FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Predicciones Globales de Fecundidad Mediante Técnicas de Aprendizaje Profundo

Trabajo Final de Grado para la Licenciatura en Estadística

Facultad de Ciencias Económicas y de Administración

Universidad de la República

Facundo Morini

Tutores

Dr. Daniel Ciganda
Mag. Francisco Piriz

Tribunal

Dra. Paola Bermolen
Dra. Natalia da Silva
Dr. Daniel Ciganda

Montevideo, Uruguay

Septiembre de 2025

Resumen

El proceso de declive sostenido de la fecundidad en todas las regiones del planeta ha agudizado la necesidad de obtener proyecciones demográficas precisas. En la actualidad, el uso de modelos bayesianos jerárquicos domina el campo de las proyecciones de fecundidad, siendo el enfoque utilizado por las Naciones Unidas y por los institutos de estadística de muchos países. Sin embargo, a pesar de presentar buenos resultados, estos modelos tienen ciertas debilidades, como la dificultad para incorporar información de otras variables que puedan mejorar los resultados de las proyecciones. En este trabajo se propone un método novedoso para la predicción de la Tasa Global de Fecundidad (TGF), planteando un enfoque de series temporales a partir de modelos de aprendizaje profundo que permiten la inclusión de información histórica relevante. Los resultados obtenidos, muestran que este enfoque es capaz de capturar la dinámica de la fecundidad en la mayoría de los poblaciones analizadas y de representar adecuadamente la incertidumbre asociada a las predicciones obtenidas.

Palabras clave— Proyecciones Demográficas, Fecundidad, Series de Tiempo, Aprendizaje Profundo.

Índice

Reproducibilidad y Código Abierto

Todo el código fuente, los conjuntos de datos utilizados y parte de los resultados de este documento están disponibles públicamente en el siguiente repositorio de GitHub:

https://github.com/FaMori/proyecciones-fecundidad-rnn

El proyecto fue desarrollado como una herramienta funcional que permite aplicar la metodología aquí descrita, ya sea sobre los datos presentados en este estudio o con conjuntos de datos propios, al igual que otras librerías de proyecciones demográficas.

Índice de cuadros

Índice de figuras

1. Introducción

Las proyecciones demográficas constituyen una herramienta fundamental para anticipar el tamaño y la composición futura de las poblaciones, siendo esenciales en la planificación de sistemas educativos y de salud, y para evaluar la sostenibilidad de los sistemas de seguridad social. La construcción de estas proyecciones se articula en torno a tres componentes: mortalidad, migración y la fecundidad. En un contexto de declive generalizado de la fecundidad, factor determinante en el envejecimiento de las poblaciones, este componente adquiere un papel clave, que ha traspasado la barrera de lo académico para convertirse en un tema de interés y preocupación general. Por ello, es necesario desarrollar y mejorar los modelos existentes para obtener proyecciones precisas de los componentes demográficos y de la incertidumbre asociada a estas predicciones.

La Tasa Global de Fecundidad (TGF) es uno de los indicadores más utilizados en la medición de la fecundidad. Esta se construye a partir de la suma de las tasas específicas de fecundidad por edad $({}_nf_x{}^{(t)})$ para un periodo determinado, que usualmente es de un año. La misma se obtiene dividiendo el número de nacimientos de mujeres en el grupo de edad $[x, x+n]$ durante el periodo t por los años-persona[1] aportados por mujeres en ese mismo rango de edad. Dichas tasas se estiman a partir de un análisis retrospectivo de la información proveniente de los registros de estadísticas vitales, la cual se complementa con datos de censos o encuestas de cobertura poblacional.

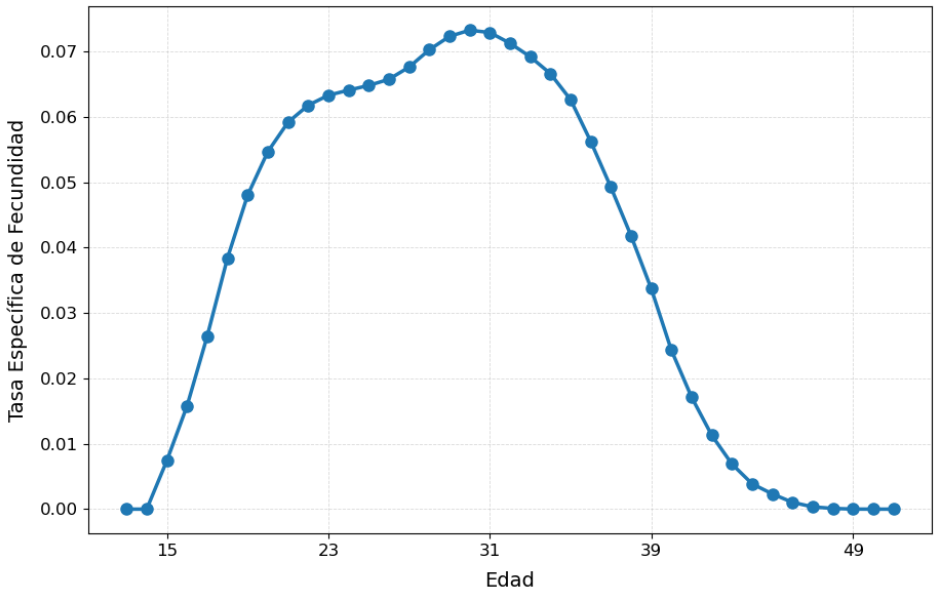

Figura 1: Tasa Específica de Fecundidad por Edad en Uruguay, año 2021. Cada punto mide la intensidad de la fecundidad a cada edad específica, durante ese año. Esto permite observar el cambio de la fecundidad según las distintas edades dentro del rango reproductivo de 15 a 49 años.

Fuente: División de Población, Naciones Unidas (NNUU).

[1]Los años-persona representan la suma de los periodos de tiempo que cada individuo de un grupo ha estado expuesto al riesgo de experimentar un evento de interés.

A partir de las mismas, se obtiene la TGF para el periodo t mediante la siguiente fórmula:

$$TGF^{(t)} = n \cdot \sum_{x} {}_nf_x{}^{(t)}$$

donde ${}_nf_x{}^{(t)}$ representa la tasa de fecundidad para el grupo de edad x. De esta manera, se construye un indicador que puede interpretarse como el número promedio de hijos que una cohorte hipotética tendría al final de su vida reproductiva si, a lo largo de ella, estuviera expuesta a las tasas de fecundidad por edad de un año en particular. Su principal ventaja radica en que, al no depender de la distribución por edades de la población, permite realizar comparaciones a lo largo del tiempo y entre distintas poblaciones.

En este ámbito, las proyecciones de la TGF de la División de Población de las Naciones Unidas (NNUU) se han consolidado como el estándar de referencia para demógrafos y especialistas en ciencias sociales. Actualmente, dichas proyecciones se elaboran mediante modelos bayesianos jerárquicos, los cuales aprovechan la información de todos los países para modelar la trayectoria individual de cada uno (Alkema et al., 2011).

Si bien estos modelos han demostrado una notable capacidad predictiva y permiten cuantificar la incertidumbre de las predicciones, también presentan limitaciones importantes. Por un lado, no incorporan de forma explícita determinantes socioeconómicos claves como el nivel de educación femenina o la prevalencia en el uso de métodos anticonceptivos, cuya influencia en los cambios de la fecundidad es ampliamente reconocida (Liu y A. Raftery, 2020). Por otro lado, descansan en supuestos debatidos, como la estabilización a largo plazo de la TGF en países de baja fecundidad, una premisa cuestionada por la evidencia empírica reciente que muestra un descenso continuado hasta niveles inferiores a un hijo por mujer en varios países. A estas debilidades se suma la rigidez de un enfoque que requiere modelar las diferentes dinámicas de la fecundidad por separado, utilizando criterios deterministas para delimitar cada etapa.

Frente a estas limitaciones, las técnicas de aprendizaje automático *(machine learning)* surgen como una alternativa prometedora, principalmente por su capacidad para modelar relaciones complejas directamente desde los datos, sin la necesidad de especificar previamente una forma funcional estricta. Dentro de esta disciplina, los algoritmos de aprendizaje profundo *(deep learning)*, que representan el área de mayor evolución y desarrollo en los últimos años, han ganado popularidad por sus excelentes resultados en áreas como el procesamiento del lenguaje natural (PLN) o el reconocimiento de imágenes, además de haber demostrado ser herramientas muy valiosas para el análisis y la predicción de datos temporales (Miller et al., 2024). En los últimos años, se han desarrollado diversas arquitecturas diseñadas específicamente para este fin, entre las que destacan modelos como *TFT* (Lim et al., 2021) basado en *Transformers*; *DeepAR* (Salinas et al., 2020) basado en redes recurrentes; o *N-BEATS* (Oreshkin et al., 2020) basado en redes *Multilayer Perceptron (MLP)*. Dichos métodos han demostrado ser competitivos incluso con conjuntos de datos pequeños, gracias a la existencia de estrategias de regularización y aumento de datos que permiten generalizar adecuadamente y mitigar el sobreajuste.

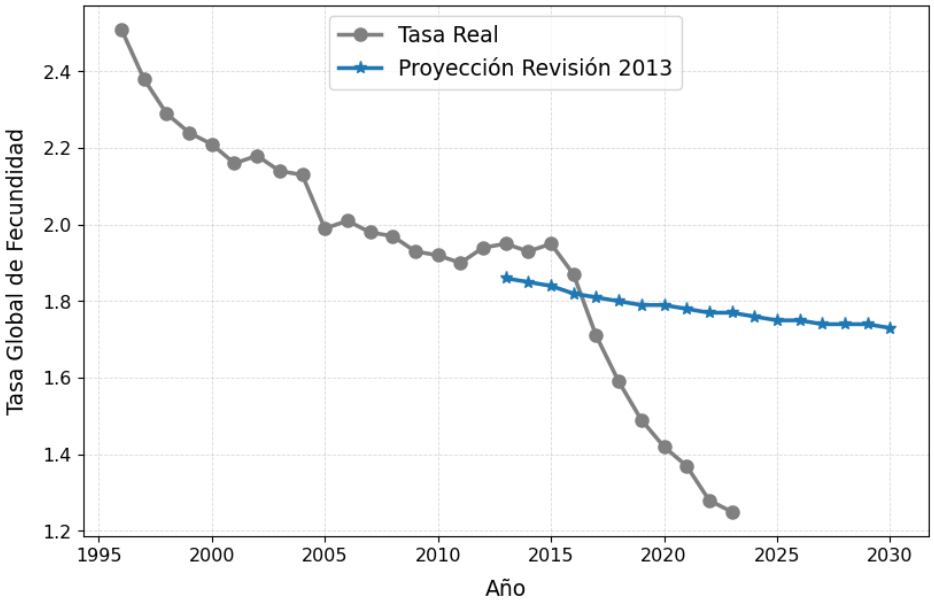

Figura 2: Tasa Global de Fecundidad (TGF) en Uruguay en las últimas décadas. El gráfico contrasta la proyección del INE (Revisión 2013) con la tasa real, cuyo descenso sostenido y más acentuado de lo previsto evidencia las limitaciones de los modelos tradicionales.

