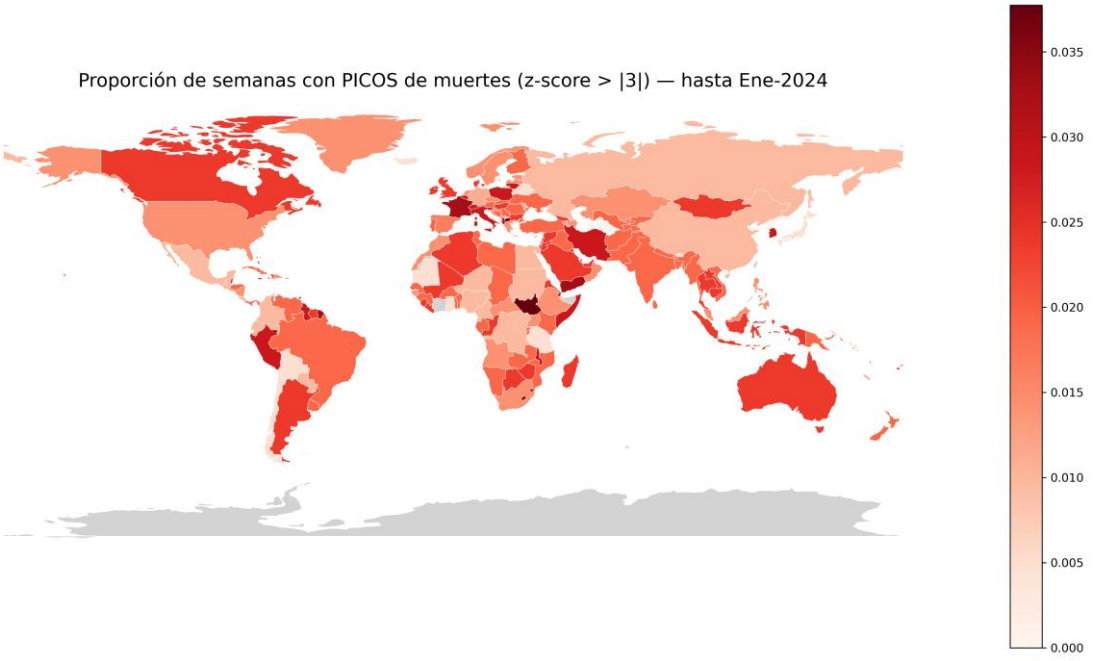

Fig. 22 Outliers - Mapa con Picos de Muertes (hasta Ene-2024)

Geográficamente, el mapa de picos de mortalidad (Fig. 22) detecta picos severos en varios puntos: **centro de África** (Sudán), América (Canadá, Perú, Argentina), **Oceanía,** Francia y parte de **Asia** (Irán).

Finalmente, el mapa de **caídas anómalas de muertes** (Fig. 23) presenta una clara anomalía en Rusia.

