CIMAT

MTY Modelando el siglo XXI

Cómputo estadístico

Proyecto

Anselmo Daniel Suarez Muñoz
Armando Salinas Lorenzana

5 de diciembre de 2024

Índice

1. Introducción

La transición hacia las energías renovables se ha convertido en una de las prioridades más apremiantes del siglo XXI, impulsada por la necesidad de combatir el cambio climático, reducir la dependencia de los combustibles fósiles y garantizar un suministro energético sostenible para una población mundial en constante crecimiento. Las fuentes de energía renovable, como la solar, eólica, hidroeléctrica, geotérmica y biomasa, no solo representan alternativas más limpias y seguras, sino que también ofrecen una oportunidad para diversificar las economías, fomentar la innovación tecnológica y mejorar la resiliencia energética global.

El interés por las energías renovables ha crecido exponencialmente en las últimas décadas, impulsado por avances tecnológicos que han reducido significativamente los costos de instalación y operación. Por ejemplo, el costo nivelado de la energía solar fotovoltaica ha disminuido en más del 80 % desde 2010, mientras que la capacidad instalada de generación eólica ha aumentado en un 300 % en el mismo período. Estas tendencias reflejan no solo la viabilidad económica de las energías renovables, sino también su capacidad para competir directamente con las fuentes convencionales de energía.

A nivel mundial, las políticas gubernamentales, los acuerdos internacionales y los compromisos de descarbonización están desempeñando un papel clave en la promoción de las energías renovables. Iniciativas como el Acuerdo de París de 2015 han establecido metas claras para limitar el calentamiento global a menos de 2 °C por encima de los niveles preindustriales, incentivando a los países a adoptar medidas que fomenten la transición energética. Asimismo, programas de subsidios, incentivos fiscales y la creciente demanda de electricidad limpia están transformando los mercados energéticos tanto en países desarrollados como en desarrollo.

No obstante, la adopción de energías renovables no ha sido uniforme en todo el mundo. Mientras que naciones como Alemania, China y Estados Unidos lideran en términos de capacidad instalada y producción, otras regiones enfrentan desafíos significativos debido a limitaciones económicas, tecnológicas y de infraestructura. En particular, los países en desarrollo a menudo carecen de los recursos financieros necesarios para implementar proyectos a gran escala, aunque paradójicamente, estas regiones suelen contar con un mayor potencial en recursos renovables, como la irradiación solar, la velocidad del viento o el potencial hidroeléctrico.

A pesar de estos desafíos, los beneficios asociados con las energías renovables son innegables. Además de reducir las emisiones de gases de efecto invernadero, estas tecnologías pueden mejorar la calidad del aire, disminuir los costos en salud pública y proporcionar acceso a electricidad a comunidades remotas que no están conectadas a redes eléctricas convencionales. Asimismo, el sector de las energías renovables es un motor de empleo, habiendo generado más de 12 millones de puestos de trabajo a nivel mundial en 2022, una cifra que continúa en crecimiento año tras año.

En este proyecto, se realiza un análisis integral del panorama de las energías renovables en los países del mundo, abordando aspectos clave como la producción, el consumo, las inversiones, las políticas gubernamentales y los factores ambientales y socioeconómicos que influyen en su desarrollo. Este análisis tiene como objetivo identificar patrones, tendencias y áreas de oportunidad que puedan servir como base para futuras estrategias de implementación y colaboración internacional.

2. Energías Renovables

2.1. Tipos de energía renovable

La **energía solar** se obtiene mediante la captura de la radiación solar a través de paneles fotovoltaicos, los cuales convierten la luz solar en electricidad, o mediante sistemas térmicos solares, que concentran el calor para generar energía mecánica o eléctrica. Su uso abarca desde la generación de elec-

tricidad y el calentamiento de agua hasta sistemas de iluminación, y se ha consolidado como la energía renovable de más rápido crecimiento. Los costos de los paneles fotovoltaicos han disminuido más del 80 % desde 2010.

Por su parte, la **energía eólica** aprovecha la fuerza del viento mediante aerogeneradores, que transforman la energía cinética del viento en electricidad. Esta forma de energía es común tanto en parques eólicos a gran escala como en sistemas locales. La energía eólica ocupa la segunda posición en importancia, con avances significativos en turbinas marinas (*offshore*) y terrestres (*onshore*).

La **energía hidroeléctrica** utiliza el agua en movimiento, proveniente de ríos o embalses, para accionar turbinas hidráulicas que convierten la energía cinética o potencial en electricidad. Esta fuente de energía es principalmente utilizada en centrales hidroeléctricas y sistemas de almacenamiento energético.

En cuanto a la **energía geotérmica**, esta aprovecha el calor interno de la Tierra, extraído de reservorios subterráneos de agua caliente o vapor, para mover turbinas y generar electricidad, además de utilizarse directamente en calefacción.

La **biomasa** utiliza materia orgánica, como madera, residuos agrícolas o urbanos, que puede ser transformada en energía mediante combustión, gasificación o fermentación para producir electricidad, calor o biocombustibles, siendo clave en procesos industriales y en el sector del transporte.

La **energía marina**, también conocida como undimotriz, captura la energía del movimiento del agua en océanos y mares, ya sea a través de olas, corrientes o mareas, para generar electricidad en zonas costeras.

La **energía de hidrógeno**, considerada renovable cuando se produce a partir de fuentes verdes, se obtiene mediante electrólisis del agua, utilizando electricidad de fuentes renovables para separar el hidrógeno del oxígeno. El hidrógeno actúa como combustible en el transporte, el almacenamiento energético y la generación de electricidad.

De manera similar, el **biogás** se produce por la descomposición anaeróbica de materia orgánica, generando metano que puede ser quemado para producir electricidad o calor.

Finalmente, la **energía oceánica térmica** aprovecha la diferencia de temperatura entre las aguas superficiales cálidas y las profundas frías para generar electricidad mediante ciclos térmicos, siendo especialmente adecuada para regiones tropicales.

Estos tipos de energía renovable ofrecen numerosos beneficios, tales como su sostenibilidad al depender de recursos naturales inagotables, la reducción de las emisiones de gases de efecto invernadero y su versatilidad en aplicaciones a diferentes escalas. Sin embargo, también enfrentan desafíos, como la intermitencia de algunas fuentes, los elevados costos iniciales de infraestructura y la complejidad de integrar estas tecnologías en las redes eléctricas existentes.

2.2. Panorama Global de la Producción de Energías Renovables

La producción de energía renovable se concentra en países que sobresalen por sus políticas energéticas, su inversión en tecnología y la disponibilidad de recursos naturales abundantes. En 2023, las energías renovables representaron cerca del 30 % de la generación eléctrica mundial, mostrando un crecimiento constante en las últimas décadas, impulsado por los avances tecnológicos, las políticas gubernamentales favorables y la disminución de los costos. Este crecimiento ha sido liderado principalmente por los siguientes países:

- **China**: Es el mayor productor mundial de energía renovable, destacándose en la producción de energía solar y eólica, además de poseer grandes infraestructuras hidroeléctricas, como la represa de las Tres Gargantas.
- **Estados Unidos**: Se caracteriza por su destacada producción de energía eólica en estados como Texas e Iowa, así como por sus avances en biomasa y energía solar, especialmente en California.

- **India**: Experimenta un rápido crecimiento en la generación de energía solar gracias a políticas gubernamentales favorables y un clima propicio. También sobresale en hidroeléctrica y biomasa.
- **Brasil**: Depende principalmente de la energía hidroeléctrica y es líder en la producción de biocombustibles, especialmente etanol derivado de caña de azúcar.
- **Alemania**: Es un pionero en Europa en el uso de energía solar y eólica, con políticas agresivas orientadas a fomentar la transición energética.
- **Noruega**: Genera casi toda su electricidad a partir de energía hidroeléctrica y está invirtiendo en energía eólica marina.
- **España**: Destaca por su producción de energía eólica y solar, siendo uno de los principales productores de Europa.
- **Australia**: Presenta un crecimiento acelerado en la adopción de energía solar y eólica, aprovechando su clima soleado y vastas extensiones de terreno.

La producción de energías renovables es un indicador clave en la transición energética global, reflejando el esfuerzo de los países y las empresas por reducir su dependencia de los combustibles fósiles y mitigar el cambio climático. Estos países ejemplifican el avance hacia una matriz energética más sostenible, liderando los esfuerzos globales para reducir las emisiones de carbono y combatir el cambio climático.

2.3. Tendencias y Perspectivas Futuras

- **Descarbonización**: Los compromisos globales para alcanzar la neutralidad de carbono están impulsando significativamente la inversión en fuentes de energía renovable.
- **Avances tecnológicos**: Innovaciones tales como baterías de mayor capacidad, paneles solares más eficientes y turbinas eólicas flotantes están incrementando tanto la viabilidad económica como técnica de las energías renovables.
- **Reducción de costos**: El costo de las energías renovables continúa disminuyendo, superando en competitividad a los combustibles fósiles en muchos mercados.
- **Almacenamiento de energía**: Soluciones como el hidrógeno verde y las baterías avanzadas permitirán almacenar energía renovable de manera más eficiente, resolviendo así el desafío de la intermitencia.
- **Inversiones globales**: Se prevé que las energías renovables atraigan la mayor parte de las inversiones en el sector energético en las próximas décadas, liderando la transformación de la industria.

La producción de energías renovables no solo está revolucionando la manera en que generamos electricidad, sino que también desempeña un papel crucial en la lucha contra el cambio climático, la creación de empleo y la diversificación de las economías. Con un crecimiento sostenido y el respaldo de políticas globales, las energías renovables se posicionan para dominar la matriz energética del futuro.

3. Acerca del conjunto de datos

El conjunto de datos sobre energías renovables y sus indicadores a nivel mundial es un recurso integral diseñado para realizar análisis e investigaciones en profundidad en el campo de las energías renovables. Este conjunto de datos incluye información detallada sobre la producción de energías renovables, factores socioeconómicos e indicadores ambientales de los siguientes países: Japón, Francia, Brasil, Canadá, India, Estados Unidos, Rusia, Alemania, China y Australia. Entre sus principales características se incluyen:

1. **Datos de energía renovable**: Cubre varios tipos de fuentes de energía renovable, como la solar, eólica, hidroeléctrica y geotérmica, detallando su producción (en GWh), capacidad instalada (en MW) e inversiones (en USD) en diferentes países y años.

2. **Indicadores socioeconómicos**: Incluye datos sobre población, PIB, consumo de energía, exportaciones e importaciones de energía, emisiones de CO2, empleos en energía renovable, políticas gubernamentales, gasto en I+D y objetivos de energía renovable.

3. **Factores ambientales**: Proporciona información sobre la temperatura media anual, las precipitaciones anuales, la irradiación solar, la velocidad del viento, el potencial hídrico, el potencial geotérmico y la disponibilidad de biomasa.

4. **Características adicionales**: Contiene características relevantes como capacidad de almacenamiento de energía, capacidad de integración a la red, precios de la electricidad, subsidios a la energía, ayuda internacional para energías renovables, puntajes de concientización pública, programas de eficiencia energética, tasa de urbanización, tasa de industrialización, liberalización del mercado energético, patentes de energía renovable, nivel educativo, acuerdos de transferencia de tecnología, programas de educación en energía renovable, capacidad de fabricación local, aranceles de importación, incentivos a la exportación, desastres naturales, estabilidad política, índice de percepción de la corrupción, calidad regulatoria, estado de derecho, control de la corrupción, índice de libertad económica, facilidad para hacer negocios, índice de innovación, número de instituciones de investigación, conferencias sobre energía renovable, publicaciones sobre energía renovable, fuerza laboral del sector energético, proporción de energía proveniente de energías renovables, asociaciones público-privadas y cooperación regional en energía renovable.

Este conjunto de datos es ideal para analistas, investigadores y formuladores de políticas que buscan estudiar tendencias, impactos y estrategias relacionadas con el desarrollo de energía renovable a nivel mundial.