Fuente: Instituto Nacional de Estadística Uruguay (INE).

A pesar de este potencial, el uso de modelos de aprendizaje profundo todavía no ha sido ampliamente explorado en demografía. Si bien existen implementaciones puntuales, como las aplicadas al campo de la mortalidad para complementar modelos tradicionales como el de *Lee-Carter* (Lindholm y Palmborg, 2022; Wang et al., 2024), los resultados obtenidos no han llegado a justificar el uso sobre enfoques más establecidos. Sin embargo, es importante destacar que dichos trabajos se han centrado en evaluar la viabilidad de estos modelos, sin profundizar en la optimización de su rendimiento.

Este trabajo busca evaluar la viabilidad del uso de métodos de aprendizaje profundo en un problema demográfico como las proyecciones globales de fecundidad. En este sentido, se propone un modelo basado en redes neuronales recurrentes que es capaz de captar la estructura interna de los datos de fecundidad a nivel global, pudiendo en un único marco flexible modelar las distintas trayectorias que presentan los países, capturar relaciones no lineales y, a su vez, permitir la incorporación de información histórica relevante para las predicciones futuras.

Si bien este trabajo se centra en la proyección de la TGF, su objetivo principal es explorar el aprendizaje profundo como una metodología novedosa en demografía. Para ello, se evalúan herramientas de esta disciplina adaptadas a los desafíos de los datos demográficos, como es el caso de la cuantificación de la incertidumbre. Este último punto es crucial, ya que las estimaciones puntuales son insuficientes para el diseño de políticas públicas robustas. Finalmente, se incorporan prácticas del aprendizaje automático, como la evaluación sistemática del rendimiento de los modelos, un enfoque poco común en la literatura sobre proyecciones demográficas.

2. La Transición de la Fecundidad

A mediados de siglo XX, las regiones menos desarrolladas de Asia, África y América Latina promediaban aproximadamente 6 nacimientos por mujer, mientras que la mayoría de los países desarrollados presentaban tasas considerablemente más bajas menores a 3 hijos por mujer; para 2021, el promedio en el mundo en desarrollo había caído a alrededor de 2.6 hijos por mujer, y grandes partes del planeta operaban por debajo del nivel de reemplazo[2] como se observa en la Figura 3.

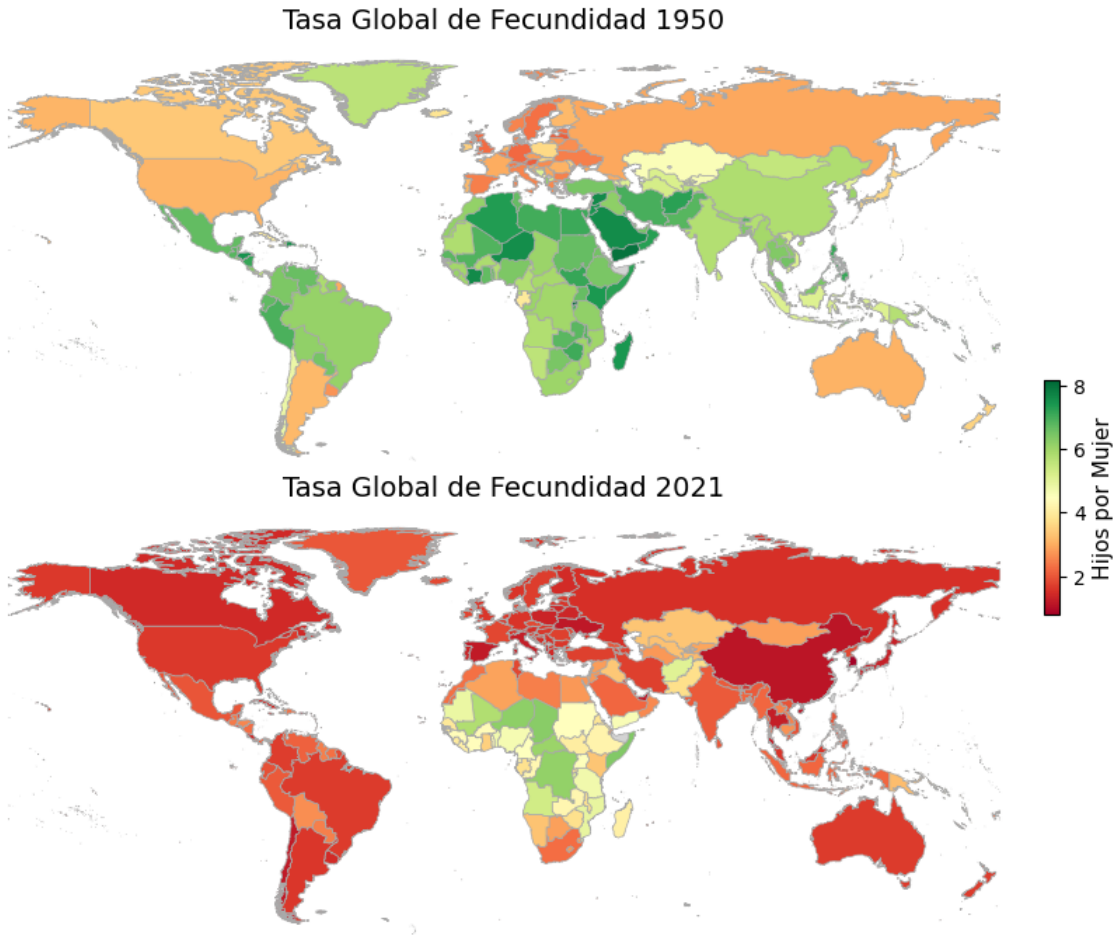

Figura 3: Tasa Global de Fecundidad (TGF) a nivel mundial, 1950 vs. 2021. Los mapas ilustran la profunda transición demográfica ocurrida en las últimas siete décadas. Para 2021 gran parte del planeta se encuentra en un descenso generalizado a niveles bajos o cercanos al de reemplazo, con África subsahariana como la principal región donde persisten niveles elevados.

Fuente: División de Población, Naciones Unidas (NNUU).

[2]Valor de la TGF de 2.1 hijos por mujer que aseguraría el reemplazo generacional en poblaciones sin contemplar la migración.

Esta drástica caída, si bien generalizada, no ha sido un proceso uniforme. La velocidad y el momento de la transición varían considerablemente entre regiones y países, lo que sugiere la influencia de una compleja interacción de determinantes. La literatura demográfica ha explorado extensamente cómo cambios a nivel socioeconómico, educativo y de salud como la mortalidad infantil o el acceso a métodos anticonceptivos actúan como motores de esta transformación.

2.1. Determinantes de la Transición

Históricamente, las sociedades anteriores a la transición demográfica se caracterizaban por niveles estables y elevados de fecundidad, un patrón observado tanto en la Europa del siglo XIX como en gran parte de África hasta bien entrada la segunda mitad del siglo XX. En este contexto, la TGF se situaba generalmente entre 6 y 7 hijos por mujer. En las economías predominantemente agrarias de la época, los hijos proveían mano de obra y constituían la principal forma de seguridad para la vejez, mientras que la elevada mortalidad infantil obligaba a las parejas a tener más hijos de los deseados como una estrategia de "sobreaseguramiento" frente a la incertidumbre de la supervivencia (Notestein, 1945). Este patrón era reforzado por una estructura social donde las oportunidades para las mujeres eran limitadas; los bajos niveles de educación y la escasa autonomía personal dejaban poco incentivo o capacidad para restringir la reproducción, perpetuando así las preferencias por familias numerosas (McDonald, 2000). En consecuencia, la fecundidad se regía principalmente por el espaciamiento biológico y prácticas tradicionales, con una limitada capacidad para su control consciente.

Este régimen de alta fecundidad comenzó a desarticularse con la industrialización, el crecimiento del ingreso y la urbanización, procesos que alteraron fundamentalmente los incentivos económicos de la paternidad: los niños dejaron de ser activos productivos para convertirse en dependientes (Bongaarts y Casterline, 2013). En paralelo a este cambio económico, la expansión de la educación amplió las oportunidades de vida de las mujeres y comenzó a generar una asociación negativa entre años de escolaridad y tamaño familiar. A estos factores estructurales se sumó la difusión de nuevas ideas, como el ideal de la familia nuclear, que se propagaron a través de medios de comunicación y redes personales, acelerando el descenso de la natalidad. La interacción de estos motores económicos, educativos e ideacionales es fundamental para comprender las distintas trayectorias de la fecundidad a nivel global.

De estos tres ejes de transformación, la expansión de la educación femenina es consistentemente señalada en la literatura como uno de los predictores más influyentes en el comportamiento reproductivo. Su impacto opera a través de múltiples mecanismos interrelacionados: un mayor nivel educativo tiende a aumentar la autonomía de las mujeres, mejora su conocimiento sobre salud reproductiva y eleva el costo de oportunidad de la maternidad al ampliar sus perspectivas laborales. Asimismo, la escolarización masiva acelera la difusión de nuevas ideas y valores, como el de la familia nuclear, modificando las aspiraciones de consumo. La evidencia empírica muestra una fuerte asociación entre ambas trayectorias; países que han experimentado un crecimiento sostenido en la educación femenina a menudo presentan descensos marcados y similares en sus niveles de fecundidad, como se ilustra en la Figura 4.