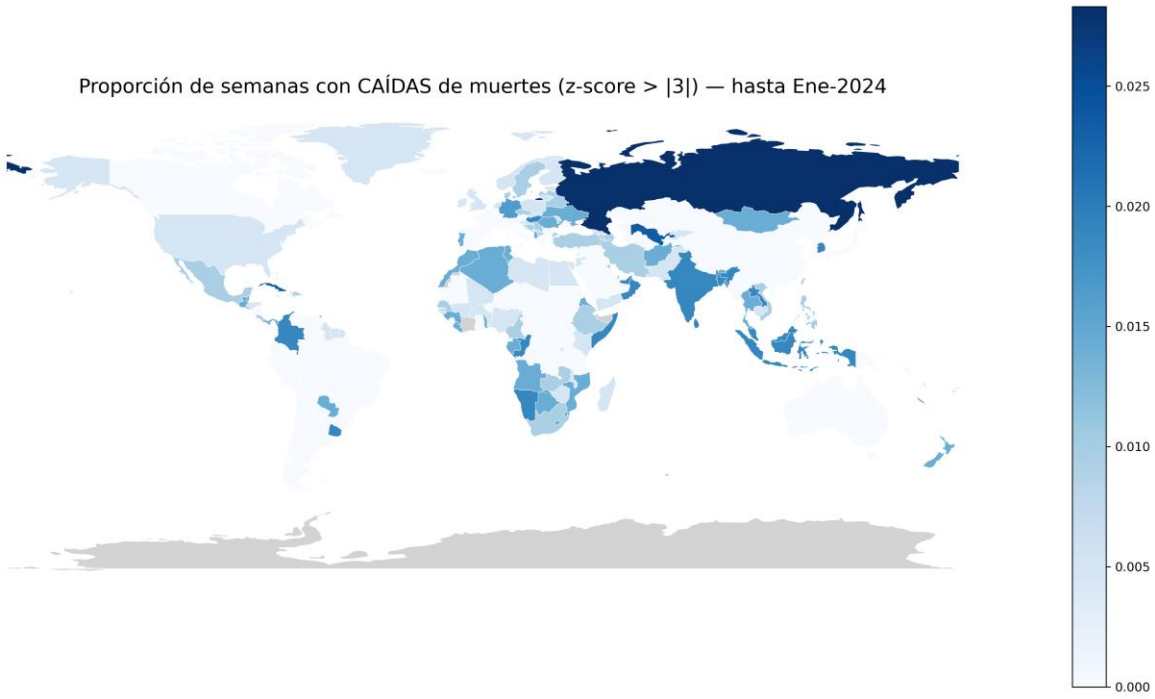

Fig. 23 Outliers - Mapa con Caídas de Muertes (hasta Ene-2024)

5.3 MODELOS DE REGRESIÓN PARA LA ESTIMACIÓN DE EFECTOS

Para ir más allá de las correlaciones y entender qué factores tuvieron un impacto estadísticamente significativo en la mortalidad, se construyeron varios modelos de regresión por Mínimos Cuadrados Ordinarios (OLS). Estos modelos permiten aislar el efecto de cada variable mientras se controlan las demás, ofreciendo una visión más precisa de su influencia.

5.3.1 Modelo 1: Factores Asociados a la Mortalidad

El primer modelo se diseñó para estimar el efecto de un conjunto de variables clave sobre la tasa de mortalidad semanal. Se utilizó un modelo *Pooled OLS* sobre el dataset de panel weekly_model_ready.csv, que agrupa las observaciones de todos los países y semanas.

- **Variable Dependiente:** new_deaths_per_million (nuevas muertes por millón de habitantes en una semana).
- **Variables Independientes:** Se incluyeron variables de gestión pandémica, epidemiológicas y estructurales. Es importante destacar la creación de la variable pvax_lag3. Esta variable representa el porcentaje de población vacunada (people_vaccinated_per_hundred), pero con un **retardo (lag) de 3 semanas**, pensado para observar el efecto de la vacuna en el sistema inmunitario.

Resultados de OLS							
Dep. Variable:	**new_deaths_per_million**			**R-squared:**		**0.210**	
Model:	**OLS**			**Adj. R-squared:**		**0.210**	
Method:	**Least Squares**			**F-statistic:**		**639.5**	
Nº. Observations:	**33568**			**AIC:**		**2.658E+05**	
Df Residuals:	**33561**			**BIC:**		**2.659E+05**	
variable	**coef**	**std err**	**z**	**P>|z|**	**ci_low**	**ci_high**	**signif**
const	-12.0013	0.296	-40.5469	0.000	-12.5814	-11.4212	*
pvax_lag3	-0.0788	0.0035	-22.8475	0.000	-0.0856	-0.0721	*
stringency_index	0.2282	0.0055	41.4009	0.000	0.2174	0.239	*
reproduction_rate	-1.3426	0.1334	-10.0652	0.000	-1.604	-1.0811	*
median_age	0.5645	0.0178	31.7712	0.000	0.5297	0.5993	*
gdp_per_capita	0.0000	0.0000	-5.3272	0.000	0.0000	0.0000	*
hospital_beds_per_thousand	-0.2791	0.0472	-5.9136	0.000	-0.3716	-0.1866	*

Tabla 1 Modelo OLS para nuevas muertes por millón

La tabla de resultados (Tabla 1) muestra que el modelo tiene una capacidad explicativa moderada (Adj. $R-squared = 0.210$) y que **todas las variables incluidas son estadísticamente significativas** (P>|z| es 0.000 para todas), lo que indica que contribuyen a explicar la variación en la mortalidad.