Estructura:

- **Filas:** Son 2500 filas que representan varios países y años.
- **Columnas:** 56 columnas que incluyen las siguientes categorías:
- **Datos de energía renovable:**
  País, Año, Tipo de energía (solar, eólica, hidroeléctrica, geotérmica, etc.), Producción (GWh), Capacidad instalada (MW), Inversiones (USD).
- **Indicadores socioeconómicos:**
  Población, PIB, Consumo de energía, Exportaciones de energía, Importaciones de energía, Emisiones de CO2, Empleos en energía renovable, Políticas gubernamentales (variable ficticia), Gasto en I+D, Objetivos de energía renovable (variable ficticia).

- **Factores ambientales:**
  Temperatura media anual, Precipitación anual, Irradiación solar, Velocidad del viento, Potencial hidroeléctrico, Potencial geotérmico, Disponibilidad de biomasa.

- **Características adicionales:**
  Capacidad de almacenamiento de energía, Capacidad de integración en la red, Precios de la electricidad, Subsidios energéticos, Ayuda internacional para renovables, Conciencia pública (puntuación de la encuesta), Programas de eficiencia energética (variable ficticia), Tasa de urbanización, Tasa de industrialización, Liberalización del mercado energético (variable ficticia), Patentes de energía renovable, Nivel educativo, Acuerdos de transferencia de tecnología, Programas de educación sobre energía renovable (variable ficticia), Capacidad de fabricación local, Aranceles de importación de equipos energéticos, Incentivos a la exportación de equipos energéticos, Desastres naturales (variable ficticia), Estabilidad política (puntaje), Índice de percepción de la corrupción, Calidad regulatoria (puntaje), Estado de derecho (puntaje), Control de la corrupción (puntaje), Índice de libertad económica, Facilidad para hacer negocios (puntaje), Índice de innovación, Número de instituciones de investigación, Número de conferencias sobre energía renovable, Número de publicaciones sobre energía renovable, Fuerza laboral del sector energético, Proporción de energía proveniente de energías renovables (%), Asociaciones público-privadas en energía (variable ficticia), Cooperación regional en energía renovable (variable ficticia).

4. Análisis de datos

Realizaremos un análisis descriptivo del conjunto de datos con el objetivo de obtener información relevante e interesante. Con el fin de identificar las principales fuentes de energía renovable que más producen los países y cómo ha evolucionado esta producción a lo largo del tiempo, hemos extraído la producción total de energía de los países incluidos en el conjunto de datos para los años 2000 y 2023 (ver figuras 1 y 2).

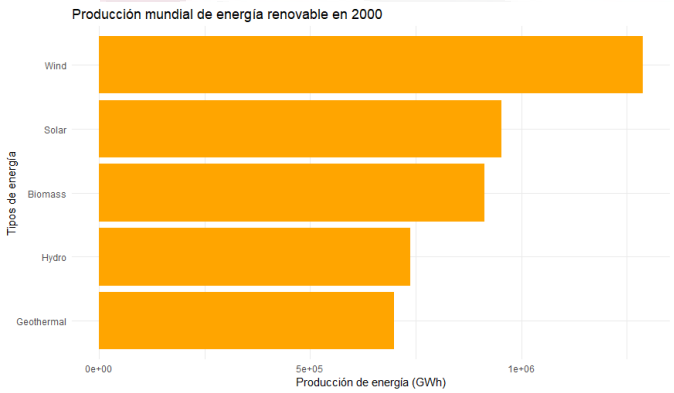

Figura 1: Producción mundial de energía por tipo en el año 2000.

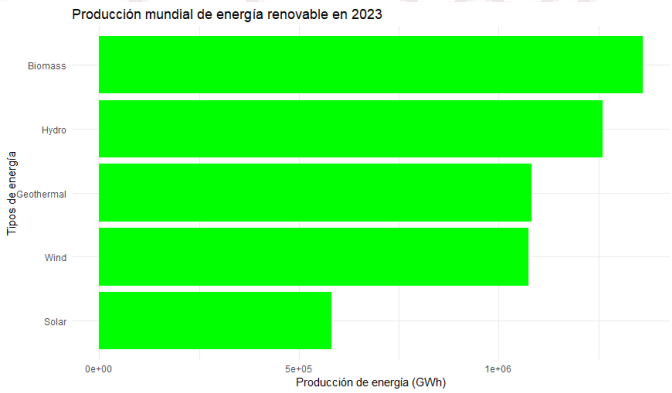

Figura 2: Producción mundial de energía por tipo en el año 2023.

Para el año 2000, se observa que las tres principales fuentes de energía eran la energía eólica, solar y biomasa. En cambio, en 2023, las principales fuentes fueron la biomasa, la hidroeléctrica y la geotermal. Curiosamente, las energías solar y eólica fueron desplazadas al final del ranking para el año 2023. Para examinar cómo han cambiado las producciones de las diferentes fuentes de energía en cada año, se puede consultar la figura (3).

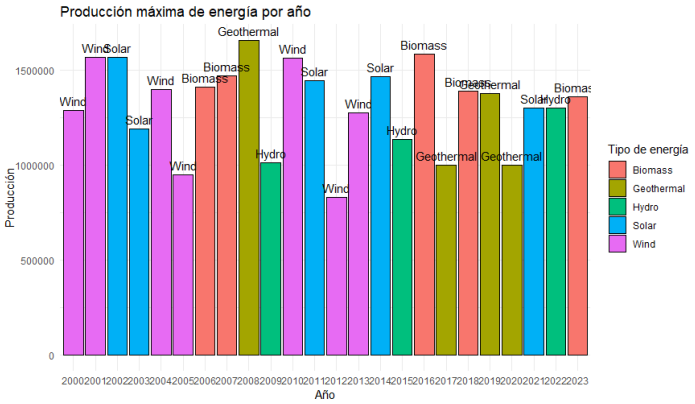

Figura 3: Tipo de energía con mayor producción para cada año, comprendido entre 2000 y 2023.

A través de esta gráfica, se puede observar que la producción de energía varió considerablemente entre los diferentes años, sin evidenciar una tendencia clara que favorezca a algún tipo de energía en particular. Sin embargo, se puede afirmar que, a lo largo de estos 23 años, la energía eólica ha sido la más producida durante seis de esos años, mientras que la hidroeléctrica ha sido la más producida en solo tres años.

En las figuras (4, 5) se presenta la producción total de energía por tipo y país para los años 2000 y 2023.

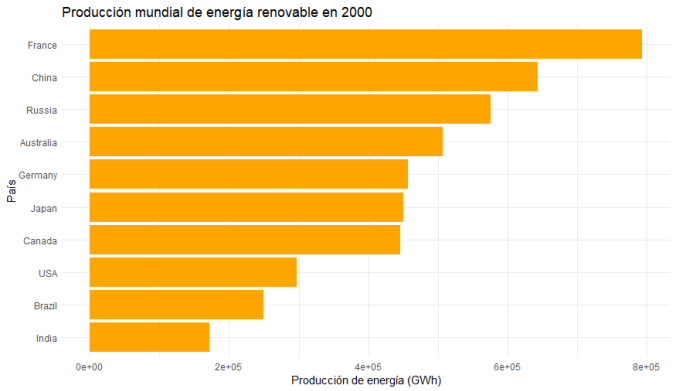

Figura 4: Producción mundial de energía por país en el año 2000.

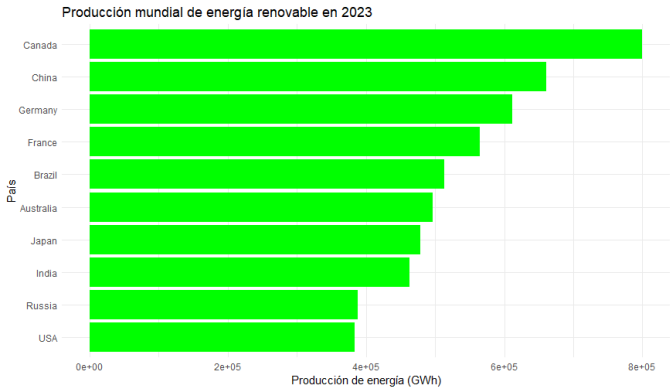

Figura 5: Producción mundial de energía por país en el año 2023.

Se observa un incremento global en la producción de energía, posiblemente relacionado con el aumento de la población mundial durante este período. En el año 2000, los países líderes en producción de energía eran Francia, China y Rusia, mientras que Estados Unidos, Brasil e India se encontraban en los últimos lugares. Para 2023, Canadá lideró la producción de energía, seguido por China, que mantuvo su puesto, y Alemania. Los últimos países en la lista fueron India, Rusia y Estados Unidos. Resulta notable que, aunque Rusia era uno de los principales productores de energía en 2000, pasó a ser uno de los países con menor producción en 2023, mientras que China logró mantener su puesto como el segundo mayor productor.

La figura 6 muestra un contraste entre la producción de energía por país en los años 2000 y 2023.

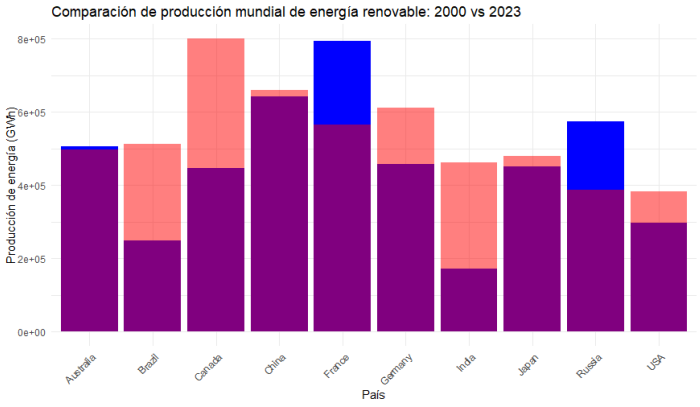

Figura 6: Contraste entre la producción de energía por país en 2000 y 2023. Las barras azules representan la producción de 2000 y las rojas la de 2023.

En este gráfico, se aprecia el cambio en la producción de energía en cada país. Es evidente que países como Brasil, Canadá, China, Alemania, India, Japón y Estados Unidos aumentaron su producción, mientras que países como Australia, Francia y Rusia experimentaron una disminución. Sin embargo, a nivel global, la producción de energía aumentó, especialmente en Brasil y Canadá. Para un análisis más detallado de la producción energética de cada país a lo largo de los años, se puede consultar la figura (7).

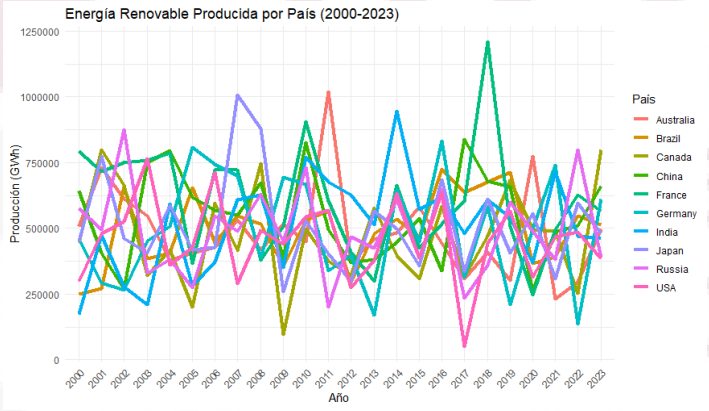

Figura 7: Series de tiempo de la producción de energía de cada país a lo largo de los años.

Finalmente, para observar las inversiones realizadas en la producción de energía, se pueden consultar las figuras 8 y 9.

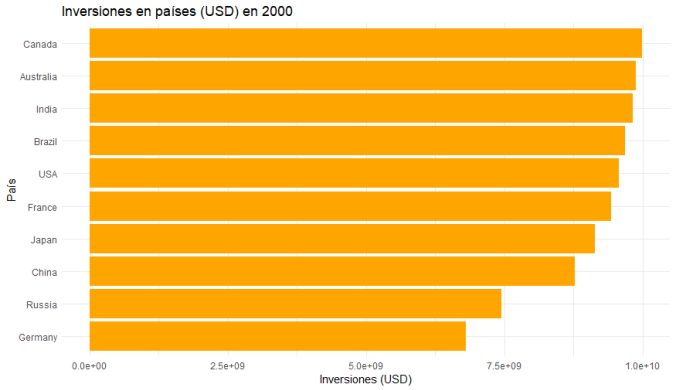

Figura 8: Inversión mundial en energía renovable en 2000.