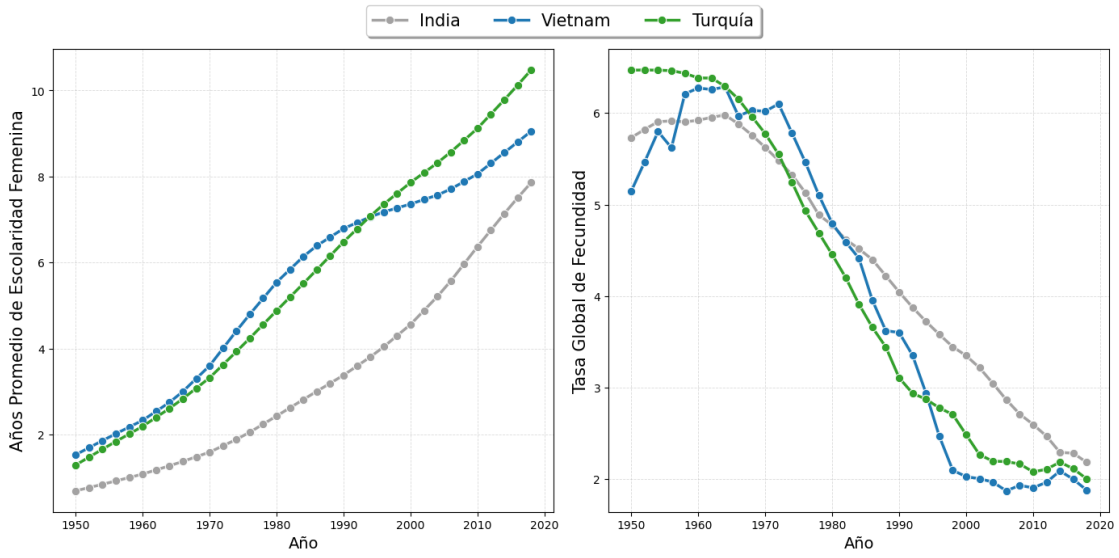

Figura 4: Evolución de la Escolaridad Femenina y la Tasa Global de Fecundidad (TGF) en India, Vietnam y Turquía (1950-2020). La figura ilustra cómo procesos similares de expansión educativa se corresponden con trayectorias de descenso de la fecundidad notablemente parecidas. El panel izquierdo muestra el crecimiento sostenido en los años de escolaridad femenina, mientras que el panel derecho refleja la consiguiente caída de la TGF en los mismos países.

Fuente: División de Población (NNUU) y Wittgenstein Centre.

Si bien la expansión educativa y el desarrollo socioeconómico son fundamentales para modificar las preferencias de fecundidad y fomentar el ideal de familias más pequeñas, estos cambios en la demanda no se traducen automáticamente en una menor natalidad. Para que las parejas puedan materializar sus nuevos deseos reproductivos, es indispensable el acceso y uso de métodos de control de la natalidad. Históricamente, en contextos de alta fecundidad, la capacidad de restringir la reproducción era limitada. La introducción y expansión de los programas de planificación familiar y los métodos anticonceptivos modernos fueron, por lo tanto, el mecanismo que permitió a las mujeres alinear su comportamiento reproductivo con sus preferencias, convirtiéndose en un catalizador crucial para la aceleración del descenso de la fecundidad.

El impacto de los programas de planificación familiar puede ser significativo incluso en países pobres con preferencias por familias numerosas. Etiopía es un ejemplo notable de esto; a pesar de tener uno de los niveles de desarrollo más bajos del mundo, la implementación de un programa de planificación familiar de alta calidad se asoció con una caída sustancial en la fecundidad deseada y un descenso considerable en la TGF total. De hecho, el país se destaca junto a otros como Malawi y Ruanda por haber logrado una reducción considerable en la fecundidad, en gran parte debido al impacto de sus programas (Bongaarts, 2020). Como se observa en la Figura 5, el rápido aumento en la prevalencia de métodos anticonceptivos modernos en Etiopía a partir de los años 90 presenta una marcada correspondencia inversa con la caída en su tasa de fecundidad, demostrando la capacidad de estos programas para acelerar la transición demográfica.

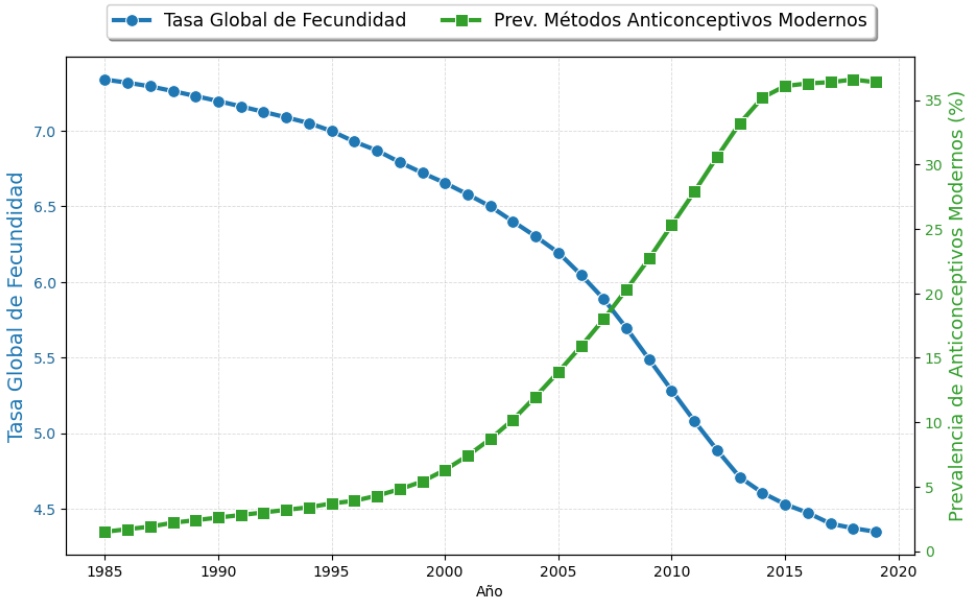

Figura 5: Tasa Global de Fecundidad y Prevalencia de Anticonceptivos Modernos en Etiopía (1985-2020). El gráfico ilustra el impacto de la inversión del país en programas de planificación familiar a partir de los años 90. Se observa cómo el pronunciado aumento en el uso de métodos anticonceptivos (línea verde) influye directamente en la trayectoria de la Tasa Global de Fecundidad (línea azul), acelerando su descenso.

Fuente: División de Población (NNUU).

Sin embargo, la fuerte influencia de determinantes como la educación y la planificación familiar tiende a atenuarse a medida que los países avanzan en su transición y alcanzan niveles de fecundidad bajos. En este nuevo contexto demográfico, donde los impulsores clásicos pierden protagonismo, la variación de la fecundidad empieza a depender de un conjunto diferente de factores. El descenso se desacelera y los cambios en el calendario reproductivo, como el aplazamiento sistemático del primer hijo, adquieren un papel central; la media de edad al primer nacimiento supera los 30 años, reduciendo la ventana de oportunidad para alcanzar el número deseado de hijos (Sobotka, 2017). Asimismo, la dinámica pasa a estar influenciada por factores culturales (nuevos estilos de vida o la aceptación de la vida sin hijos) y por el diseño de políticas públicas que faciliten la compatibilidad entre la vida familiar y laboral, como licencias parentales y servicios de cuidado infantil accesibles.

2.2. Fases de la Tasa Global de Fecundidad

La evolución heterogénea de la fecundidad y sus determinantes a lo largo del tiempo ha consolidado un marco conceptual predominante para explicar este proceso. Dicho marco, utilizado como base por organismos como las Naciones Unidas, segmenta la transición en un modelo de tres fases distintas. La trayectoria de algunos países se ajusta a esta estructura, siendo Vietnam un caso que ha atravesado por las distintas fases en las últimas décadas y que ilustra con claridad dicho comportamiento (Figura 6).

- **1. Fase I - Pre-Transicional:** La primera fase del modelo corresponde al régimen de alta fecundidad que prevalecía históricamente. Como se analizó, este sistema era una respuesta a un entorno de elevada mortalidad infantil, economías agrarias y limitadas oportunidades para las mujeres. Demográficamente, esta etapa se define por niveles de fecundidad estables, con una TGF que fluctuaba en torno a 6 o 7 hijos por mujer sin mostrar un declive sostenido. Una gran parte de los países del mundo ya han superado esta fase, la cual comenzó a concluir en las décadas de 1960 y 1970 para gran parte de Asia y América Latina, mientras que en los países de África subsahariana persistió hasta la ultimas décadas.

- **2. Fase II - Transición de la Fecundidad:** La segunda fase se define por el inicio y desarrollo del declive sostenido de la TGF, desde los altos niveles de la Fase I hasta alcanzar, o incluso situarse por debajo, del nivel de reemplazo. A diferencia del período anterior, esta etapa se caracteriza por una caída acelerada y generalmente irreversible de la fecundidad.

- **3. Fase III - Post-Transicional:** La fase post-transicional se inicia cuando la TGF desciende por debajo del nivel de reemplazo, situación actual de la mayoría de los países del mundo. Técnicamente, esta etapa supone un relativo equilibrio, en el cual la tasa presenta cierta recuperación cuando se alcanzan niveles bajos de fecundidad. Este período se distingue de la fase anterior por tres características clave: la centralidad del aplazamiento de la fecundidad, la primacía de nuevos factores culturales e institucionales, y un giro de las políticas públicas desde un enfoque anti-natalista hacia uno pro-natalista.

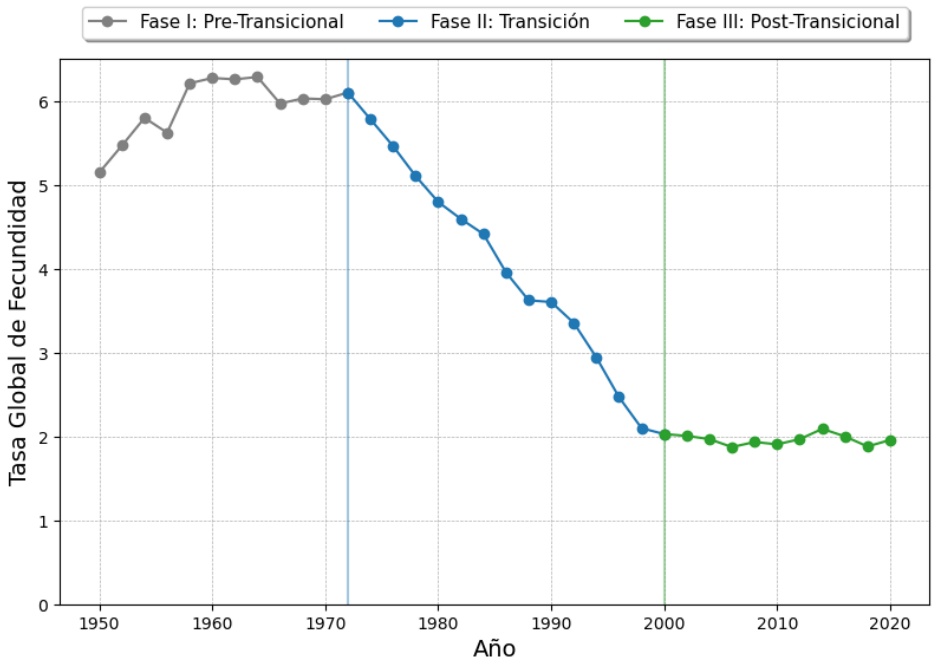

Figura 6: Datos históricos de la TGF en Vietnam desde 1950 hasta 2020. La trayectoria de este país es uno de los casos que cumple con claridad el comportamiento descrito por el modelo de tres fases.