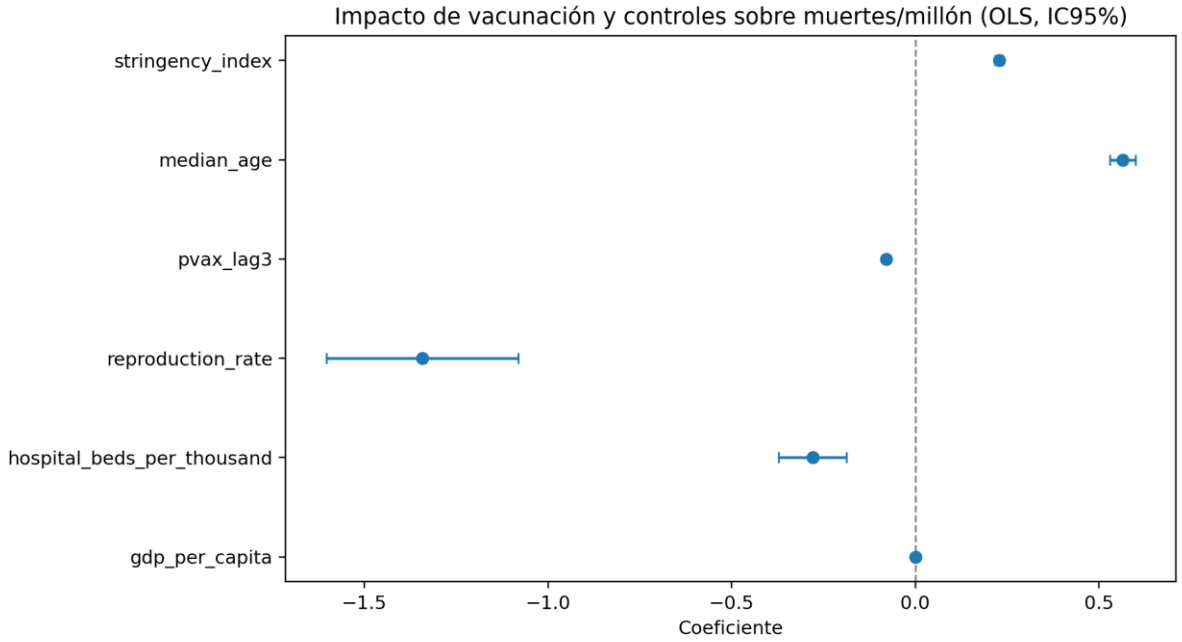

Fig. 24 OLS – Coeficiente de variables de impacto sobre muertes por millón

El gráfico de coeficientes (Fig. 24) permite interpretar visualmente el impacto y la magnitud de cada factor:

- Factores con Efecto Protector (Reducen la mortalidad):

  Estos factores están asociados a una disminución en la tasa de muertes por millón, manteniendo el resto de las variables constantes:

  - **Vacunación (pvax_lag3):** El coeficiente de **-0.0788** indica que por cada punto porcentual que aumenta la población vacunada (con un retardo de 3 semanas), la tasa de muertes por millón disminuye en **0.0788**. Esto es un indicador del efecto protector de la vacunación.
  - **Camas de hospital:** Con un coeficiente de **-0.2791**, esta variable muestra que una mayor capacidad hospitalaria está asociada a una menor mortalidad.
  - **Tasa de Reproducción:** Su coeficiente de **-1.3426** es el más fuerte en magnitud, pero contraintuitivo. Esto puede deberse a que los mayores picos de transmisión coincidieron con la variante Ómicron, que fue menos letal.

  - **PIB per cápita:** El coeficiente es negativo, pero extremadamente pequeño (**-0.00001773**). Aunque estadísticamente significativo, su impacto práctico es ínfimo.