Figura 9: Inversión mundial en energía renovable en 2023.

En estas figuras podemos observar las inversiones realizadas por cada país en los años 2000 y 2023, expresadas en dólares. En primer lugar, podemos destacar que, en el año 2000, los países que más invirtieron fueron Canadá, seguido de Australia e India. Los países con menores inversiones fueron China, Rusia y Alemania. Para el año 2023, los países que más invirtieron fueron Rusia, Japón y China, mientras que los que menos invirtieron fueron Australia, Alemania y Brasil. Es interesante notar cómo Rusia y China han incrementado significativamente sus inversiones en este período, pasando de ser los países con menores inversiones a ser los de mayor inversión. Aunque Rusia es el país con la mayor inversión en 2023, no es el mayor productor de energía, ocupando el penúltimo puesto en términos de producción. Un caso similar ocurre con Japón. Algo curioso es que, a pesar de que algunos países han realizado mayores inversiones en energías renovables, esto no siempre se traduce en una mayor producción de energía, como es el caso de Rusia y Japón. En cambio, Canadá, que no se encuentra entre los países con mayores inversiones, es uno de los que más produce energía.

Para conocer los países que más consumieron energía en los años 2000 y 2023, podemos observar las figuras 10 y 11.

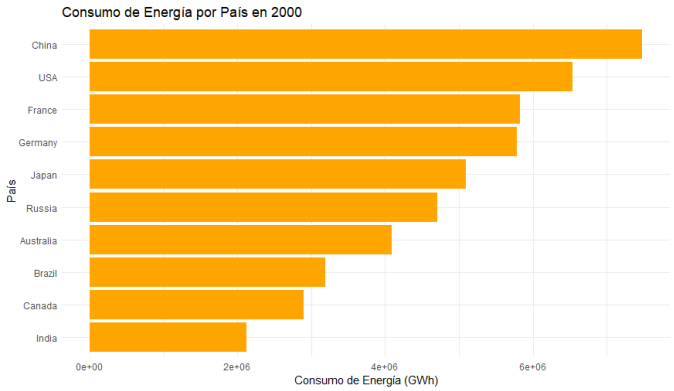

Figura 10: Consumo de energía mundial por país en el año 2000.

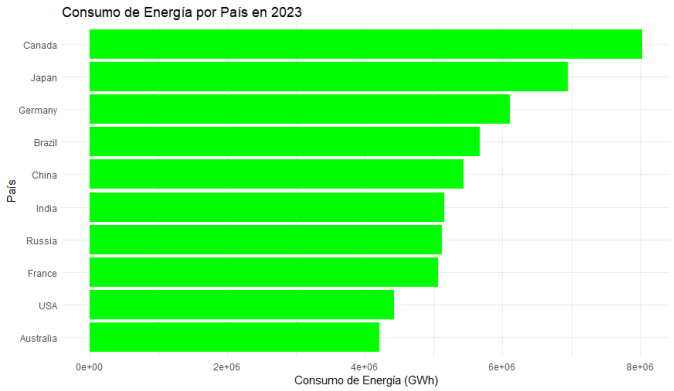

Figura 11: Consumo de energía mundial por país en el año 2023.

En el año 2000, los países que más consumieron energía fueron China, Estados Unidos y Francia, mientras que los países con menor consumo fueron Brasil, Canadá y China. Para el año 2023, los países que más consumieron energía fueron Canadá, Japón y Alemania, mientras que los que menos consumieron fueron Francia, Estados Unidos y Australia. Es destacable que Japón, a pesar de no producir mucha energía en comparación con otros países, es uno de los mayores consumidores de energía en 2023.

También es relevante que Estados Unidos no figura entre los países que más producen ni entre los que más consumen energía.

Para un mejor contraste entre los países, podemos observar la figura 12.

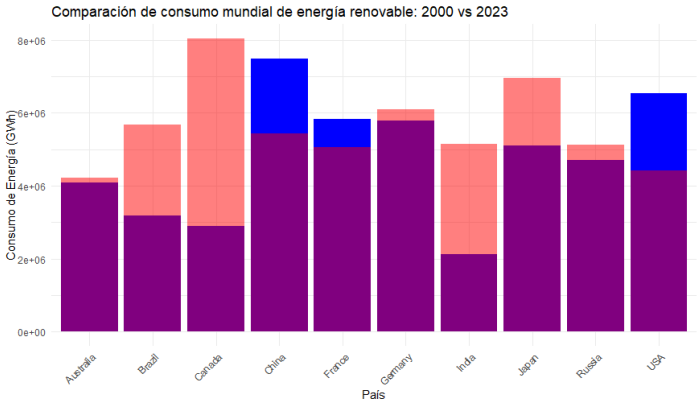

Figura 12: Contraste del consumo de energía de los países entre los años 2000 y 2023. Las barras azules representan la energía consumida en el año 2000 y las rojas en el año 2023.

Se puede observar que los países que incrementaron su consumo de energía del 2000 al 2023 fueron Brasil, Canadá, India, Japón y Australia, mientras que los países que redujeron su consumo fueron China, Francia y Estados Unidos.

Para observar las exportaciones de energía renovable realizadas por los países en los años 2000 y 2023, se pueden consultar las figuras 13 y 14.

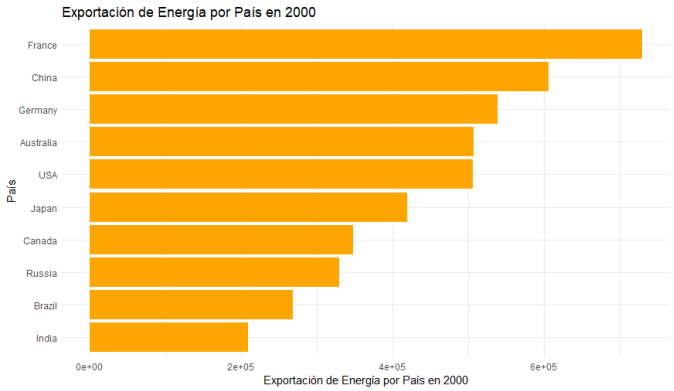

Figura 13: Exportación de energía mundial por país en el año 2000.

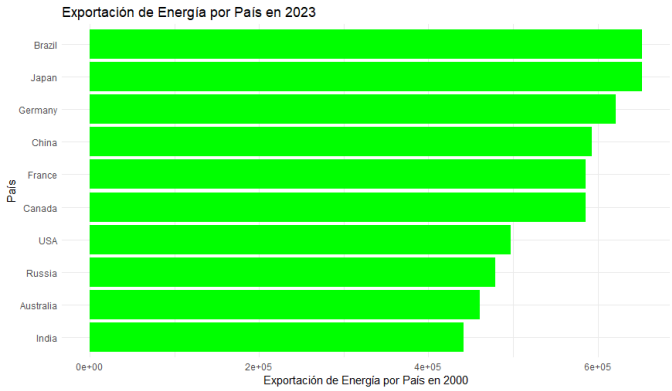

Figura 14: Exportación de energía mundial por país en el año 2023.

En el año 2000, los países líderes en exportación de energías renovables fueron Francia, China y Alemania, mientras que en 2023, los países más destacados en exportación fueron Brasil, Japón y Alemania. Es interesante observar que países como Japón y Alemania, que son grandes consumidores de energía, también están entre los principales exportadores de energía en 2023. Por otro lado, Canadá, aunque es el país que más produce energía, consume gran parte de su producción y no se destaca en la exportación de energía renovable.

En el año 2000, Francia era el país que más producía y exportaba energía, pero con el tiempo ha sido superado por otros países en términos de exportación.

Para observar el contraste entre las exportaciones de energía, podemos consultar la figura 15.

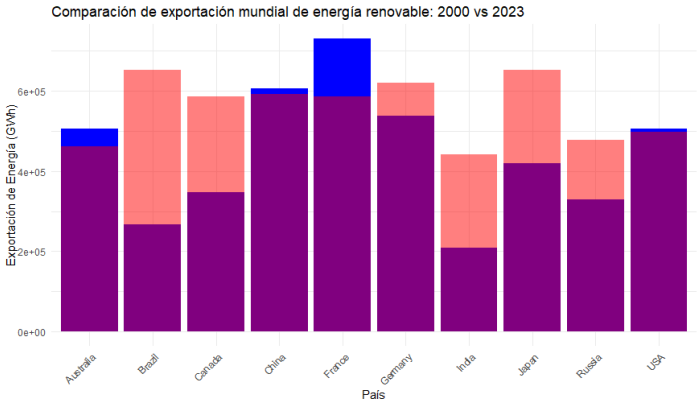

Figura 15: Contraste de la exportación de energía de los países entre los años 2000 y 2023. Las barras azules representan la exportación de energía en el año 2000 y las rojas en el año 2023.

Entre los países que más han incrementado sus exportaciones de energías renovables se encuentran Brasil, Canadá, Alemania, India, Japón y Rusia. Los países que han reducido sus exportaciones incluyen Australia, China, Francia y Estados Unidos.

5. Metodología

5.1. Modelo ARIMA

El modelo $ARIMA(p, d, q)$ es un modelo estadístico utilizado para el análisis y pronóstico de series temporales. Combina tres componentes principales: autoregresión (AR), diferenciación (I) y media móvil (MA), lo que lo hace especialmente adecuado para capturar patrones temporales en series estacionarias o transformarlas en estacionarias mediante diferenciación.

El modelo se denota como $ARIMA(p, d, q)$, donde:

- p: número de términos autoregresivos,
- d: número de diferencias necesarias para estacionarizar la serie,
- q: número de términos de media móvil.

La forma general del modelo ARIMA es:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d Y_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) e_t$$

donde B es el operador de rezago, ϕ son los coeficientes autoregresivos, θ son los coeficientes de media móvil y e_t es el término de error.

El modelo ARIMA describe los valores actuales como una combinación de valores pasados (AR), cambios recientes en los datos (I) y errores aleatorios pasados (MA). Su objetivo es encontrar una combinación óptima de p, d y q que minimice el error de predicción. Con esta estructura, el modelo ARIMA se convierte en una herramienta fundamental para el análisis de datos temporales, combinando simplicidad, flexibilidad y precisión en la modelación.

5.1.1. Raíz del Error Cuadrático Medio

El **RMSE** (Root Mean Square Error o Raíz del Error Cuadrático Medio) es una métrica comúnmente utilizada para evaluar la calidad de los modelos de predicción, especialmente en contextos de series temporales y regresión. Mide la magnitud promedio de los errores entre los valores predichos por un modelo y los valores observados en los datos originales. Matemáticamente, se define como:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2},$$

donde n es el número total de observaciones, y_i es el valor observado en el instante i, y $\hat{y}_i$ es el valor predicho por el modelo en el instante i. El RMSE proporciona una medida de cuánto, en promedio, los valores predichos por el modelo se desvían de los valores reales, expresándose en las mismas unidades que la variable de interés, lo que facilita su interpretación en contextos prácticos. Es una métrica estándar y ampliamente utilizada que penaliza los errores grandes debido a la naturaleza cuadrática de su fórmula, permitiendo así comparar diferentes modelos de manera efectiva. Sin embargo, el RMSE es sensible a valores atípicos, ya que los errores grandes tienen un impacto desproporcionado en el resultado, y no distingue entre errores positivos y negativos, considerando únicamente su magnitud. Por tanto, aunque el RMSE es útil para medir la precisión de los modelos y comparar su desempeño, se debe interpretar cuidadosamente en presencia de valores extremos.

5.2. Selección de variables

5.2.1. Método Forward Selection

El método de selección hacia adelante (**Forward**) comienza con un modelo vacío y agrega variables una por una según su contribución al modelo. El proceso sigue los siguientes pasos:

1. **Inicio:** Comienza con un modelo vacío, sin predictores.
2. **Evaluación:** Se evalúa cada predictor de manera individual para determinar cuál mejora más el modelo (por ejemplo, usando R^2, AIC, BIC o p-valor).
3. **Inclusión:** Se agrega al modelo el predictor que más mejora la métrica seleccionada.
4. **Repetición:** Este proceso se repite hasta que no haya variables que mejoren significativamente el modelo.