Fuente: División de Población (NNUU), comienzos de fases estimados.

El modelo de tres fases ofrece un marco simplificado que ha sido fundamental para entender el comportamiento histórico de la fecundidad. Sin embargo, como todo modelo, es una visión que no logra capturar la creciente complejidad de las dinámicas actuales, especialmente en la fase post-transicional, donde un número creciente de países exhibe tasas muy por debajo del nivel de reemplazo.

Como se ilustra en la Figura 7, las trayectorias de los países con baja fecundidad presentan un comportamiento marcadamente heterogéneo. Mientras algunos, como Alemania, muestran una relativa estabilidad con un leve repunte reciente, otros, como Uruguay, rompen una tendencia estable previa para iniciar un nuevo descenso en la fecundidad en los últimos años. Un caso extremo es el de la República de Corea, cuya fecundidad ha seguido una caída continua hasta niveles sin precedentes, situándose por debajo de un hijo por mujer, lo que lo ha convertido en un gran foco de estudio en la demografía. Estas trayectorias, lejos de la estabilización propuesta originalmente por la fase post-transicional, marcan las limitaciones del modelo. La evidencia empírica de las últimas décadas no solo cuestiona la idea de una convergencia hacia un equilibrio, sino que demuestra la emergencia de nuevos patrones demográficos no previstos por la teoría clásica. Este escenario de divergencia exige un replanteamiento de los paradigmas establecidos y la búsqueda de modelos más flexibles que se adapten a estas nuevas realidades.

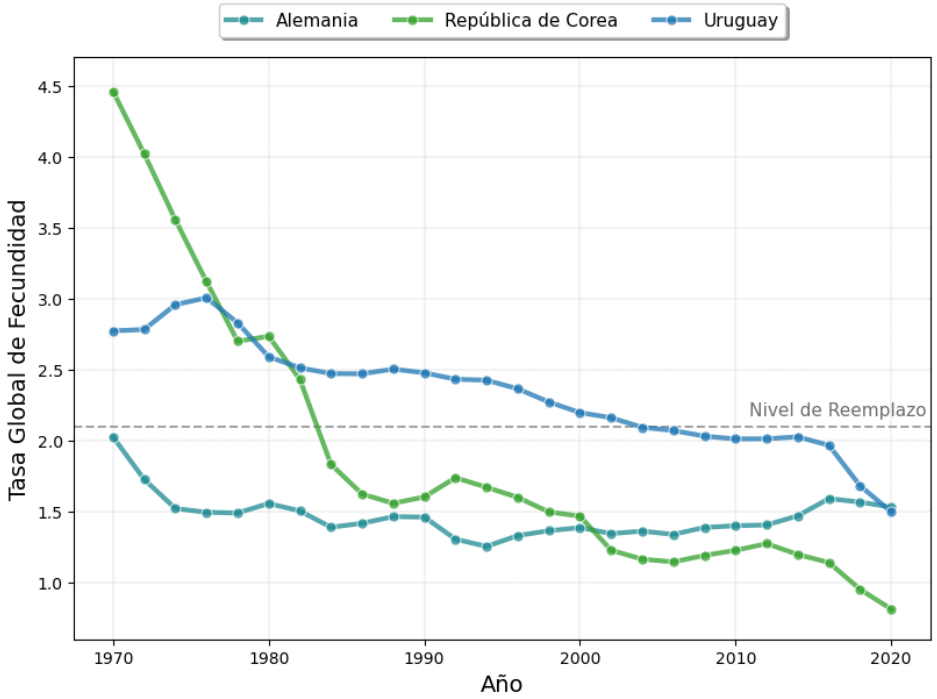

Figura 7: Trayectorias en la Fecundidad Post-Transicional (1970-2020). Se ilustra el comportamiento heterogéneo de la TGF en países que han superado el nivel de reemplazo. Mientras Alemania muestra una relativa estabilización con un leve repunte, Uruguay presenta un descenso en los últimos años, mientras la República de Corea ejemplifica una caída continua a niveles ultra-bajos. Estas trayectorias cuestionan la idea de una convergencia a un equilibrio estable en la fase post-transicional.

Fuente: División de Población (NNUU).

2.3. Proyecciones de la Fecundidad: Modelos Bayesianos Jerárquicos

A pesar de que el marco de tres fases presenta ciertas debilidades en el contexto actual, este sigue siendo la base conceptual sobre la que se construye el modelo probabilístico estándar utilizado por las Naciones Unidas (Alkema et al., 2011). En lugar de aplicar un único modelo para toda la trayectoria de la fecundidad, este enfoque segmenta el proceso, aplicando un modelo estadístico diferente para la fase de transición y para la fase post-transicional, basándose en criterios deterministas para definir el paso de una a otra.

Dado que el interés se centra en las proyecciones, la Fase I (pre-transicional) no se modela explícitamente. El modelo asume que cualquier país que se encuentre aún en esta fase al final del periodo de observación comenzará su transición en el periodo inmediatamente posterior. Por lo tanto, el marco de proyección se concentra exclusivamente en las Fases II y III.

Modelado de la Transición de la Fecundidad:

Para los países que se encuentran en pleno declive de su fecundidad, el modelo establece el inicio de esta fase mediante la identificación del último gran pico antes del descenso sostenido. Específicamente, la regla busca el pico de fecundidad más reciente que sea superior a 5.5 hijos por mujer, y cuya magnitud no difiera en más de 0.5 hijos del máximo histórico registrado para ese país. Si esta condición no se cumple (lo que podría ocurrir en países con transiciones muy tempranas), se utiliza el primer año de la serie de datos como punto de partida. Para esta fase, el modelo propuesto es un paseo aleatorio con una deriva no constante. La idea es que, en cada periodo, la TGF disminuye una cierta cantidad (la deriva) más un término de error aleatorio. La fórmula se puede expresar como:

$$f_{c,t+1} = f_{c,t} - d_{c,t} + \epsilon_{c,t}$$

donde $f_{c,t}$ es la TGF del país c en el periodo t, $d_{c,t}$ es el término que modela el decrecimiento de la tasa en el periodo de transición y $\epsilon_{c,t}$ es el componente aleatorio el cual se supone que se distribuye normal con varios parámetros que modelan el desvio.

Por otro lado el componente $d_{c,t}$ se modela por medio de una función paramétrica que especifica el decrecimiento en 5 años de la TGF en función de la tasa en ese periodo y un vector de parámetros θ. En este caso se utiliza una función doble logística g como se observa en la Figura 8 donde $\theta_c = (\Delta_{c1}, \Delta_{c2}, \Delta_{c3}, \Delta_{c4}, d_c)$ son los parámetros específicos para cada país, que marcan el ritmo de disminución de la tasa.

Dado que la cantidad de observaciones por país en esta fase puede variar según si el mismo ha recién comenzado o terminado la transición, se justifica el uso de modelos bayesianos jerárquicos, en pos de compartir información valiosa sobre la dinámica de la TGF entre países. En este caso, la jerarquía utilizada es mundo-país, lo que permite al modelo ajustar los parámetros no solo con la experiencia

pasada de un país, sino también incorporando la información del resto del mundo.

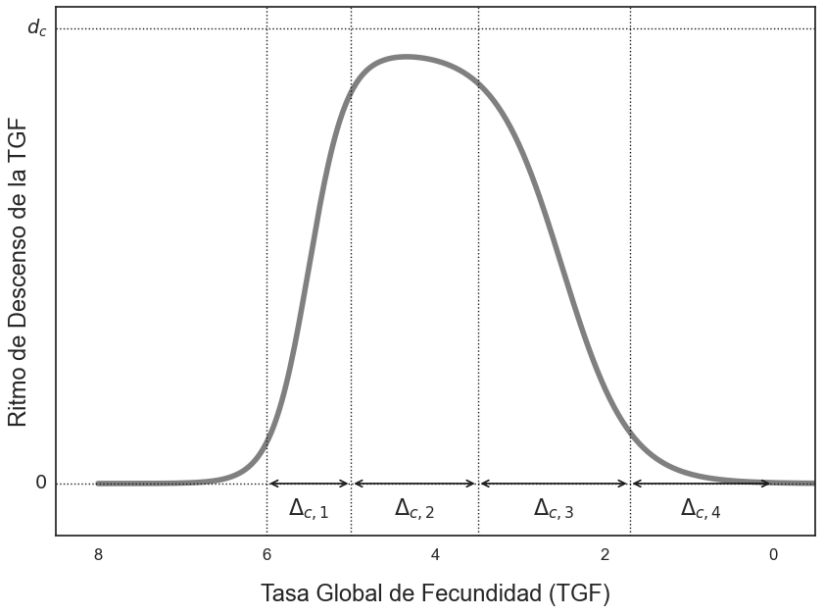

Figura 8: Función doble logística $g(f_{c,t}; \theta_c)$ que modela el ritmo del decaimiento en la TGF para un país c. La curva ilustra cómo la velocidad del descenso varía a lo largo de la transición: es inicialmente lenta en los niveles altos de fecundidad, se acelera en la etapa intermedia y vuelve a desacelerarse al llegar a valores más bajos.

Fuente: Elaborado en base a A. E. Raftery y Ševčíková (2023)

En este enfoque, no se emplea una jerarquía más desarrollada, como regiones, debido a que consideran que el proceso de la TGF no siempre es similar entre países vecinos, pudiendo presentar diferencias significativas.

Modelado de la Fase Post-Transición:

Una vez que un país ha completado su transición, el modelo cambia a un proceso autorregresivo de primer orden (AR(1)). El comienzo de esta fase se establece de forma determinista como el primer período donde se observan dos aumentos consecutivos de la TGF por debajo de un nivel de 2 hijos por mujer. Donde la TGF fluctúa en torno al 'nivel final de fecundidad' específico de cada país μ_c quedando la siguiente expresión:

$$f_{c,t+1} = \mu_c + \rho_c(f_{c,t} - \mu_c) + \epsilon_{c,t}$$

donde $|\rho_c| < 1$ para ser un proceso estacionario y $\epsilon_{c,t}$ es el componente aleatorio especifico para cada país que distribuye normal con media 0 y desvió σ_ϵ. Estos parámetros se estiman mediante un modelo jerárquico bayesiano con la misma estructura jerárquica (mundo-país).

El enfoque utilizado por NNUU es un marco robusto, fundamentado en la teoría demográfica y perfeccionado durante años. No obstante, su propia estructura expone rigideces inherentes que justifican la exploración de nuevas metodologías.