- Factores con Efecto de Riesgo (Aumentan la mortalidad):
  - **Edad Media:** Es el factor de riesgo más importante, con un coeficiente de **0.5645**. Por cada año que aumenta la edad media de un país, la tasa de muertes por millón se incrementa en más de medio punto, confirmando la alta vulnerabilidad de las poblaciones envejecidas.
  - **Índice de Restricciones:** Su coeficiente positivo (**0.2282**) es un claro ejemplo de **causalidad inversa**. No significa que las restricciones causaran más muertes, sino que los gobiernos aplicaban las medidas más severas precisamente **cuando la mortalidad estaba aumentando**.

5.3.2 Modelo 2: Creación de un Índice de Eficiencia Sanitaria

Para evaluar el desempeño de la gestión sanitaria de cada país, se construyó un segundo modelo OLS. A diferencia del modelo anterior, este se ejecutó sobre un DataFrame agregado, con una única fila por país que resume sus indicadores medianos a lo largo de toda la pandemia.

El objetivo de este modelo era predecir la tasa de mortalidad mediana (***deaths_pm***) que se "esperaría" para un país, dadas sus características estructurales (***PIB, edad media, camas de hospital, etc***.) y de gestión pandémica (***vacunación, restricciones***).

La clave de este análisis reside en el estudio de los **residuos** del modelo, que representan la diferencia entre la mortalidad real observada y la mortalidad predicha por el modelo.

- Un **residuo negativo** significa que un país tuvo ***menos*** muertes de las esperadas, lo que sugiere una **gestión eficiente**.
- Un **residuo positivo** significa que un país tuvo ***más*** muertes de las esperadas, sugiriendo una **gestión menos eficiente**.

A partir de estos residuos, se construyó un **"Índice de Eficiencia"** estandarizado (z-score), donde una puntuación más alta indica un mejor desempeño.

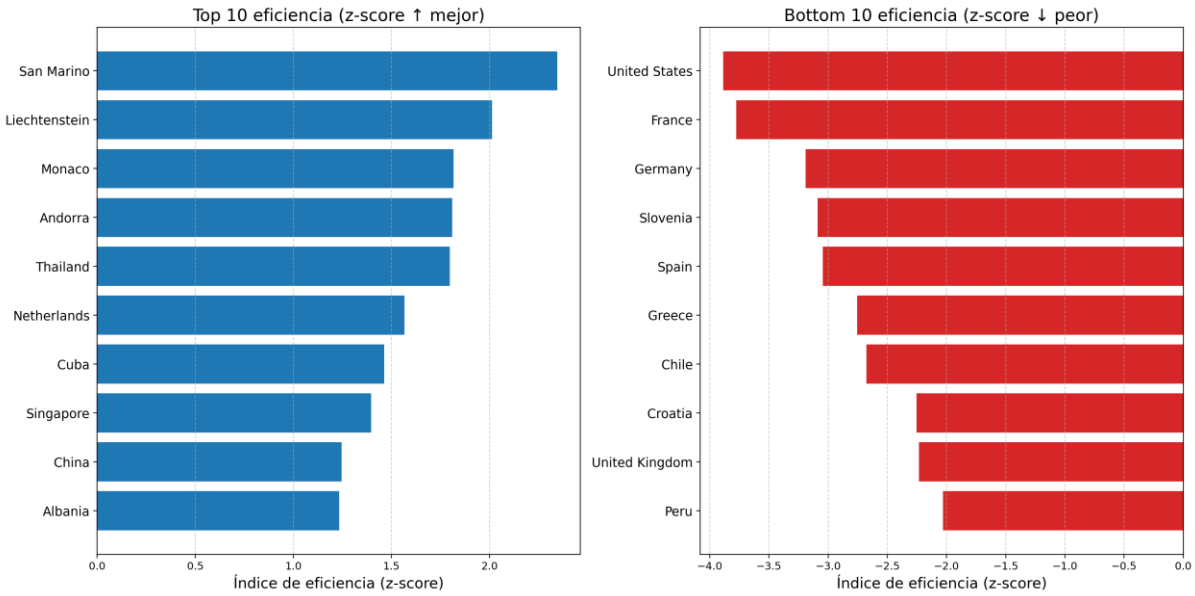

Fig. 25 Índice de Eficiencia Sanitaria - Top 10 y Bottom 10

El gráfico de ranking (Fig. 25**¡Error! No se encuentra el origen de la referencia.**) muestra l os 10 países con el mejor y peor desempeño según este índice:

- **Top 10 (Más eficientes):** El grupo de países con mayor eficiencia es heterogéneo y no responde a un único perfil. Se pueden identificar varios patrones de éxito:
    - **Microestados europeos ricos:** Países como **San Marino, Liechtenstein, Mónaco y Andorra** ocupan las primeras posiciones, donde su pequeño tamaño y alta renta pudieron facilitar una gestión sanitaria y fronteriza muy ágil.
    - **Países con sistemas de salud pública robustos:** Naciones como **Singapur, Tailandia y Cuba** también destacan, indicando que una fuerte organización sanitaria fue una estrategia eficiente.
    - **Estrategia de supresión:** La presencia de **China** se explica por su política de "COVID Cero" mantenida durante gran parte del periodo.
- **Bottom 10 (Menos eficientes):** Sorprendentemente, este grupo está dominado por **grandes economías de Europa Occidental y Norteamérica**, como **Estados Unidos, Francia, Alemania, España y Reino Unido**. Estos índices revelan que, a pesar de sus recursos y ventajas estructurales, su tasa de mortalidad fue significativamente más alta de lo que el modelo predecía para ellos.

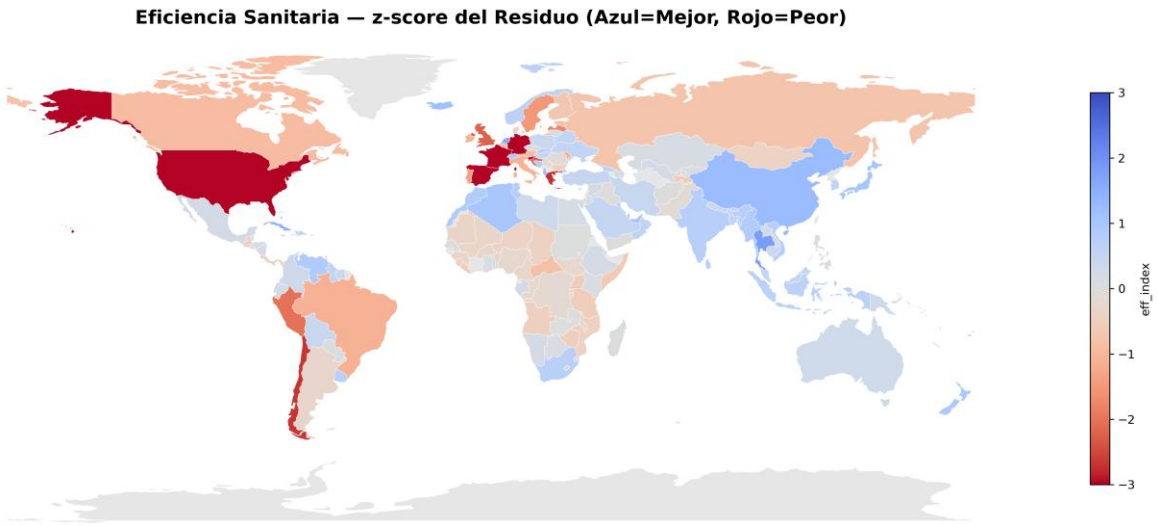

Fig. 26 Índice de Eficiencia Sanitaria – Mapa mundial z-score del Residuo

El mapa de Eficiencia Sanitaria (Fig. 26) permite visualizar estos resultados geográficamente. Se confirma la **concentración de un alto desempeño (azul)** en la región de Asia-Pacífico y una **baja eficiencia (rojo)** en el **oeste de Europa** y **Estados Unidos**.

5.3.3 Análisis de Efectos de Interacción: Vacunación y Edad

Para profundizar en la relación entre la vacunación y la mortalidad, se investigó si el efecto protector de las vacunas era constante en todos los países o si variaba en función de la estructura demográfica. Se planteó la hipótesis de que la vacunación podría ser especialmente efectiva en países con poblaciones más envejecidas. Para ello, se realizaron dos análisis complementarios.