Entre las ventajas de este método, se destaca su eficiencia computacional para bases de datos con muchas variables, y su utilidad cuando existe un gran número de variables irrelevantes. No obstante, su principal desventaja es que puede pasar por alto combinaciones de predictores que serían óptimas en conjunto.

5.2.2. Método Backward Selection

El método de selección hacia atrás (**Backward**) comienza con un modelo completo, que incluye todas las variables, y elimina una por una aquellas que tienen menor impacto. El proceso sigue los siguientes pasos:

1. **Inicio:** Comienza con un modelo que incluye todos los predictores.
2. **Evaluación:** Se evalúa la importancia de cada variable (por ejemplo, con los p-valores de las pruebas t o R^2).
3. **Eliminación:** Se elimina la variable menos significativa (aquella que tiene el p-valor más alto o menor contribución al modelo).
4. **Repetición:** Este proceso se repite hasta que todas las variables restantes sean significativas o se alcance un criterio predefinido.

Entre las ventajas de este método, se encuentra que garantiza que todas las variables sean consideradas inicialmente y funciona bien cuando el número de predictores es bajo. Sin embargo, una de sus desventajas es que requiere más recursos computacionales cuando hay muchas variables y puede eliminar variables útiles en combinación con otras.

5.2.3. Regresión Ridge

La regresión Ridge modifica la función de minimización de los errores al agregar un término de penalización que es proporcional al cuadrado de los coeficientes:

$$\text{Minimizar: } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Esta penalización controla la magnitud de los coeficientes β_j, evitando valores excesivamente grandes en situaciones con colinealidad entre predictores. Aunque reduce la magnitud de los coeficientes, no los lleva exactamente a cero, por lo que incluye todos los predictores en el modelo. Es especialmente útil cuando hay muchas variables correlacionadas o cuando el número de predictores es cercano al tamaño de la muestra.

5.2.4. Regresión Lasso

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) agrega una penalización proporcional al valor absoluto de los coeficientes:

$$\text{Minimizar: } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

A diferencia de Ridge, Lasso puede forzar algunos coeficientes a ser exactamente cero, lo que efectivamente realiza una selección de variables. Es ideal para conjuntos de datos con alta dimensionalidad, ya que ayuda a simplificar el modelo al seleccionar automáticamente las variables más relevantes. También puede manejar colinealidad, pero tiende a seleccionar una variable representativa del grupo de predictores correlacionados, descartando las demás.

5.3. Imputación de datos

5.3.1. Imputación por la Media

Se reemplazan los valores faltantes de una variable con la media de los valores observados de esa misma variable. Es un método sencillo y rápido, pero puede subestimar la variabilidad de los datos y sesgar las relaciones entre variables. Es útil en casos con pocos datos faltantes y cuando la precisión no es crítica.

5.3.2. Imputación por Regresión

Se ajusta un modelo de regresión utilizando las variables observadas para predecir los valores faltantes. Aprovecha las relaciones lineales entre las variables para una imputación más precisa, pero puede introducir sesgo si las relaciones entre las variables no son lineales o si hay multicolinealidad.

5.3.3. Imputación por Regresión Estocástica

Similar a la imputación por regresión, pero agrega un término de error aleatorio al valor imputado para preservar la variabilidad de los datos. Ayuda a evitar la sobreestimación de relaciones entre variables. Es especialmente útil cuando se busca realizar análisis inferenciales más robustos después de la imputación.

5.3.4. Imputación Hot-Deck (Vecino Más Cercano)

Se imputan los valores faltantes utilizando observaciones completas que son similares en otras variables. La similitud se define típicamente mediante medidas de distancia (por ejemplo, distancia euclídea o de Gower). Este método preserva las distribuciones originales de los datos y las relaciones entre variables, pero puede ser computacionalmente intensivo en conjuntos de datos grandes.

6. Resultados

6.1. Ajuste de un modelo ARIMA

Al observar los datos, podemos notar que contamos con series de tiempo de la producción de energía en cada uno de los países desde el año 2000 hasta 2023. Por lo tanto, se ajustó un modelo ARIMA a estas series de tiempo para cada país. A continuación, se presenta el ajuste del modelo para los países de Estados Unidos y Japón.

Estados Unidos

Se observa la serie de tiempo de la producción de energía para Estados Unidos.

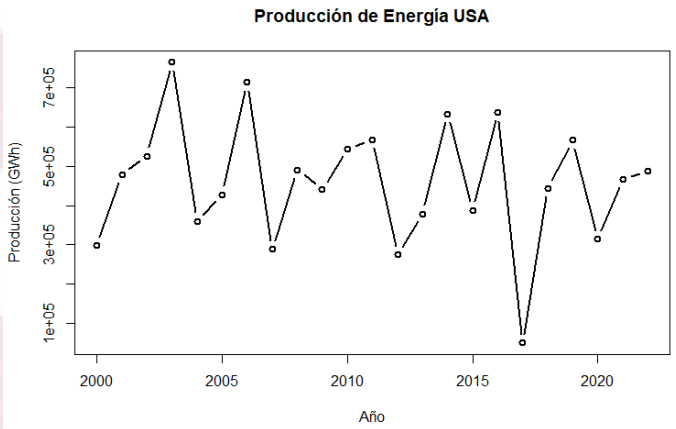

Figura 16: Serie de tiempo de la producción de energía renovable para el país de Estados Unidos.

Se puede notar que la producción de energía en este país no muestra una tendencia positiva ni negativa, por lo que se puede considerar no realizar ninguna diferenciación. La serie parece oscilar alrededor de un valor promedio, y no se observa de manera clara la presencia de estacionalidad. Tras la observación visual, se ajustó un modelo ARIMA probando diferentes combinaciones de parámetros (p, d, q), seleccionando el modelo que minimizó el error de ajuste. El mejor modelo, dado los resultados obtenidos, es el siguiente:

- **Modelo seleccionado:** ARIMA(0,0,1) con media distinta de cero.
- **Coeficientes estimados:**
  - Coeficiente MA(1): $\hat{\theta}_1 = -0.4556$ (error estándar: 0.1750).
    Este coeficiente indica que los errores del modelo están correlacionados con los errores del período anterior, con una magnitud moderada y un signo negativo.
  - Media: $\hat{\mu} = 460,640.62$ (error estándar: 16,712.78).
    Este valor representa el promedio estimado de la serie de tiempo.
- **Varianza del error residual:** $\sigma^2 = 2.199 \times 10^{10}$.
  Esto representa la variabilidad no explicada por el modelo.
- **Criterios de información:**
  - Log-verosimilitud: -305.57

- AIC: 617.13
- AICc (corregido): 618.39
- BIC: 620.54

Estos criterios permiten comparar el modelo con otros posibles candidatos. Valores más bajos indican un mejor ajuste.

- **Errores de predicción en el conjunto de entrenamiento:**
  - Error medio (ME): $-2,510.842$.
    Indica que, en promedio, las predicciones están ligeramente sesgadas hacia abajo.
  - Raíz del error cuadrático medio (RMSE): $141,698$.
    Mide el promedio de los errores al cuadrado, siendo una métrica estándar de error absoluto.
  - Error absoluto medio (MAE): $106,412.8$.
    Representa el promedio de las desviaciones absolutas entre los valores observados y los predichos.
  - Porcentaje de error medio (MPE): $-32.771\,\%$.
    Indica que, en promedio, el modelo subestima los valores reales en un $32.77\,\%$.
  - Porcentaje de error absoluto medio (MAPE): $49.092\,\%$.
    Muestra que, en promedio, el error absoluto es equivalente al $49.09\,\%$ de los valores reales.
  - MASE: 0.4969.
    Representa el error medio absoluto normalizado respecto a un modelo de referencia basado en el promedio.
  - Coeficiente de autocorrelación en el primer rezago de los residuos (ACF_1): 0.0073.
    Indica que no hay autocorrelación significativa en los residuos, lo cual es un buen indicador de ajuste.

Podemos observar el comportamiento de los residuos del modelo, como se muestra en la figura 17.

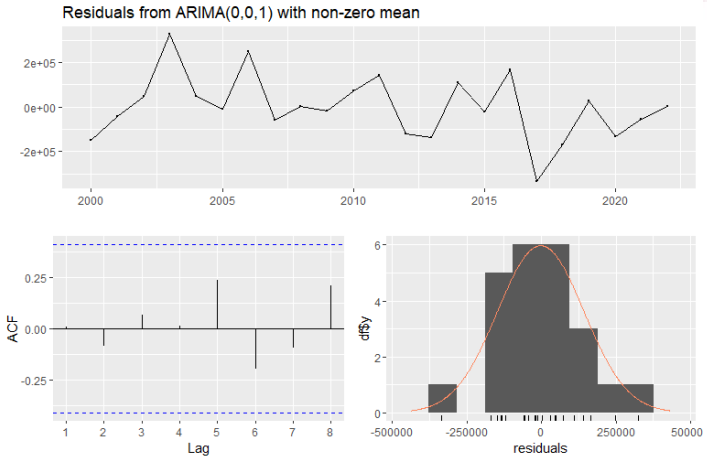

Figura 17: Residuales del modelo ARIMA ajustado.

Se puede apreciar que la serie de los errores no presenta una estructura definida, no muestra tendencias y, en su lugar, oscila alrededor de un valor promedio. Si observamos la figura ACF, se muestra la

función de autocorrelación de los residuos para distintos retardos (lags). Si los residuos son ruido blanco, todas las autocorrelaciones (representadas por las barras) deben encontrarse dentro de las líneas de confianza (líneas azules), como se observa en esta serie. También se presenta una figura que representa la distribución de los residuos y la compara con una curva normal superpuesta. Si el modelo es adecuado, los residuos deben aproximarse a una distribución normal. En este caso, los residuos se asemejan a una distribución normal, lo cual es un buen indicio de que el modelo es adecuado.

Además, se realizó una prueba de Ljung-Box con el fin de verificar la normalidad en los residuos. Los resultados son los siguientes:

```
Ljung-Box test

data:  Residuals from ARIMA(0,0,1) with non-zero mean
Q* = 2.1557, df = 4, p-value = 0.7071

Model df: 1.   Total lags used: 5
```

Podemos observar que el valor de p-valor es 0.7071, lo cual nos ayuda a determinar si los residuos presentan autocorrelación. Un valor alto (como en este caso, mayor a 0.05) sugiere que no hay evidencia significativa de autocorrelación. Por lo tanto, no se rechaza la hipótesis nula, que establece que:

- **Hipótesis nula** (H_0): Los residuos no presentan autocorrelación.
- **Hipótesis alternativa** (H_1): Los residuos presentan autocorrelación significativa.

Por lo tanto, el modelo ARIMA(0,0,1) ajustado produce residuos que se comportan como ruido blanco, es decir, sin autocorrelación significativa.

A continuación, realizamos la predicción para el año 2023, la cual se muestra en la figura 18.

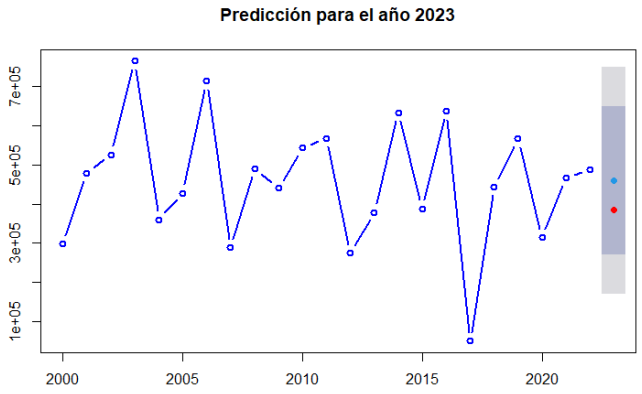

Figura 18: Predicción para el año 2023 con el modelo ARIMA(0,0,1) ajustado. El punto rojo representa el valor real y el punto azul el valor predicho.