En primer lugar, su naturaleza univariada, si bien aporta simplicidad, omite por diseño la valiosa información contenida en covariables socioeconómicas que, como la propia teoría demográfica reconoce, son determinantes en la dinámica de la TGF y podrían mejorar el poder predictivo. En segundo lugar, el modelo depende de una estructura predefinida, con formas funcionales (como la doble logística) y reglas deterministas para definir las transiciones entre fases. Esta estructura, basada en criterios expertos, se convierte en una debilidad en el contexto demográfico actual, donde emergen patrones de fecundidad sin precedentes que desafían los supuestos teóricos más tradicionales, como el de la estabilización en la fase post-transicional.

Estas limitaciones abren la puerta a evaluar paradigmas alternativos. Entre ellos, destaca el aprendizaje automático y, particularmente, el aprendizaje profundo. Estos modelos suelen emplearse bajo el enfoque de aprendizaje supervisado, cuyo objetivo es aproximar una función f que relacione un conjunto de variables de entrada X (covariables) a una variable de interés y.

$$y = f(X) + \epsilon$$

A diferencia de los modelos paramétricos, no se asume una forma funcional específica para f. En su lugar, la función se construye de forma flexible a partir de las relaciones internas que el algoritmo detecta en los datos durante el entrenamiento, permitiendo la incorporación de múltiples variables sin necesidad de imponer una estructura predefinida.

3. Modelado de la Fecundidad Mediante Redes Neuronales Recurrentes

La aplicación de modelos de aprendizaje profundo a la proyección de la fecundidad, si bien prometedora dadas las limitantes de los modelos actuales, exige una metodología que aborde los desafíos inherentes a la naturaleza de los datos. La Tasa Global de Fecundidad (TGF) es una serie temporal, y a diferencia de los problemas de regresión clásicos donde las observaciones se asumen independientes, esta posee una dependencia secuencial que requiere un tratamiento específico.

En este contexto, la elección del aprendizaje profundo como paradigma se fundamenta en un conjunto de ventajas clave. Este enfoque ofrece un marco de trabajo que se distingue por:

1. **La capacidad de incorporar covariables de forma flexible:** A diferencia de los modelos más tradicionales, las redes neuronales permiten la inclusión de múltiples determinantes socioeconómicos para enriquecer las proyecciones sin imponer relaciones predefinidas entre ellos.

Asimismo, esta flexibilidad permite construir modelos para analizar relaciones complejas utilizando únicamente datos históricos, sin la necesidad de proyectar valores a futuro de estas covariables.

2. **La ausencia de supuestos estructurales rígidos**: Como se mencionó, estos modelos no asumen una forma funcional específica para la función f. En su lugar, la aprenden a partir de los datos, lo que los hace ideales para capturar las relaciones complejas y no lineales que caracterizan la dinámica de la fecundidad.

3. **La especialización en el modelado de datos secuenciales:** El campo del aprendizaje profundo ofrece arquitecturas especializadas, como las Redes Neuronales Recurrentes (RNN), diseñadas explícitamente para procesar secuencias y capturar la dependencia temporal característica de las series de tiempo.

4. **La versatilidad para cuantificar la incertidumbre**: La misma arquitectura neuronal puede ser adaptada con relativa facilidad para generar predicciones probabilísticas, permitiendo cuantificar la incertidumbre, un requisito fundamental en el campo de las proyecciones demográficas.

Sin embargo, para materializar estas ventajas, la implementación de un modelo de aprendizaje profundo en este dominio exige una metodología que aborde una serie de desafíos específicos. Es fundamental adaptar el paradigma de aprendizaje supervisado a la estructura secuencial de las series de tiempo, seleccionar una arquitectura capaz de capturar dependencias temporales, gestionar la alta cantidad de parámetros del modelo para evitar el sobreajuste con datos de longitud limitada.

3.1. Enfoque de Modelado Global

La naturaleza flexible de los modelos de redes neuronales, caracterizada por un elevado número de parámetros, presenta un desafío particular en la predicción de series temporales, especialmente en campos como las ciencias sociales y la econometría, donde los conjuntos de datos suelen ser de tamaño limitado. Esta combinación de alta complejidad del modelo y relativa escasez de datos incrementa el riesgo del sobreajuste, uno de los problemas más comunes en el aprendizaje automático. Este fenómeno ocurre cuando el modelo, en lugar de aprender la señal subyacente de los datos, memoriza el ruido y las particularidades de la muestra de entrenamiento, lo que resulta en una pérdida de su capacidad para generalizar a datos no observados.

Para mitigar este riesgo y, a la vez, aprovechar la capacidad de los modelos complejos, en los últimos años ha cobrado fuerza el paradigma de los modelos de predicción globales (Global Forecasting Models) (Hewamalage, Bergmeir y Bandara, 2022). A diferencia del enfoque tradicional que ajusta un modelo individual para cada serie, la estrategia global consiste en entrenar un único modelo con un conjunto de cientos o miles de series temporales. Bajo el supuesto de que las series comparten un mecanismo de generación de datos común, este enfoque permite que el modelo aprenda patrones más generales y robustos. Esta metodología se adapta de forma ideal al problema planteado, donde se

dispone de trayectorias de fecundidad para un gran número de países, pero con una cantidad limitada de observaciones históricas para cada uno.

Una de las ventajas principales de este enfoque es su capacidad para aprender representaciones latentes de la dinámica de la fecundidad. El punto central de este aprendizaje colectivo es que la experiencia histórica de los países que ya han completado su transición informa directamente las proyecciones de aquellos que aún se encuentran en pleno descenso. La naturaleza flexible de la red neuronal le permite capturar los comportamientos desiguales que caracterizan a los distintos países. Al entrenar al modelo con la totalidad de las trayectorias, permite que una única estructura se adapte a dinámicas muy diferentes, desde caídas pronunciadas hasta estabilizaciones o leves recuperaciones, sin necesidad de imponer reglas deterministas.

Asimismo como en parte ya se mencionó, el entrenamiento conjunto con las series de todos los países actúa como un potente mecanismo de regularización. Al exponer al modelo a una gran diversidad de trayectorias, se reduce el riesgo de sobreajuste a las particularidades o al ruido específico de la serie de un único país. Este efecto regularizador incentiva a la red a aprender patrones más generales y robustos de la evolución de la fecundidad, mejorando su capacidad de generalización.

3.2. Estructuración de Datos para Series Temporales

Más allá de los datos que se utilizan para entrenar el modelo, un paso fundamental es la transformación de los mismos para que puedan ser procesados por los algoritmos de aprendizaje profundo. Como se mencionó, el entrenamiento en aprendizaje supervisado se basa en aprender una función que mapea un conjunto de entradas X a una salida y. En el contexto de series temporales, esto exige una reformulación del problema: la "entrada" X se define como una secuencia de datos históricos y la "salida" y como la secuencia de valores futuros que se desean predecir.

Para realizar esta transformación de un problema temporal a uno de regresión, se recurre a la técnica de *ventanas deslizantes*. Este método consiste en crear un conjunto de datos de entrenamiento mediante la creación de ventanas sucesivas de tamaño fijo que se desplazan a lo largo del tiempo. Para el enfoque multivariado y de horizonte múltiple que se plantea en este trabajo, cada ventana se compone de:

- *Entrada (Input)*: Una secuencia de i observaciones históricas que incluye tanto los valores pasados de la TGF como un conjunto de k covariables asociadas a esos mismos períodos.
- *Etiqueta (Label)*: La secuencia de h observaciones futuras de la TGF que se desea predecir.

Al deslizar esta ventana en pasos sucesivos a través de la serie temporal de cada país, se genera un gran conjunto de pares $(entrada, etiqueta)$ que constituyen los datos de entrenamiento para el modelo. De esta manera, el problema de predicción se reformula como la tarea de aprender una función f que mapea secuencias de datos históricos a secuencias de valores futuros. Para un instante t, esta relación se puede expresar formalmente como:

$$(y_{t+1}, \ldots, y_{t+h}) = f(y_{t-i+1}, \ldots, y_t, X_{t-i+1}, \ldots, X_t) + \epsilon$$

La elección del tamaño de la ventana de entrada, i, es un paso crucial, ya que debe ser lo suficientemente largo para capturar la estructura temporal y las dependencias relevantes del proceso que se busca modelar. A su vez este marco puede generalizarse para problemas más complejos, como la predicción de múltiples series de forma simultánea (salida multivariante) o la inclusión de covariables cuyos valores futuros son conocidos.

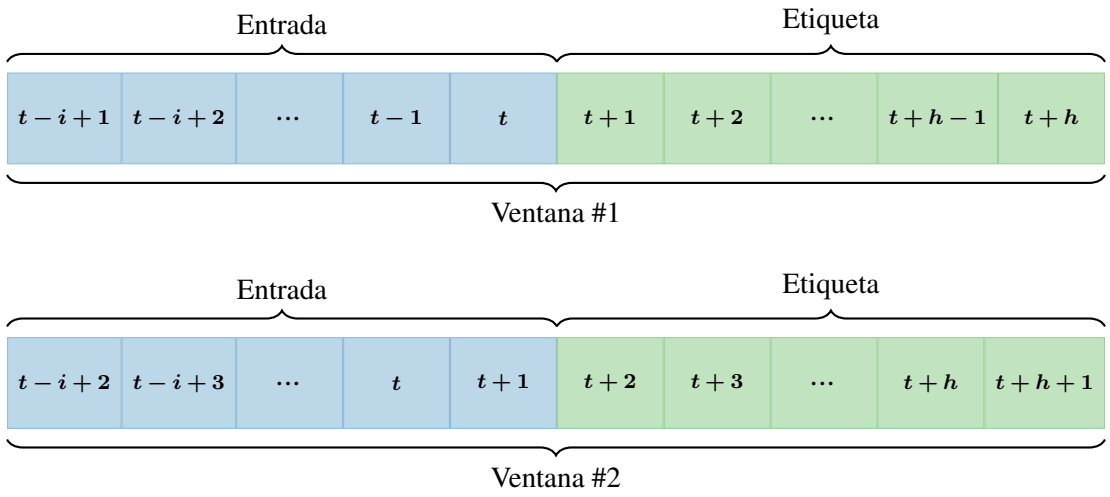

Figura 9: Ilustración del método de ventanas deslizantes a un paso para series temporales. Cada nueva ventana se genera desplazándose un paso hacia adelante, manteniendo un tamaño de ventana fijo.

3.3. Arquitectura del Modelo

El mundo de los modelos de aprendizaje profundo para datos secuenciales y los aplicables a series de tiempo está en constante evolución, surgiendo todo el tiempo nuevos tipos de arquitecturas que demuestran funcionar para el problema. Desde redes convolucionales, recurrentes, transformers o la aparición de modelos fundacionales de series temporales (TSFM), el abanico de arquitecturas posible se extiende cada año. En este trabajo dado que se trata de una evaluación inicial y poco explorada en el contexto demográfico, se optó por cierta simplicidad dentro de la materia con lo cual se centró principalmente en el uso de redes recurrentes, una arquitectura "clásica" dentro del aprendizaje profundo.