5.3.3.1 *Modelo con Término de Interacción*

Primero, se construyó un modelo OLS que incluía un **término de interacción (pvaxXage)**, resultado de multiplicar las variables estandarizadas de vacunación (**pvax_lag3_std**) y edad media (**median_age_std**).

Resultados de OLS			
Dep. Variable:	**new_deaths_per_million**	**R-squared:**	**0.223**
Model:	**OLS**	**Adj. R-squared:**	**0.223**
Method:	**Least Squares**	**F-statistic:**	**755.5**
Nº. Observations:	**33568**	**AIC:**	**2.653E+05**
Df Residuals:	**33562**	**BIC:**	**2.653E+05**

variable	coef	std err	z	P>|z|	ci_low	ci_high	signif
const	5.5078	0.097	56.533	0.000	5.317	5.699	*
pvax_lag3_std	-2.5754	0.113	-22.878	0.000	-2.796	-2.355	*
median_age_std	4.3264	0.098	44.167	0.000	4.134	4.518	*
pvaxXage	-1.9523	0.109	-17.881	0.000	-2.166	-1.738	*
stringency_index_std	4.1273	0.107	38.677	0.000	3.918	4.336	*
reproduction_rate_std	-0.2540	0.057	-4.450	0.000	-0.366	-0.142	*

Tabla 2 *Modelo Interacción: OLS para nuevas muertes por millón*

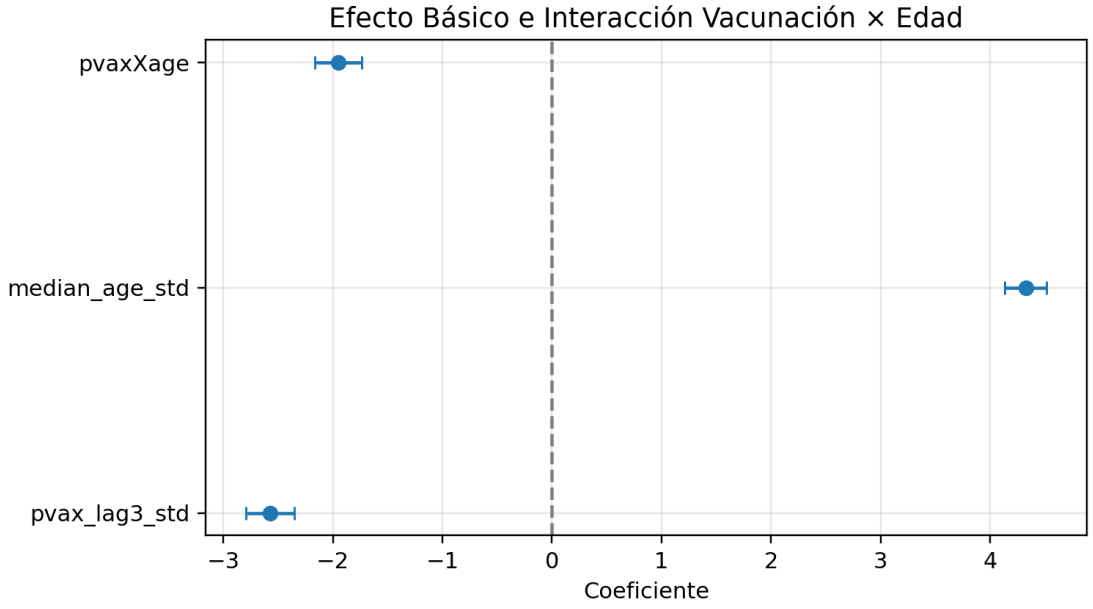

Fig. 27 OLS – Efecto básico y vacunación de impacto sobre muertes por millón

El análisis de los resultados de la Tabla 2 y Fig. 27 es concluyente:

- **Efectos Principales:** Tanto la **vacunación** (pvax_lag3_std, coef: -2.58) como la **edad media** (median_age_std, coef: 4.33) son altamente significativas. Tal como se esperaba, la vacunación muestra un fuerte efecto protector (coeficiente negativo), mientras que una mayor edad media se asocia con una mayor mortalidad (coeficiente positivo).