Con esta predicción, se obtuvo un valor de RMSE $= 75,910.73$.

Japón

Observamos la serie de tiempo de la producción de energía para Japón.

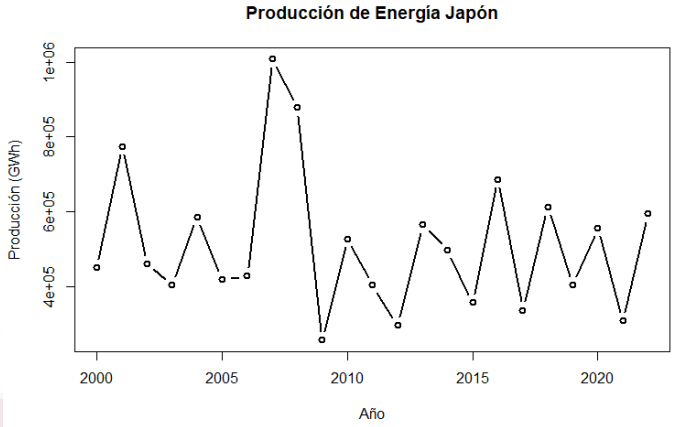

Figura 19: Serie de tiempo de la producción de energía renovable para Japón.

Se puede notar que, al igual que en el caso anterior, la serie de tiempo no presenta una tendencia definida, sino que tiende a oscilar alrededor de un valor medio. Por lo tanto, no es necesario realizar ninguna diferenciación. Tampoco es evidente si existe estacionalidad. Posteriormente, tras probar distintas configuraciones de los parámetros (p, d, q) en modelos ARIMA, se seleccionó el modelo con el menor valor del Criterio de Información de Akaike (AIC). El modelo seleccionado fue el siguiente:

- **Modelo seleccionado:** ARIMA(0,0,0) con media distinta de cero.
- **Coeficientes estimados:**
  - Media: $\hat{\mu} = 513,914.43$ (error estándar: 38,136.73).
    Este valor representa el promedio estimado de la serie de tiempo.
- **Varianza del error residual:** $\sigma^2 = 3.497 \times 10^{10}$.
  Esto representa la variabilidad no explicada por el modelo.
- **Criterios de información:**
  - Log-verosimilitud: -311.32
  - AIC: 626.64
  - AICc (corregido): 627.24
  - BIC: 628.91

  Estos criterios permiten comparar el modelo con otros posibles candidatos. Valores más bajos indican un mejor ajuste.
- **Errores de predicción en el conjunto de entrenamiento:**
  - Error medio (ME): -1.8475×10^{-10}.
    Indica que, en promedio, las predicciones están prácticamente sin sesgo.
  - Raíz del error cuadrático medio (RMSE): $182,891.1$.
    Mide el promedio de los errores al cuadrado, siendo una métrica estándar de error absoluto.

- Error absoluto medio (MAE): $143,533.2$.
  Representa el promedio de las desviaciones absolutas entre los valores observados y los predichos.
- Porcentaje de error medio (MPE): $-12.00193\,\%$.
  Indica que, en promedio, el modelo subestima los valores reales en un $12.00\,\%$.
- Porcentaje de error absoluto medio (MAPE): $30.3195\,\%$.
  Muestra que, en promedio, el error absoluto es equivalente al $30.32\,\%$ de los valores reales.
- MASE: 0.6073.
  Representa el error medio absoluto normalizado respecto a un modelo de referencia basado en el promedio.
- Coeficiente de autocorrelación en el primer rezago de los residuos (ACF_1): -0.1207.
  Indica que existe una ligera autocorrelación negativa en los residuos, lo cual podría señalar una leve dependencia entre los errores en los períodos consecutivos.

Notamos que el mejor modelo fue un modelo ARIMA(0,0,0), lo que indica un modelo sin componentes autorregresivos, sin diferenciación y sin componentes de media móvil. Este modelo es simplemente un modelo de media constante con la suposición de que la serie temporal es estacionaria en su nivel y no requiere ninguna manipulación adicional para predecir los valores futuros.

En la práctica, este modelo es el modelo base más simple y se usa a menudo como referencia o punto de partida. Su uso es útil cuando los datos son altamente estacionarios y no requieren una estructura más compleja para su modelado.

Podemos observar el comportamiento de los residuales del modelo, como se muestra en la figura 27.

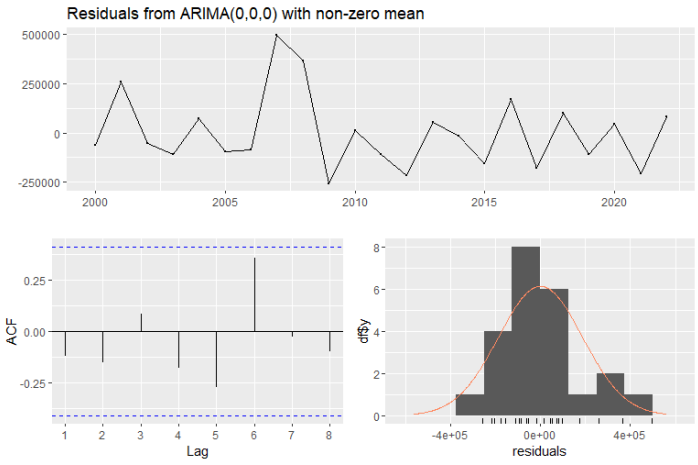

Figura 20: Residuales del modelo ARIMA ajustado.

Al observar los residuales, podemos darnos cuenta de que tampoco presentan alguna estructura ni tendencia; simplemente oscilan alrededor de un valor muy cercano a cero. Para confirmar esto, realizamos una prueba de Ljung-Box:

```
Ljung-Box test

data:  Residuals from ARIMA(0,0,0) with non-zero mean
Q* = 4.5339, df = 5, p-value = 0.4754

Model df: 0.   Total lags used: 5
```

Podemos asegurar que los residuos no están correlacionados y que lo que observamos es ruido blanco, con una significancia de 0.05. De esta forma, corroboramos esto con el histograma mostrado de los errores y una aproximación de una función de distribución normal. Finalmente, al observar el gráfico de la función de autocorrelación (ACF), también confirmamos que los residuos son ruido blanco, debido a que las autocorrelaciones se encuentran dentro de la banda de confianza.

Por lo tanto, se realizó la predicción para el año 2023, que podemos observar en la figura 28.

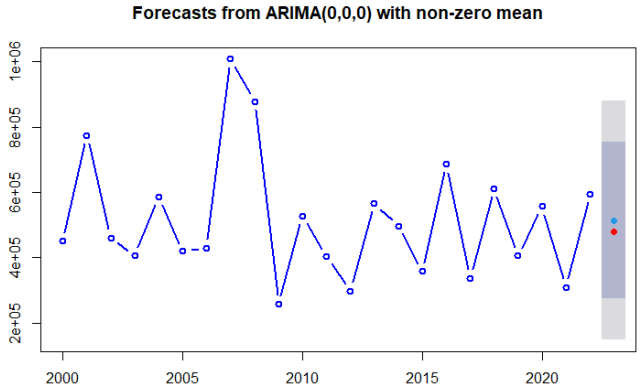

Figura 21: Predicción para el año 2023 con el modelo ARIMA(0,0,0) ajustado. El punto rojo representa el valor verdadero y el azul el predicho.

Con esta predicción se obtuvo un valor $RMSE = 34,844.66$.

Para estos dos ejemplos, los modelos seleccionados fueron los indicados; sin embargo, tenemos más países, como Alemania, Australia, China, Francia, etc. Cada serie de tiempo para cada país se ajustó a un modelo ARIMA no muy diferente a los parámetros mostrados para estos dos ejemplos, ya que no hay mucha diferencia en el comportamiento de la producción de energía entre los países. Si observamos las series de tiempo de todos los países, como se muestra en la figura 29,

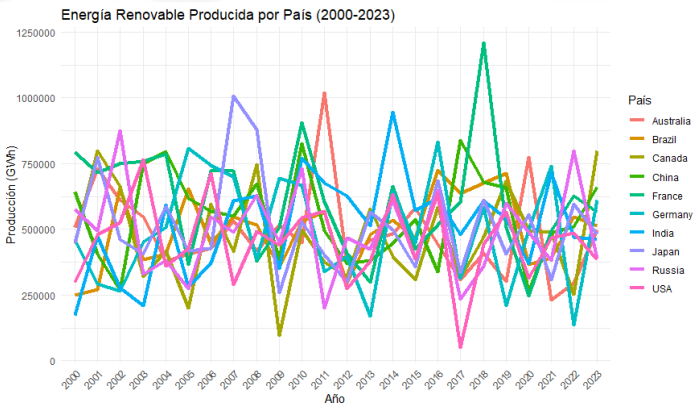

Figura 22: Series de tiempo de cada país para la producción de energía a lo largo de los años.

Podemos apreciar un comportamiento similar entre todas las series de tiempo: todas son estacionarias, no tienen tendencias, no presentan varianzas cambiantes ni una media que dependa del tiempo. Por tanto, podemos pensar que son series de tiempo que no muestran gran complejidad. De esta forma, realizamos

una predicción de energía renovable para los países en el año 2024, y podemos compararlos con las producciones del año 2000 (ver figuras 30 y 31):

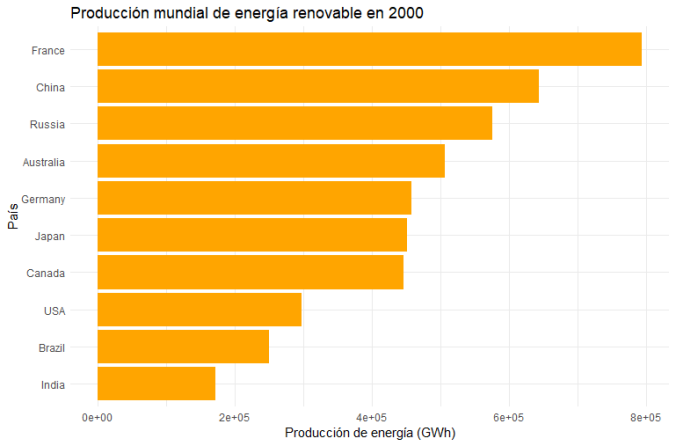

Figura 23: Producción de energía mundial para cada país en el año 2000.

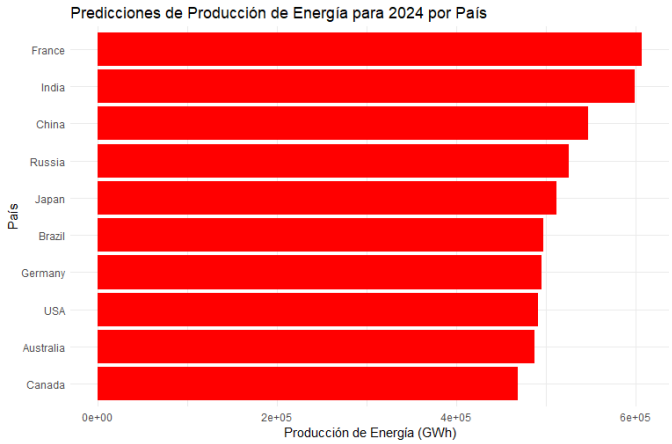

Figura 24: Predicción de la producción de energía mundial para cada país en el año 2024.

Podemos apreciar un considerable crecimiento en la producción en los 24 años. Es notable el gran crecimiento de India en la producción de energías renovables, y cómo Francia vuelve a ser el país con mayor producción. Algo interesante es que Canadá, que en 2023 era el principal productor de energía, en 2024 tuvo una de las producciones más bajas. Sin embargo, para el año 2024, las diferencias entre los países en cuanto a producción se reducen, y no hay tanta diferencia en comparación con el año 2000. Este comportamiento de Canadá podría explicarse por la naturaleza cíclica de la producción, ya que en todas las series de tiempo observamos fluctuaciones consecutivas de producción, como si en un año produjeran demasiado y el siguiente bajaran la producción. Dado que la producción de Canadá fue la más alta en 2023, para 2024 se esperaba una disminución en la producción. Estas diferencias entre países en cuanto a la producción en los años 2023 y 2024 pueden observarse en la figura 32.