3.3.1. Redes Neuronales Recurrentes

Las redes neuronales recurrentes (RNN) son un tipo de algoritmo de aprendizaje profundo especialmente diseñado para trabajar con datos secuenciales. Mediante un proceso recurrente, en cada instante

t se ingresa una entrada de información x_t y se computa un estado oculto h_t. La idea detrás de este estado oculto es el de generar una representación comprimida de la información relevante hasta ese instante t mediante un vector numérico. De esta manera, se construye una red neuronal que tiene "memoria" y puede utilizar información relevante del pasado para determinar estados futuros.

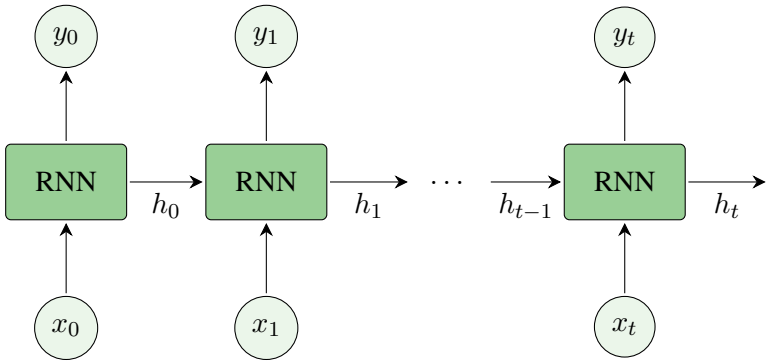

Figura 10: Esquema de red neuronal recurrente usual, en cada instante t se ingresa una entrada x_t devolviendo un valor y_t donde parte de la información vuelve a $t + 1$ mediante el estado oculto h_t.

Internamente, la red ajusta sus parámetros (*pesos y sesgos*) a través de un proceso de entrenamiento guiado por la minimización de una función de pérdida, que mide la discrepancia entre las predicciones y los valores reales. Para ello, las RNN utilizan una adaptación del algoritmo de retropropagación estándar, conocida como retropropagación a través del tiempo (*Backpropagation Through Time*). Sin embargo, al propagar la señal de error hacia atrás a través de secuencias muy largas, los gradientes tienden a disminuir exponencialmente hasta volverse casi nulos. Este fenómeno, conocido como desvanecimiento del gradiente (*vanishing gradient*), limita severamente la capacidad del modelo para aprender dependencias a largo plazo.

Para prevenir el problema del desvanecimiento del gradiente, se han desarrollado arquitecturas más sofisticadas, siendo las redes *Long Short-Term Memory* (LSTM) y las *Gated Recurrent Unit* (GRU) los dos mecanismos más utilizados en la actualidad. Ambas arquitecturas fueron evaluadas en este trabajo, seleccionándose finalmente las unidades GRU como el componente recurrente principal. Esta decisión se fundamentó en el mejor rendimiento de dichas unidades en el problema y una menor cantidad de parámetros.

Las unidades GRU, propuestas por Cho et al. (2014), vienen a solucionar en parte el problema del desvanecimiento del gradiente. La idea central es controlar la información que pasa a las siguientes iteraciones como forma de mitigar el problema de pérdida de memoria. La forma como se decide que información es relevante mantener es a través de dos compuertas:

La *compuerta de actualización* gestiona cuánta información del estado anterior mantener y cuánto se actualiza con la nueva información ingresada:

$$U_t = \sigma(W_{xu} \cdot x_t + W_{hu} \cdot h_{t-1} + b_u)$$

La *compuerta de reinicio* decide cuánta información del estado oculto anterior h_{t-1} debe ser olvidada, permitiendo al modelo eliminar información irrelevante:

$$R_t = \sigma(W_{xr} \cdot x_t + W_{hr} \cdot h_{t-1} + b_r)$$

Donde la función sigmoide σ se encarga de generar valores entre 0 $(olvidar)$ y 1 $(mantener)$, siendo h_{t-1} es el estado oculto anterior, x_t es la entrada en ese instante, mientras que W y b son los pesos y sesgos de las compuertas parámetros a ajustar mediante retropropagación. A través de estas compuertas se genera un candidato a estado oculto $\tilde{h}_t$, creándose el estado h_t basado en el candidato y en el estado anterior h_{t-1}:

$$\tilde{h}_t = \tanh(W_{xh} \cdot x_t + W_{hh} \cdot R_t \odot h_{t-1} + b_h)$$
$$h_t = U_t \odot h_{t-1} + (1 - U_t) \odot \tilde{h}_t$$

Donde la función tangente hiperbólica $tanh$ permite introducir no linealidades y escalar los posibles valores del vector de estados y $\odot$ es el producto Hadamard (multiplicación elemento a elemento).

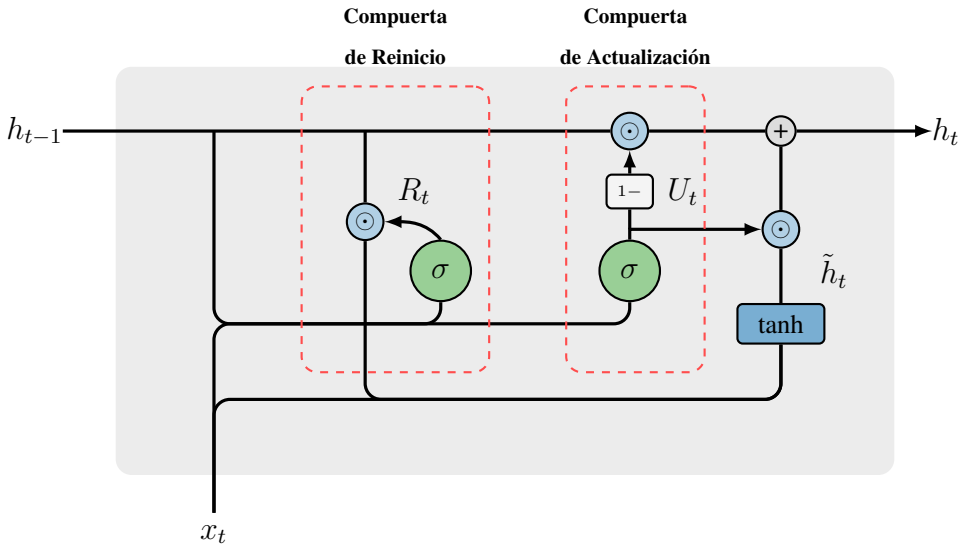

Figura 11: Esquema interno de una unidad **GRU**. La *compuerta de reinicio* R_t controla cuánta información del estado anterior h_{t-1} llega a la $tanh$ para formar el candidato $\tilde{h}_t$; la *compuerta de actualización* U_t decide la interpolación entre h_{t-1} y $\tilde{h}_t$ para producir la salida h_t. Los círculos verdes representan activaciones sigmoides, los $\odot$ celestes multiplicaciones elemento a elemento, el bloque azul la función $tanh$ y el circulo gris la suma final.

3.3.2. Arquitectura Encoder Decoder

Si bien el modelo RNN presentado anteriormente es eficaz para procesar secuencias, su estructura fundamental mapea una entrada en un instante t a una salida en ese mismo instante. La necesidad de generar proyecciones multi-horizonte utilizando únicamente información histórica tanto de la TGF como de sus covariables requiere una arquitectura capaz de mapear una secuencia de entrada a una secuencia de salida de longitud potencialmente distinta. Para esta tarea, una de las estructuras más utilizadas es la de encoder-decoder, también conocida como *sequence-to-sequence* (Seq2Seq).

La idea central de esta arquitectura es separar la red en dos componentes principales:

- *Codificador (Encoder)*: Su función es procesar la ventana de datos históricos completa, que contiene los valores pasados de la TGF y de las covariables y comprimir toda esta información en un único vector numérico de tamaño fijo. Este vector, conocido como vector de contexto o estado latente, debe capturar la esencia de la trayectoria pasada del país.
- *Decodificador (Decoder)*: Su tarea es tomar ese vector de contexto y, a partir de él, generar la secuencia de h predicciones futuras de la TGF.

Para generar la secuencia de salida utilizando RNN, en este caso el decoder opera de manera autorregresiva. Esto significa que produce la predicción un paso a la vez. La predicción para el primer paso futuro (y_{t+1}) se genera a partir del vector de contexto, más un "token" inicial que en problemas de series de tiempo puede ser el valor último conocido de la serie a predecir. Luego, para predecir el segundo paso (y_{t+2}) el decoder utiliza como entrada su propia predicción anterior ($\hat{y}_{t+1}$), y así sucesivamente para todo el horizonte de predicción. Este mecanismo le permite generar secuencias de salida de longitud variable de forma dinámica.

Si bien el proceso autorregresivo de introducir las propias predicciones del modelo en pasos sucesivos se utiliza en la fase de predicción (fuera de la muestra de entrenamiento), durante el entrenamiento presenta un desafío: si el modelo comete un error en las primeras épocas (primeros pasos de entrenamiento), este error se propagará y acumulará en los pasos siguientes, desestabilizando el aprendizaje. Para mitigar esto, se utiliza una técnica llamada *teacher forcing*. En lugar de alimentar al decoder con su propia predicción, en el entrenamiento se le proporciona el valor real del paso anterior. Esto fuerza al modelo a aprender a partir de información correcta en cada paso, acelerando la convergencia. En la práctica, se suelen emplear estrategias mixtas, donde se comienza con una alta probabilidad de usar *teacher forcing* (darle el valor real en lugar del predicho) y se reduce gradualmente a medida que avanza el entrenamiento, para que el modelo se acostumbre a trabajar con sus propias predicciones.