- **Efecto de Interacción** (pvaxXage): El término de interacción es estadísticamente significativo y presenta un coeficiente negativo de **-1.95**. Esto confirma la hipótesis: el efecto protector de la vacunación se vuelve **aún más fuerte** a medida que aumenta la edad media de un país. En otras palabras, la vacunación fue más decisiva para reducir la mortalidad en las naciones demográficamente más envejecidas.

5.3.3.2 *Análisis por Cuartiles de Edad*

Para visualizar este descubrimiento de una manera más intuitiva, se realizó un segundo análisis. Se dividieron los países en cuatro grupos (cuartiles) según su edad media, desde los más jóvenes (Q1) hasta los más envejecidos (Q4), y se ejecutó una regresión OLS para cada grupo.

Resultados de OLS por cuartiles

age_q	coef	ci_low	ci_high
Q1 (más joven)	-0.0086	-0.0119	-0.0052
Q2	-0.0266	-0.0340	-0.0192
Q3	-0.1190	-0.1381	-0.0999
Q4 (más mayor)	-0.0888	-0.1093	-0.0682

Tabla 3 Modelo OLS: vacunación por cuartiles de edad media

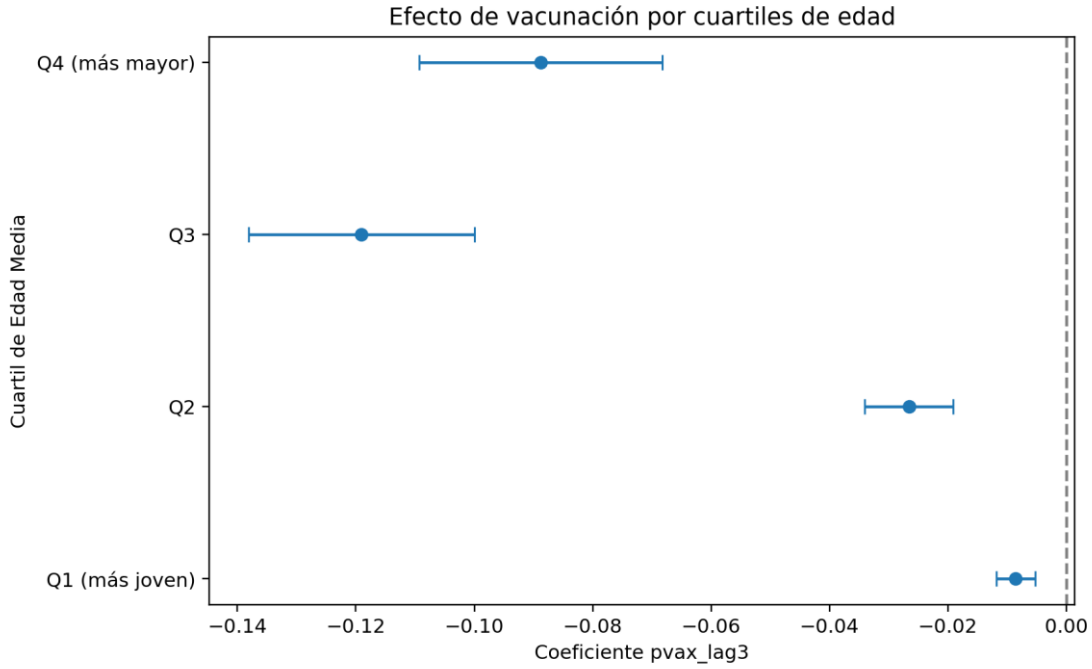

Fig. 28 OLS – Efecto vacunación por cuartiles de edad media

Las Tabla 3 y Fig. 28 muestran el coeficiente de la variable de vacunación (pvax_lag3) para cada uno de los cuatro cuartiles de edad. Los resultados confirman que el efecto de la vacunación depende fuertemente de la demografía, pero revelan una **relación no lineal** interesante:

- Para el cuartil de países más jóvenes (Q1), el efecto de la vacunación, aunque protector (coef: **-0.0086**), es muy pequeño.
- El efecto protector aumenta drásticamente en el segundo (Q2, coef: **-0.027**) y, sobre todo, en el **tercer cuartil (Q3)**, donde alcanza su **máxima efectividad** con un coeficiente de **-0.119**.
- Curiosamente, para el grupo de países más envejecidos (Q4), el efecto protector, aunque sigue siendo muy fuerte (coef: **-0.089**), es **ligeramente menor que en el Q3**. Esto podría sugerir que, en poblaciones extremadamente envejecidas, otros factores de riesgo (como la fragilidad anciana) tienen un peso tan dominante que el impacto marginal de la vacunación, si bien sigue siendo crucial, es levemente inferior en comparación con el grupo anterior.

6 RECOMENDACIONES Y LÍNEAS FUTURAS

Este estudio ofrece varias recomendaciones aplicables para la gestión de futuras crisis sanitarias. La segmentación de países en perfiles homogéneos permite, por ejemplo, abandonar las estrategias únicas y diseñar políticas diferenciadas que optimicen la asignación de recursos según las vulnerabilidades específicas de cada arquetipo de nación. Asimismo, la confirmación empírica de que el efecto protector de la vacunación se magnifica en poblaciones envejecidas subraya la necesidad de priorizar a estos grupos demográficos para minimizar la mortalidad. Finalmente, el Índice de Eficiencia desarrollado se postula como una valiosa herramienta de benchmarking, permitiendo a los gobiernos evaluar su desempeño en comparación con países de características similares para identificar y adoptar las estrategias más exitosas.

Como futuras mejoras, se podrían abarcar varias líneas de profundización. Sería de gran interés replicar el estudio a un nivel subnacional, permitiendo analizar la heterogeneidad del impacto de la pandemia dentro de las propias naciones. Por último, la inclusión de nuevas

variables, como datos de movilidad poblacional o la prevalencia de variantes específicas del virus, podría enriquecer los modelos y ofrecer una visión aún más completa de la pandemia.

7 CONCLUSIÓN

A lo largo de esta memoria, se ha abordado el complejo análisis del denso conjunto de datos generado por la pandemia de COVID-19 para extraer patrones, evaluar factores y generar conocimiento aplicable. Partiendo del dataset de Our World in Data, se ha desarrollado un pipeline de datos que incluyó fases de limpieza, una robusta imputación de valores nulos y la agregación de los datos a una escala semanal consistente. Sobre esta base preparada, se aplicó una metodología mixta que combinó el análisis exploratorio, el clustering no supervisado (K-Means) y el modelado de regresión (OLS) para cuantificar el impacto de diferentes factores sobre la mortalidad.

Los principales hallazgos confirman la segmentación del mundo en dos grandes perfiles de respuesta, donde el desarrollo socioeconómico y la demografía resultaron ser clave. Se ha demostrado empíricamente el fuerte efecto protector de la vacunación, destacando que su impacto es significativamente mayor en países con poblaciones más envejecidas. Adicionalmente, se ha propuesto un Índice de Eficiencia Sanitaria original, cuyos resultados revelan que la riqueza de un país no fue garantía de una gestión eficiente, mostrando que grandes potencias de Europa Occidental y Norteamérica tuvieron un desempeño inferior al que sus capacidades predecían.

En definitiva, este trabajo puede ser útil para comprender la crisis sufrida a nivel global por la pandemia y las decisiones de salud pública tomada por cada nación, sirviendo como base para gestionar desde un punto de vista más sólido futuras crisis.

8 REFERENCIAS

[1] Lab, Global Change Data, «Dataset de la COVID-19,» [En línea]. Available: https://catalog.ourworldindata.org/garden/covid/latest/compact/compact.csv.

[2] J. Sundström, «world.geo.json,» 2016. [En línea]. Available: https://github.com/johan/world.geo.json.

9 ANEXO I: CÓDIGO

El código utilizado puede encontrarse en este repositorio de ***GitHub:***

https://github.com/Jose-Manuel-Segovia-Valdivia/TFM_COVID-19_ANALYSIS.git

Se ha trabajado con 4 scripts .py de Python ordenados correctamente.