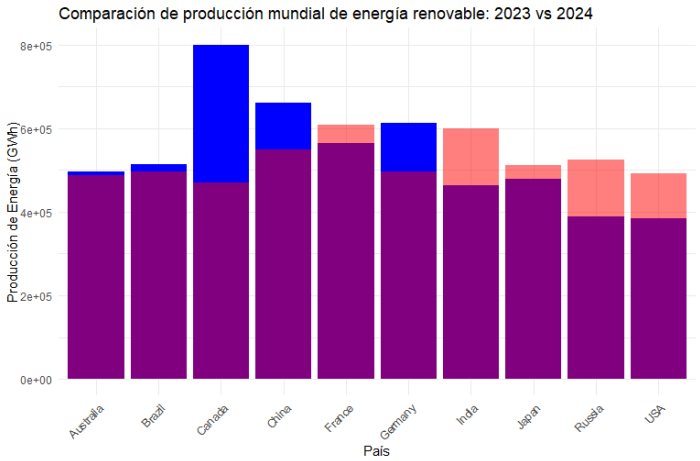

Figura 25: Contraste de la producción de energía de los países entre los años 2023 y la predicción del año 2024. Las barras azules representan la energía producida en el año 2023 y las rojas en el año 2024.

Aquí podemos ver claramente cómo los países de Francia, India, Japón, Rusia y Estados Unidos tuvieron una mayor producción en relación con el año anterior, mientras que países como Australia, Brasil, Canadá, China y Alemania presentan una disminución en la producción, especialmente Canadá.

6.2. Selección de variables

Como parte del estudio también se se realizó la selección de subconjuntos para determinar las variables más importantes, tomando como variable de respuesta a la producción (en GWh) total de energía de los paises, para realizar dicha tarea se optó por el método de selección de subcinjuntos adelante y hacía atrás, ya que este método es más eficiente para un grán número de variables predictoras.

En los siguientes gráficos se observan las relaciones de los predictores con los diferentes criterios para la selección del número de predictores por el método Forward stepwise, para cada criterio se observa un óptimo diferente en el número de predictores seleccionado.

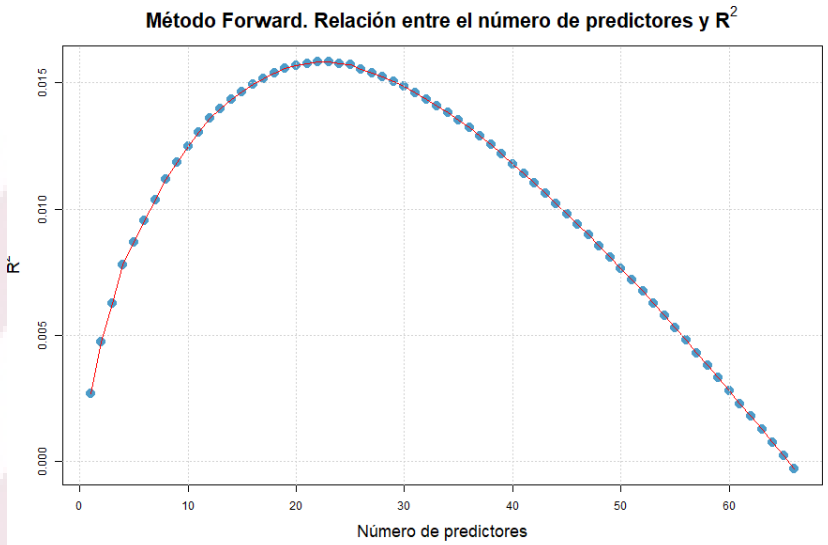

Figura 26: Relación entre el número de predictores y el R^2 con el método Forward stepwise para los 68 predictores

Los resultados arrojan que el mejor modelo según el R^2 ajustado tiene 22 predictores, lo cual se logra apreciar con gran claridad en la figura(26), donde vemos que el máximo está, precisamente entre 20 y 30 predictores

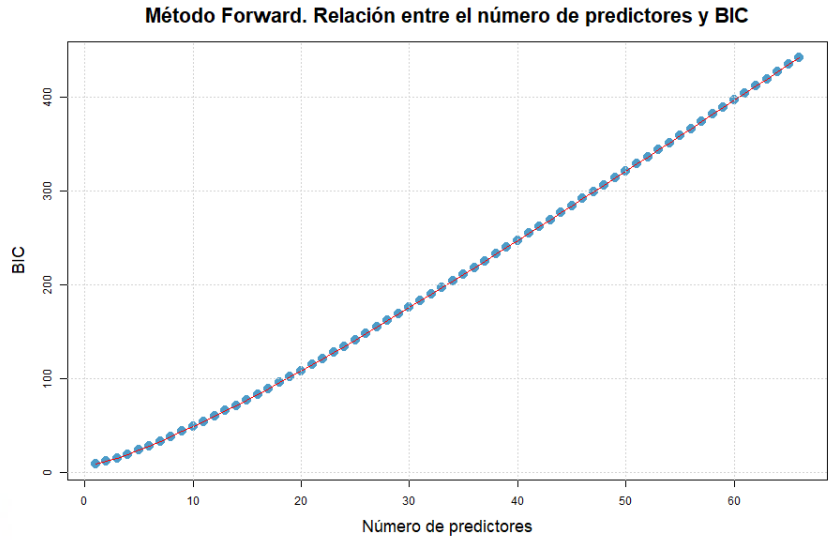

Figura 27: Relación entre el número de predictores y el BIC con el método Forward stepwise para los 68 predictores

Por otra parte, los resultados arrojaron que el mejor modelo según el BIC tiene 1 predictor, lo cual es evidente si se observa la figura(27)

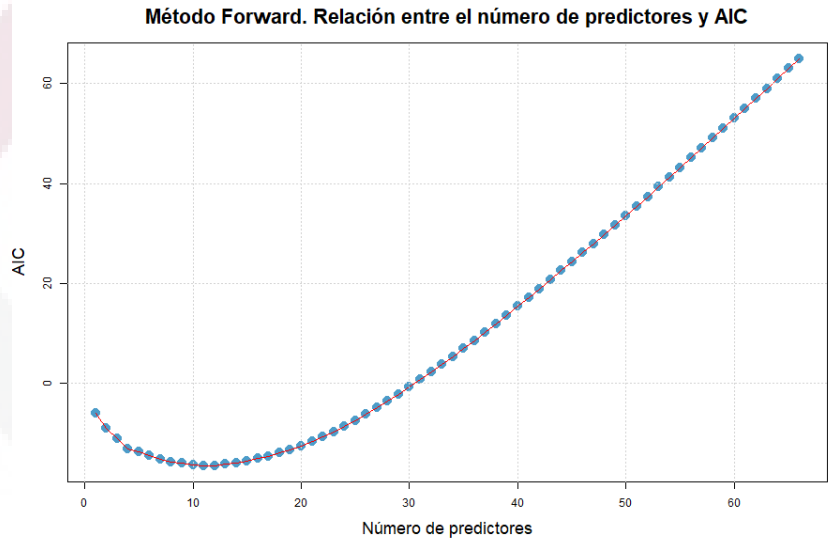

Figura 28: Relación entre el número de predictores y el AIC con el método Forward stepwise para los 68 predictores

Por último, con los resultados obtenidos con el criterio AIC sugieren que el mejor modelo contiene 12 predictores, ya que si observamos detenidamente la figura(28) nos podemos percatar que el mínimo se encuentra entre 10 y 20 predictores.

Cuadro 1: Variables seleccionadas según los criterios de selección de subconjuntos Forward Stepwise

Mejor R^2 ajustado	**Mejor BIC**	**Mejor C_p**
Capacidad instalada (MW)	Temperatura anual promedio	Población
Población		Temperatura anual promedio
PIB		Disponibilidad de biomasa
Empleos en energía renovable		Precios de electricidad
Políticas gubernamentales		Tasa de urbanización
Temperatura anual promedio		Tasa de industrialización
Disponibilidad de biomasa		Capacidad de manufactura local
Precios de electricidad		Índice de libertad económica
Subsidios energéticos		Fuerza laboral en el sector energético
Tasa de urbanización		Alianzas público-privadas en energía
Tasa de industrialización		Cooperación regional en energía renovable
Liberalización del mercado		Tipo de energía: hidroeléctrica
Programas educativos		País: EE.UU.
Capacidad de manufactura local		
Incentivos de exportación		
Estado de derecho		
Índice de libertad económica		
Fuerza laboral en energía		
Alianzas público-privadas		
Cooperación regional		
Tipo de energía: hidroeléctrica		
País: EE.UU.		

En la tabla(1) se encuantras las variables seleccionadas por el método Forwaed stepwise de acuerdo a cada criterio considerao, R^2, BIC y C_p respectivamente. Es importante notar que algunas variables están presentes en ambos criterios,R^2 y C_p , como Población, Temperatura anual promedio, Disponibilidad de biomasa, Precios de electricidad, Tasa de urbanización, Tasa de industrialización, Índice de libertad económica, Fuerza laboral en el sector energético, Alianzas público-privadas y Tipo de energía: hidroeléctrica.

En los gráficos siguientes se muestran las relaciones entre los predictores y los distintos criterios utilizados para seleccionar el número de predictores mediante el método Backward stepwise. Cada criterio indica un número óptimo diferente de predictores seleccionados.

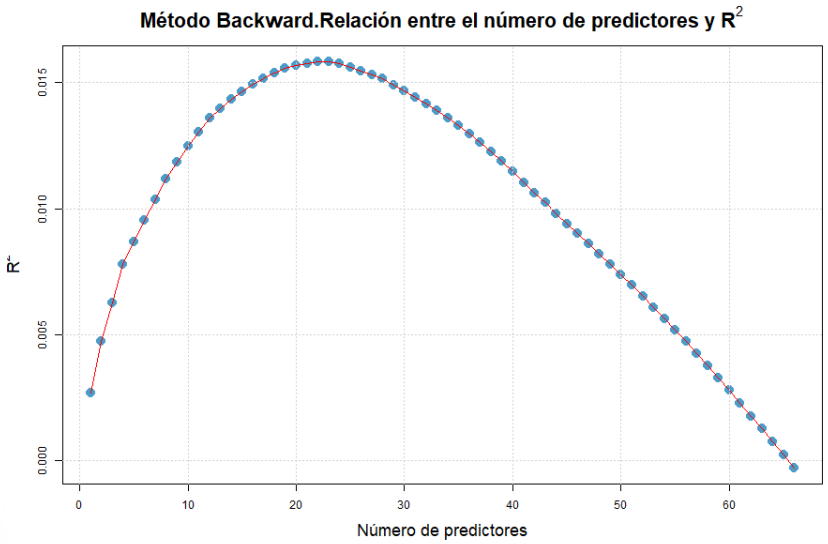

Figura 29: Relación entre el número de predictores y el R^2 con el método Backward stepwise para los 68 predictores

Los resultados indican que el modelo óptimo, de acuerdo con el R^2 ajustado, incluye 22 predictores. Esto se observa claramente en la figura (29), donde el valor máximo del R^2 ajustado se encuentra, precisamente, dentro del rango de 20 a 30 predictores.