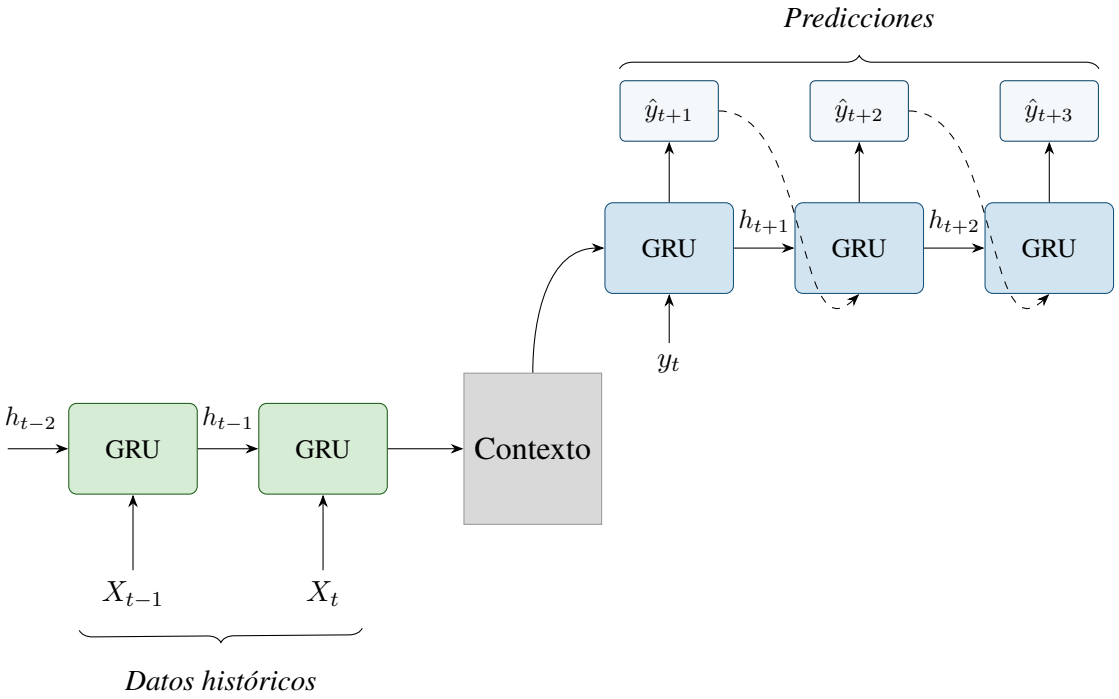

Figura 12: Arquitectura Encoder-Decoder con unidades GRU para la predicción de series de tiempo. El modelo utiliza un encoder para procesar los datos históricos y un decoder autorregresivo para generar las predicciones futuras.

3.3.3. Mecanismo de Atención

La arquitectura encoder-decoder es eficaz para mapear secuencias, pero su rendimiento puede potenciarse si el modelo aprende a asignar una importancia selectiva tanto a las características que componen la entrada en cada instante (covariables) como a los diferentes pasos temporales de la secuencia histórica. La idea central es aprender un conjunto de pesos que ponderan la importancia de diferentes elementos de una entrada para una tarea específica.

En este trabajo, el mecanismo se aplica directamente sobre las covariables históricas de entrada. Para cada instante de tiempo t, la red aprende un vector de pesos de atención α_t mediante una función *softmax*. Este vector pondera la importancia de cada covariable, generando un nuevo vector de características ponderadas $\tilde{X}_t$ que se utiliza como entrada para el codificador.

$$\tilde{\mathbf{X}}_t = \alpha_t \odot \mathbf{X}_t$$

Una ventaja adicional y significativa de los mecanismos de atención es que añaden un grado de interpretabilidad a modelos que, de otro modo, son considerados “cajas negras”. Al inspeccionar los pesos de atención, es posible obtener una idea de a qué partes o covariables de entrada le está prestando más

atención el modelo para realizar sus predicciones en distintos escenarios, ofreciendo una ventana a su funcionamiento interno.

3.3.4. Cuantificación de la Incertidumbre

La arquitectura *encoder-decoder* con unidades GRU, descrita en las secciones anteriores, conforma un modelo completo capaz de generar predicciones puntuales. Entrenado con una función de pérdida estándar, como el Error Cuadrático Medio (MSE), el modelo aprendería a predecir el valor esperado o la media de la distribución futura de la TGF. Sin embargo, para las proyecciones demográficas, una única estimación puntual resulta insuficiente, ya que oculta el rango de escenarios futuros que son plausibles y limita la capacidad de los responsables para diseñar políticas robustas.

Un aspecto fundamental en cualquier modelo de predicción, y de especial relevancia en las proyecciones demográficas, es la capacidad de cuantificar la incertidumbre asociada a las estimaciones. Los modelos de aprendizaje profundo, a pesar de ser deterministas en su formulación estándar, pueden extenderse para generar predicciones probabilísticas. En el ámbito de aprendizaje automático se suelen distinguir dos fuentes principales de incertidumbre que un modelo debe aspirar a capturar:

- *Incertidumbre Aleatoria (Irreducible)*, que representa la variabilidad o el "ruido" inherente al proceso de generación de datos. Esta incertidumbre es una propiedad intrínseca del sistema y no puede eliminarse, ni siquiera con un modelo perfecto o una cantidad infinita de datos.
- *Incertidumbre Epistémica (Reducible)*, que refleja la falta de conocimiento sobre la especificación correcta del modelo o sus parámetros. Esta surge en gran parte a la disponibilidad de datos de entrenamiento limitados y, a diferencia de la aleatoria, puede disminuirse a medida que se recopilan más datos. En el contexto de las redes neuronales, que tienen espacios de parámetros de muy alta dimensionalidad, la incertidumbre epistémica cobra relevancia, ya que pueden existir múltiples configuraciones de pesos que explican los datos de entrenamiento de manera igualmente plausible.

Para abordar estas fuentes de incertidumbre, se han desarrollado diversas técnicas. Enfoques como las Redes Neuronales Bayesianas (BNN) que en lugar de aprender un único valor para los pesos de la red, aprenden una distribución de probabilidad sobre ellos o los ensambles de modelos (ensembles) que consisten en entrenar múltiples modelos y analizar la dispersión de sus predicciones son particularmente efectivos para estimar la incertidumbre epistémica.

Por otro lado, métodos como las redes neuronales probabilísticas buscan capturar el componente aleatorio de los datos. Para esto, la idea central de este tipo de arquitecturas es, en lugar de predecir un valor puntual, tratar de aprender la distribución de probabilidad condicional, $P(y_{t+1:t}|X_{t-i-1:t})$. Dentro de este paradigma, se distinguen dos grandes familias: los métodos paramétricos, como DeepAR, que entrenan a la red para que prediga los parámetros (e.g., media y varianza) de una distribución de

probabilidad predefinida (como la Gaussiana); y los métodos no paramétricos, que no hacen supuestos sobre la forma de la distribución. En este último enfoque, adoptado en este trabajo, el modelo se entrena para predecir directamente los cuantiles de la distribución futura.

La transición de un modelo determinista a uno probabilístico en el contexto del aprendizaje profundo no requiere, en general, una reinvención radical de las arquitecturas de red. Las mismas arquitecturas que son efectivas para extraer características temporales en modelos deterministas, como GRUs o Transformers, también forman la base de los modelos probabilísticos. El cambio fundamental reside en la capa de salida del modelo y en la función de pérdida utilizada durante el entrenamiento. En lugar de tener una capa de salida que produce un único valor escalar, la capa de salida se diseña para producir los parámetros que definen la distribución de probabilidad predictiva.

En el enfoque no paramétrico, la red se diseña para predecir directamente los cuantiles condicionales de la distribución. Para ello, la capa de salida del modelo se modifica para que tenga N neuronas, donde cada una corresponde a un cuantil específico que se desea estimar (e.g. $q_1, q_2, ...q_N$). La función de pérdida para un único cuantil, también conocida como *pinball loss*[3], está dada por la fórmula:

$$\mathcal{L}_q(y, \hat{y}_q) = \begin{cases} q \cdot (y - \hat{y}_q) & \text{si } y > \hat{y}_q \quad \text{(subestimación)} \\ (1 - q) \cdot (\hat{y}_q - y) & \text{si } y \leq \hat{y}_q \quad \text{(sobreestimación)} \end{cases}$$

donde en este caso y es el valor real observado de la TGF, $\hat{y_q}$ es la predicción del modelo para el cuantil q.

Para una predicción multihorizonte de h pasos con N cuantiles, la pérdida total se generaliza como la suma de las pérdidas para cada cuantil en cada horizonte de predicción:

$$\mathcal{L} = \sum_{k=1}^{h} \sum_{q=1}^{N} \mathcal{L}_q(y_{t+k}, \hat{y}_{q,t+k})$$

[3]Es posible demostrar matemáticamente que al minimizar esta función de pérdida, la predicción del modelo convergerá al cuantil condicional teórico deseado (Koenker y Hallock, 2001).

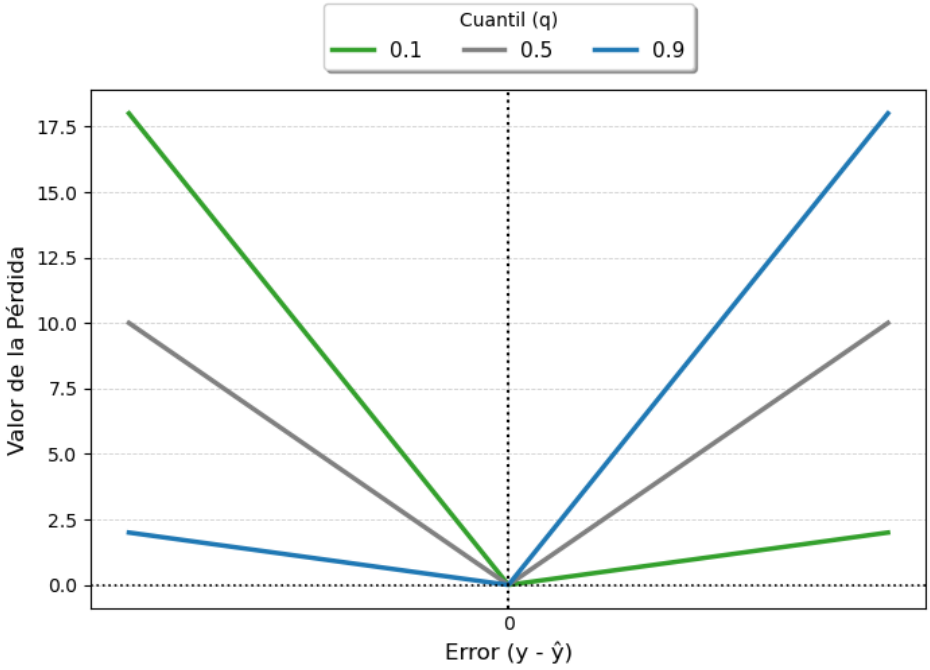

Figura 13: Función de pérdida cuantílica (*pinball loss*). Para el cuantil 0.5 (mediana), la pérdida es simétrica y equivale al Error Absoluto Medio (MAE). En cambio, para cuantiles asimétricos como el 0.9, la función penaliza más la subestimación ($y > \hat{y}$) que la sobreestimación, incentivando al modelo a que sus predicciones sean más altas.

Las estrategias para capturar la incertidumbre no son mutuamente excluyentes; por el contrario, su combinación es una práctica extendida para obtener estimaciones más completas. La literatura recoge, entre otras, combinaciones de ensambles de modelos con predicciones cuantílicas, así como métodos agnósticos al modelo como la predicción conformal. En este trabajo se adopta un esquema de doble vía: (i) se emplea la pérdida cuantílica previamente mencionada; y (ii) se complementa con un *ensamble* de modelos múltiples instancias con arquitectura idéntica entrenadas de forma independiente con el objetivo de robustecer las predicciones y, en conjunto, ofrecer una caracterización más rica de la incertidumbre.