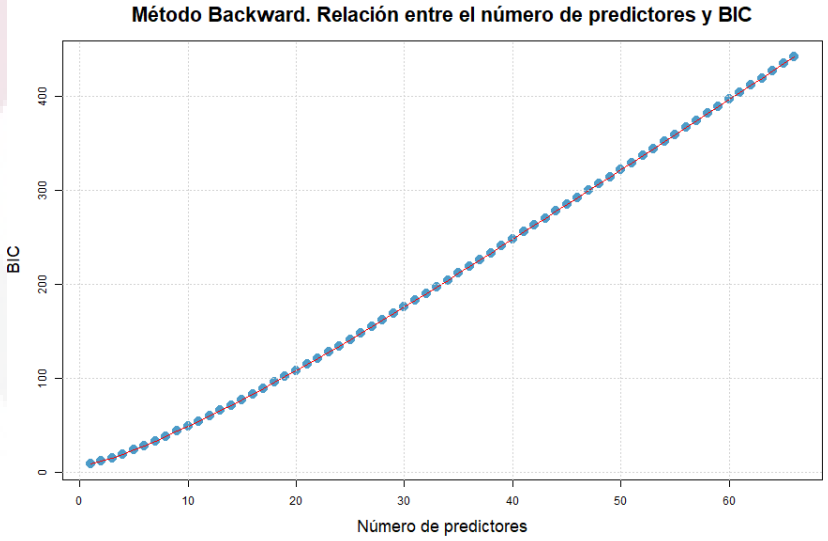

Figura 30: Relación entre el número de predictores y el BIC con el método Backward stepwise para los 68 predictores

Por otro lado, los resultados muestran que el modelo óptimo según el BIC incluye un único predictor. Esto se evidencia claramente en la figura (30).

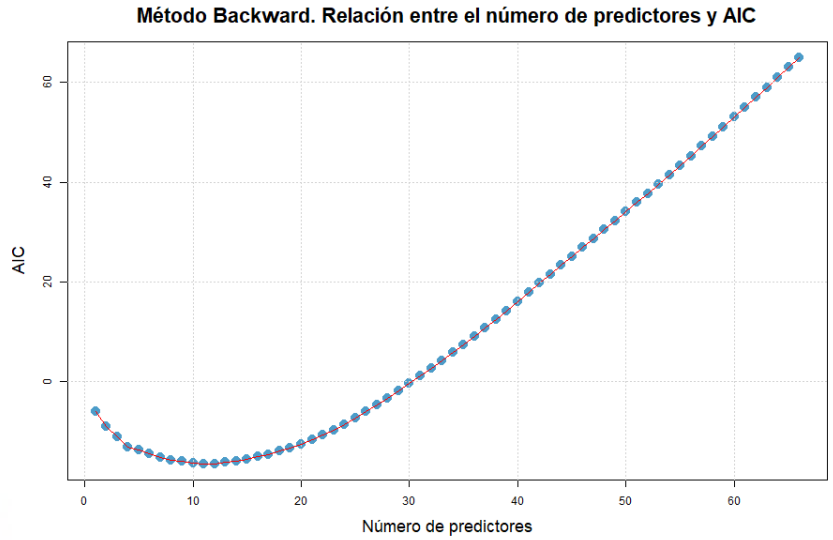

Figura 31: Relación entre el número de predictores y el AIC con el método Backward stepwise para los 68 predictores

Por último, los resultados obtenidos utilizando el criterio AIC indican que el mejor modelo contiene 12 predictores. Esto se puede apreciar en la figura (28), donde el valor mínimo se encuentra dentro del rango de 10 a 20 predictores.

Cuadro 2: Variables seleccionadas según los criterios de selección de subconjuntos por Backward stepwise

Mejor R^2 ajustado	**Mejor BIC**	**Mejor C_p**
Capacidad instalada (MW)	Temperatura anual promedio	Población
Población		Temperatura anual promedio
PIB		Disponibilidad de biomasa
Empleos en energía renovable		Precios de electricidad
Políticas gubernamentales		Tasa de urbanización
Temperatura anual promedio		Tasa de industrialización
Disponibilidad de biomasa		Capacidad de manufactura local
Precios de electricidad		Índice de libertad económica
Subsidios energéticos		Fuerza laboral en el sector energético
Tasa de urbanización		Alianzas público-privadas en energía
Tasa de industrialización		Cooperación regional en energía renovable
Liberalización del mercado		País: EE.UU.
Programas educativos		
Capacidad de manufactura local		
Incentivos de exportación		
Estado de derecho		
Índice de libertad económica		
Fuerza laboral en energía		
Alianzas público-privadas		
Cooperación regional		
Tipo de energía: hidroeléctrica		
País: EE.UU.		

En la tabla(2) se encuantras las variables seleccionadas por el método Backward stepwise de acuerdo a cada criterio considerao, R^2, BIC y C_p respectivamente Las variables que están presentes en ambos criterios, son: Población, Temperatura anual promedio, Disponibilidad de biomasa, Precios de electricidad, Tasa de urbanización, Tasa de industrialización, Capacidad de manufactura local, Índice de libertad económica, Fuerza laboral en el sector energético, Alianzas público-privadas en energía, Cooperación regional en energía renovable y País: EE.UU.

Variables seleccionadas por ambos métodos, son Población, Temperatura anual promedio, Disponibilidad de biomasa, Precios de electricidad, Tasa de urbanización, Tasa de industrialización, Capacidad de manufactura local, Índice de libertad económica, Fuerza laboral en el sector energético, como se observa en la tabla(3)

Cuadro 3: Variables seleccionadas según los criterios de selección de subconjuntos (Forward y Backward Stepwise)

Método Forward	**Método Backward**
Población	Población
Temperatura anual promedio	Temperatura anual promedio
Disponibilidad de biomasa	Disponibilidad de biomasa
Precios de electricidad	Precios de electricidad
Tasa de urbanización	Tasa de urbanización
Tasa de industrialización	Tasa de industrialización
Índice de libertad económica	Índice de libertad económica
Fuerza laboral en el sector energético	Fuerza laboral en el sector energético
Alianzas público-privadas	Alianzas público-privadas en energía
Tipo de energía: hidroeléctrica	Cooperación regional en energía renovable
	País: EE.UU.

Una tarea importante es conocer cuáles son dichos predictores, es decir saber cuales son las variables más importantes para el modelo ya que nos interesa predecir eh interpretar con esta base de datos. Para realizar esta tarea se realizo regresión Lasso para un conjunto de valores de λ, en el cual se obtuvo un λ óptimo y apartir de él se construyó el modelo por medio de regresión Lasso.

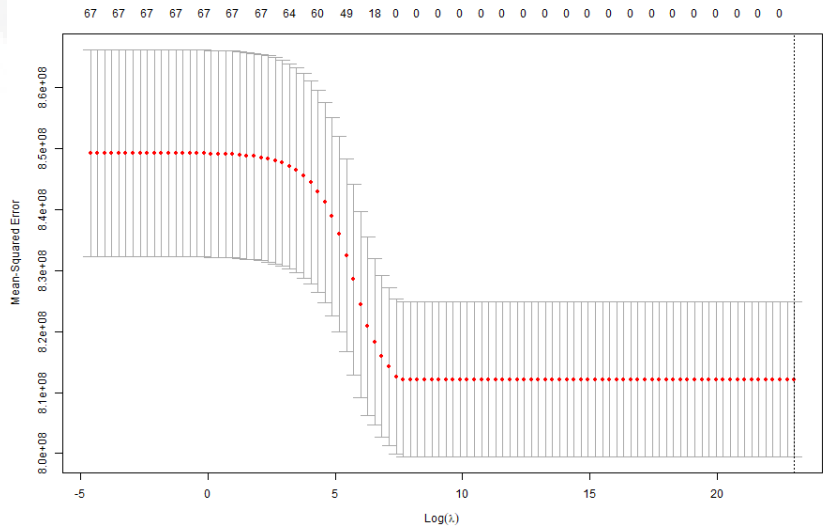

Figura 32: Gráfica MSE de prueba para cada valor de λ

De a cuerdo con los resultados obtenidos con el error de prueba MSE para cada valor de λ se obtuvo que un λ óptimo tiene el valor de 533.669 tiene el error de prueba más pequeño, lo cual se aprecia enla figura(32), también observamos que el eje horizontal está en logaritmo, por lo que $\log(533.669) \approx 6.2797$ que aproximadamente donde se muestra el mínimo de la función.

Posteriormente con este valor de $\lambda = 533.669$ óptimo se estimó el modelo de regresión Lasso obteniendo los siguientes resultados para los coeficientes

Variable	**Coeficiente**	**Significancia**
(Intercepto)	4.897919×10^{4}	
Capacidad instalada (MW)	-2.605683×10^{-2}	
Inversiones (USD)	-2.798848×10^{-8}	
Población	-4.203079×10^{-7}	
PIB	$-7.822208 \times 10^{-12}$	
Importaciones de energía	1.692486×10^{-3}	
Empleos en energías renovables	1.821450×10^{-3}	
Políticas gubernamentales	-4.629274×10^{2}	
Gasto en I+D	-3.773106×10^{-8}	
Temperatura media anual	9.519692×10^{1}	
Precipitación anual	6.621298×10^{-3}	
Disponibilidad de biomasa	-9.510842	
Precios de la electricidad	2.709779×10^{3}	
Subsidios energéticos	-2.192991×10^{-7}	
Tasa de urbanización	-3.039248×10^{1}	
Programas educativos sobre renovables	2.572473×10^{2}	
Capacidad de manufactura local	1.563312×10^{1}	
Incentivos de exportación de equipos	2.561883	
Calidad regulatoria	-4.699674×10^{1}	
Estado de derecho	-1.054721×10^{2}	
Control de la corrupción	-3.104592×10^{1}	
Índice de libertad económica	-1.614618	
Fuerza laboral del sector energético	6.312911×10^{-5}	
Asociaciones público-privadas en energía	5.956655×10^{2}	
Cooperación regional en energía	1.346638×10^{3}	
País: EE.UU.	-6.130102×10^{2}	
Tipo de energía: Biomasa	5.731401×10^{2}	
Tipo de energía: Solar	-6.972572×10^{2}	

De la tabla anterior sólo extraemos las que tienen coeficiente positivo en la base 10, ya que se consideran las que aportan más al modelo, los resultados se adjuntan en la siguiente tabla

Cuadro 5: Variables con exponente positivo, exponente 1 y sin exponente en la base 10

Variable	**Coeficiente**
(Intercepto)	4.897919×10^4
Temperatura media anual	9.519692×10^1
Precios de la electricidad	2.709779×10^3
Programas educativos sobre renovables	2.572473×10^2
Capacidad de manufactura local	1.563312×10^1
Calidad regulatoria	4.699674×10^1
Control de la corrupción	3.1045×10^1
Incentivos de exportación de equipos	3.1045×10^1
Estado de derecho	3.1045×10^1
Asociaciones público-privadas en energía	5.956655×10^2
Cooperación regional en energía	1.346638×10^3
Tipo de energía: Biomasa	5.731401×10^2
Disponibilidad de biomasa	-9.510842
Índice de libertad económica	-1.614618

Las variables por ambos métodos son las siguientes

Cuadro 6: Variables correspondientes al modelo Lasso, Forward y Backward

Variable	**Forward y Backward**
Temperatura media anual	**Temperatura media anual**
Precios de la electricidad	Disponibilidad de biomasa
Programas educativos sobre renovables	**Precios de la electricidad**
Capacidad de manufactura local	Tasa de urbanización
Calidad regulatoria	Tasa de industrialización
Control de la corrupción	**Capacidad de manufactura local**
Incentivos de exportación de equipos	**Índice de libertad económica**
Estado de derecho	Fuerza laboral en el sector energético
Asociaciones público-privadas en energía	Poblacion
Cooperación regional en energía	-
Tipo de energía: Biomasa	-
Disponibilidad de biomasa	-
Índice de libertad económica	-

Presentamos las graficas de la inclución de las variables para el modelo de regresión Lasso en la figura(33) y para la regresión Ridge en la figura (35), además de la validación cruzada que se incluye en la figura(34)

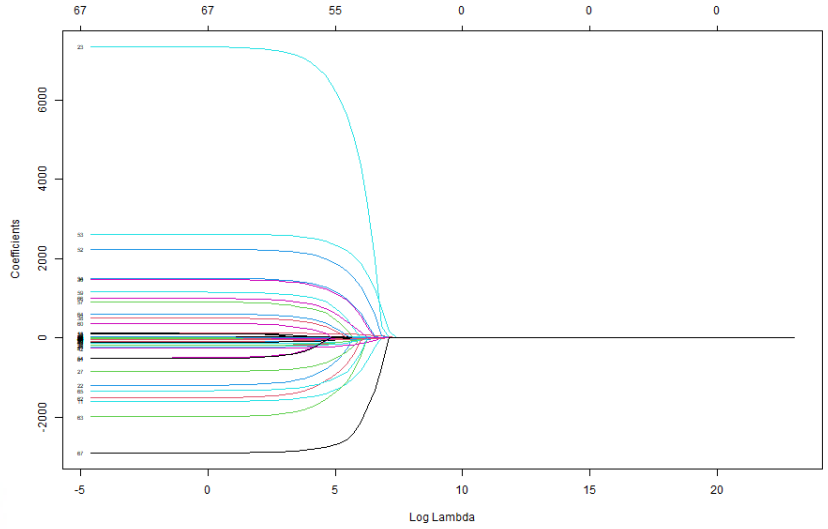

Figura 33: Gráfica inclusión de variables para cada valor de λ, para la regresión Lasso

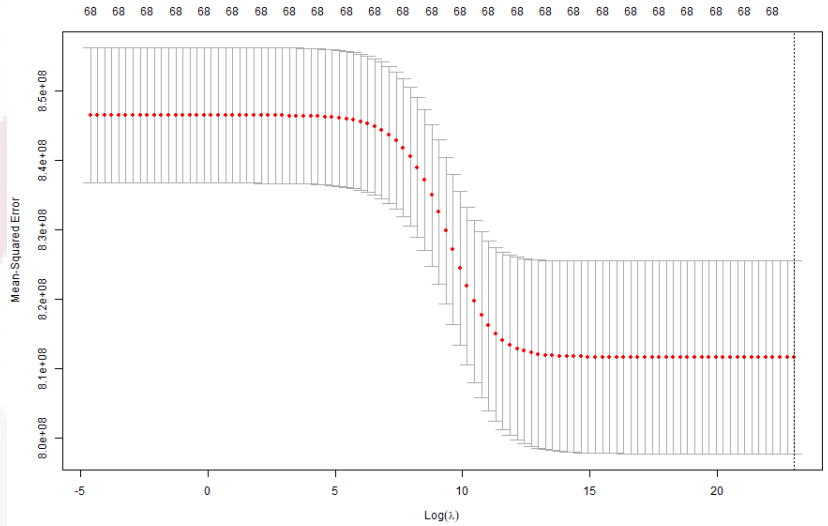

Figura 34: Gráfica Validación cruzada Regresión Ridgge

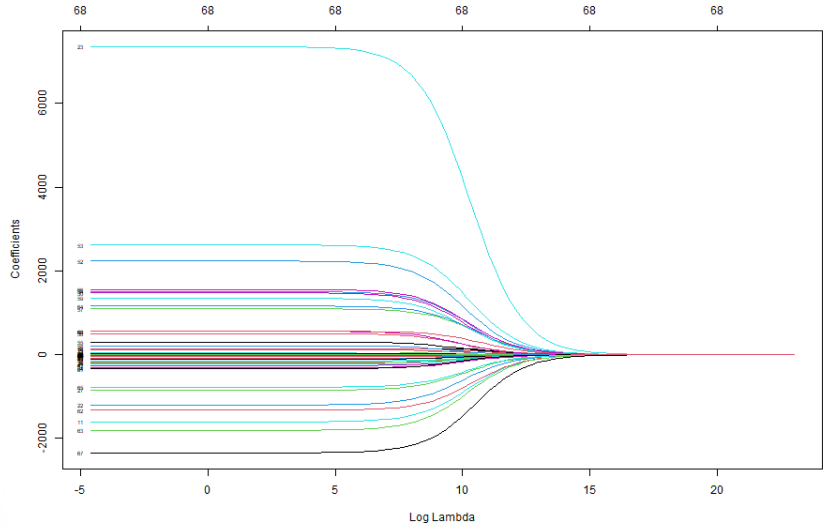

Figura 35: Gráfica inclusión de variables para cada valor de λ, para la regresión Ridge

6.3. Imputación de datos en la Variable Predictora

En esta sección se presenta el proceso de imputación de datos realizado en la variable predictora. Se decidió eliminar aleatoriamente el 30 % de los datos en dicha variable para ser posteriormente imputados utilizando distintos métodos. La base de datos, que incluye variables categóricas transformadas en variables dummy, cuenta con 2500 muestras y 69 variables. De las 2500 filas, se eliminaron 750 y se sustituyeron por valores NAN. Posteriormente, los datos faltantes fueron imputados utilizando los métodos de media, regresión, regresión estocástica y el método hot-deck (vecinos más cercanos), empleando como medida de distancia la distancia de Gower, dado que nuestro conjunto de datos incluye variables continuas y binarias. Los resultados de las imputaciones fueron evaluados mediante las métricas MAE, MSE y RMSE. Los resultados obtenidos se presentan en la Tabla 7.

Cuadro 7: Comparación de métricas de imputación

Método	MAE	MSE	RMSE
Media	25,460.95	945,070,008	30,741.99
Regresión	24,815.33	820,312,608	28,641.1
Regresión estocástica	31,364.66	1,501,964,566	38,755.19
Hot-deck	32,081.61	1,574,385,297	39,678.52

De los resultados presentados en la tabla anterior, se puede observar el desempeño de cada método en las distintas métricas. Se destaca que el mejor método fue la regresión, que obtuvo los menores valores en todas las métricas, seguido por la imputación por la media, luego la regresión estocástica y, por último, el método hot-deck. Cabe señalar que, aunque podría esperarse que la imputación por la media tenga el peor desempeño y el método hot-deck el mejor, este no fue el caso para este conjunto de datos. Esto se debe al comportamiento de los valores de la variable de respuesta. Como se observó en las series de tiempo y los resultados de los modelos ARIMA, estos valores oscilan alrededor de una media con cierta varianza, sin presentar tendencias o dinámicas más complejas. Por lo tanto, la imputación por la media no es una opción irrazonable en este contexto. Quizás un mejor resultado se hubiera obtenido si se hubiera añadido varianza a los valores imputados (imputación de Cohen), simulando mejor el comportamiento de los datos.

A continuación, se muestran las distribuciones de la variable de respuesta, tanto con los datos originales como con los datos imputados. En la Figura 36, se presenta la distribución de la variable de respuesta utilizando los datos originales.

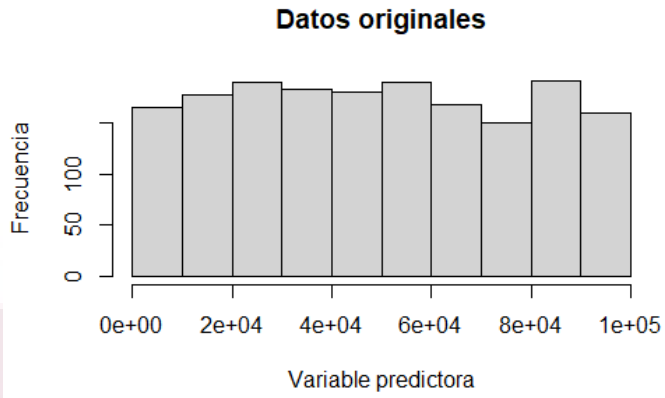

Figura 36: Distribución de la variable de respuesta con los datos originales.

En la Figura 37, se observan las distribuciones de la variable de respuesta para cada uno de los métodos de imputación.

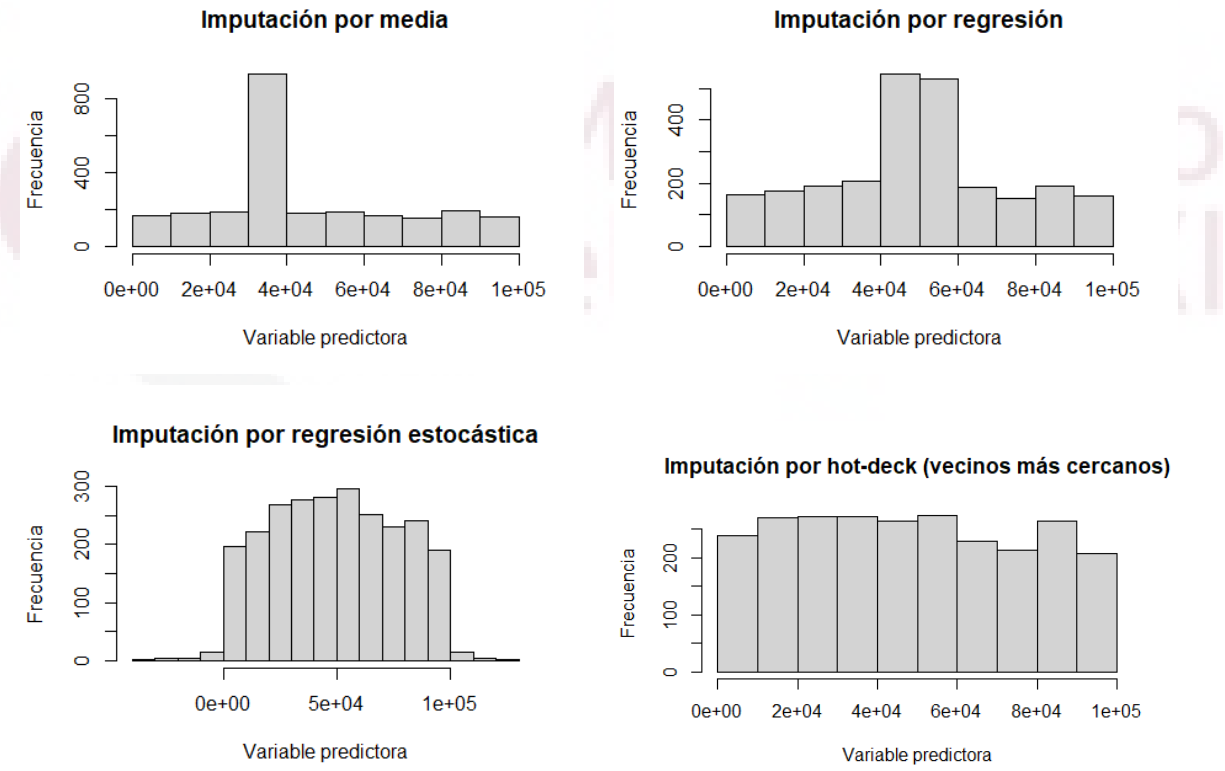

Figura 37: Distribuciones de la variable de respuesta para cada método de imputación.

Como se puede observar, las métricas obtenidas para los datos imputados pueden resultar engañosas. Aunque la regresión y la imputación por la media fueron los métodos con las mejores métricas, las

distribuciones de estas imputaciones no son similares a la distribución original. Específicamente, las imputaciones por media y regresión sobrestiman uno o dos intervalos de datos. En contraste, aunque la regresión estocástica y el método hot-deck presentaron peores puntuaciones en las métricas, sus distribuciones resultaron ser más adecuadas y más cercanas a la original, especialmente en el caso del método hot-deck, cuya distribución es muy similar a la original.

7. Conclusiones

Como conclusión podemos destacar que el comportamiento de las series de tiempo de la producción de energía para los países no presenta una dinámica compleja, por lo que pueden ser modelados fácilmente con modelos ARIMA con ordenes bajos en su parte autoregresivas y de medias moviles. Con el ajuste de estos modelos a las series de tiempo se espera que países como australia, brazil, canada, china y alemana siminuyan su producción de energía para el año 2024 mientras que los países como francia, india, japón, rusia y estados unidos incrementen su producción. Por otra parte bajo el análisis de las variables en los datos se concluye que las principales variables relacionadas con la producción de energía en los países es la temperatura promedio, capacidad de manufactura local, precios de electricidad y el índice de libertad económica.

La producción de energía renovable está influida principalmente por factores económicos y tecnológicos. Los precios de electricidad y la capacidad de manufactura local tienen un impacto directo al incentivar la inversión y reducir costos, mientras que el índice de libertad económica facilita el desarrollo de proyectos al mejorar el entorno regulatorio. Por otro lado, la temperatura media anual afecta de forma indirecta, siendo relevante según la fuente de energía predominante, como la solar. En conjunto, estas variables destacan la importancia de un entorno económico y tecnológico favorable para impulsar la transición energética.