3.3.5. Ensambles Cuantílicos con Redes Recurrentes

La construcción de *ensambles* requiere diversificar sus modelos, lo que usualmente se logra mediante técnicas de remuestreo de datos (como *bagging*) o alterando la configuración y la semilla de aleatorización de cada modelo. Para este trabajo, se implementó la segunda estrategia a través de los siguientes pasos:

1. **Diversificación de Modelos**: La estrategia consiste en configurar un ensamble de 10 modelos. Para asegurar la diversidad, cada modelo recibe una semilla de aleatorización (*seed*) distinta y una configuración de algunos hiperparámetros levemente diferente. Esto permite que cada modelo explore el espacio de soluciones de manera diferente.

2. **Entrenamiento Independiente**: El entrenamiento de cada modelo del ensamble transcurre de forma aislada, utilizando su configuración específica para aprender una representación ligeramente diferente de los datos.

3. **Agregación de Predicciones**: El pronóstico final resulta de promediar las predicciones cuantílicas de todos los modelos. Este paso no solo robustece la estimación puntual (el cuantil 0.50 o mediana), sino que también genera intervalos de predicción más estables y fiables, capturando de manera más efectiva la incertidumbre epistémica del modelo.

Modelos base del ensamble: arquitectura encoder–decoder

Cada uno de los modelos que conforman el *ensamble* están compuestos por un diseño de red neuronal recurrente (RNN) con una arquitectura encoder-decoder, diseñada específicamente para tareas de secuencia a secuencia como la predicción multi-horizonte.

La arquitectura del modelo, ilustrada en la figura 14, procesa la información a través de los siguientes módulos principales:

1. **El Codificador (Encoder):** Este módulo es responsable de procesar la secuencia de datos históricos de entrada (la ventana de i pasos), que contiene las trayectorias pasadas de la TGF y de las covariables socioeconómicas y demográficas. Su objetivo es comprimir toda esta información en un vector de contexto que encapsula la información relevante. El proceso dentro del codificador se desarrolla en los siguientes pasos:

    - **Embedding de Características Categóricas:** La información geográfica, introducida al modelo como una variable categórica (e.g., la subregión de pertenencia de un país), es transformada en una representación vectorial densa (un *embedding*). Esto permite que la red sea capaz de condicionar sus predicciones y capturar patrones específicos de la fecundidad para por ejemplo cada región dentro de una misma arquitectura unificada.
    - **Atención sobre las Características de Entrada:** Antes de ser procesadas por las unidades recurrentes, las covariables de entrada atraviesan un mecanismo de atención. Este módulo consiste en una pequeña red neuronal *feed-forward* que, para cada instante de tiempo, calcula un conjunto de pesos para cada covariable. Estos pesos se multiplican elemento a elemento por el vector de características original.
    - **Capas Recurrentes GRU:** La secuencia de entrada, ya enriquecida con los *embeddings* y ponderada por la atención, es procesada por una capa de unidades GRU. La función de estas capas es capturar las dependencias temporales y generar el estado oculto final, que servirá como el vector de contexto para el decodificador.

2. **El Decodificador (Decoder):** Este módulo utiliza el vector de contexto para generar la secuencia de h predicciones futuras de la TGF.

    - **Generación Autorregresiva:** El decodificador produce las predicciones de manera secuencial. Para el primer paso futuro, utiliza el contexto y el último valor conocido de la TGF. Para los pasos subsiguientes, se alimenta de su propia predicción anterior, en un proceso autorregresivo. Durante el entrenamiento, se emplea la técnica de ***teacher forcing***, suministrando el valor real observado con una probabilidad decreciente para estabilizar el aprendizaje.

3. **Capa de Salida:** La salida de la unidad GRU del decodificador en cada paso de tiempo se proyecta a través de una capa lineal final. De acuerdo con la estrategia de regresión cuantílica adoptada, esta capa no produce una única estimación puntual, sino un vector de N valores. Cada uno de estos valores corresponde a un cuantil específico (e.g., 0.05, 0.50, 0.95) de la distribución predictiva, permitiendo así la construcción de intervalos de predicción.

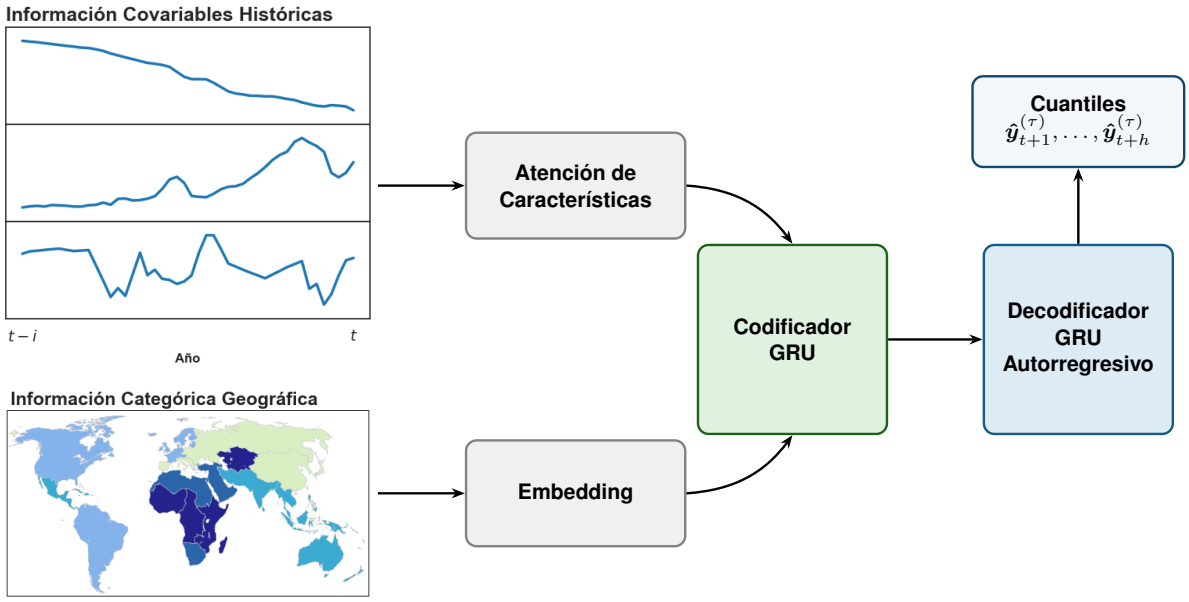

Figura 14: Esquema de la arquitectura del modelo Encoder-Decoder implementado. El modelo procesa las covariables históricas a través de un mecanismo de atención y la información geográfica mediante una capa de embedding. Ambas representaciones son integradas por el codificador GRU, cuyo vector de contexto es utilizado por el decodificador autorregresivo para generar las predicciones cuantílicas de la TGF.

3.3.6. Técnicas de Regularización

Adicional a las capas mostradas en la arquitectura, el modelo incorpora diversas técnicas de regularización para mitigar el riesgo de sobreajuste. Una de las estrategias más utilizadas en el entrenamiento de redes neuronales es el *Dropout*, que consiste en "apagar" o poner a cero, de forma aleatoria durante cada pasada de entrenamiento, un porcentaje de las neuronas de una capa. Esta técnica impide que la red dependa excesivamente de neuronas o características específicas, forzándola a aprender representaciones más distribuidas y robustas. En este trabajo, se empleó una variante conocida como *Feature Dropout*, que en lugar de apagar neuronas individuales, desactiva aleatoriamente características de entrada completas a lo largo de toda la secuencia temporal para una muestra de entrenamiento dada.

Como técnica complementaria, se añadió una pequeña cantidad de ruido blanco gaussiano a las características de entrada durante el entrenamiento. Este método actúa como una forma de aumento de datos (*data augmentation*), creando nuevas versiones ligeramente perturbadas de los datos de entrenamiento en cada época. Esta práctica obliga al modelo a ser menos sensible a variaciones irrelevantes y a aprender mapeos más suaves y estables, mejorando así su capacidad de generalización.

3.4. Metodología de Entrenamiento

En el desarrollo de cualquier modelo de aprendizaje automático, y en particular en las redes neuronales, el tratamiento de los datos y la configuración del entrenamiento son una de las etapas de mayor importancia, que a menudo determinan el éxito del modelo tanto o más que la propia complejidad de la arquitectura utilizada. En esta sección se detalla todo el proceso realizado para el entrenamiento del modelo de la TGF.

3.4.1. Fuente de Datos y Covariables Seleccionadas

El conjunto de datos utilizado para este trabajo se construyó a partir de tres fuentes principales: la División de Población de las Naciones Unidas, el *Wittgenstein Centre for Demography and Global Human Capital*, y la base de datos del Banco Mundial[4]. Para la construcción del modelo global, se seleccionaron 193 países, excluyendo principalmente microestados y territorios con poblaciones muy reducidas o con una cobertura insuficiente de las covariables analizadas. El periodo de estudio en este caso será desde 1950 a 2021, que luego se dividirá para testeo y validación del modelo. Esto resulta en un conjunto de datos total de aproximadamente 15.000 observaciones (puntos de datos país-año).

La variable objetivo del modelo, la Tasa Global de Fecundidad (TGF), se obtuvo de las estimaciones históricas publicadas por la División de Población de las Naciones Unidas. Se eligió esta fuente por ser una de las más completas y estandarizadas, construida a partir de una síntesis de registros vitales, censos y encuestas como la *Demographic and Health Surveys* (DHS). Sin embargo, es crucial señalar que estos datos no son puramente empíricos, sino que son el resultado de interpolaciones y ajustes construidos sobre supuestos conceptualmente similares a los que utiliza el modelo utilizado por Naciones Unidas para las proyecciones. Como se analizará en la sección de resultados, esta característica podría

[4]Portal de datos: División de Población de las Naciones Unidas, Wittgenstein Centre, Banco Mundial.

generar una evaluación sobreoptimista del rendimiento dicho modelo, un factor clave a considerar al interpretar la comparación.

A esta serie principal se le sumó un conjunto de covariables socioeconómicas y demográficas. La elección del conjunto de covariables finales se basó en probar un conjunto de variables con amplia discusión bibliográfica en el mundo de la demografía que tienen un impacto en la dinámica de la fecundidad como las que se mencionó en la sección de transición de la fecundidad, para luego realizar un filtro final seleccionando el subconjunto con mejor desempeño en base a la métrica de validación.

Covariable	**Descripción**	**Fuente**	**Cobertura**	**Observaciones**
Educación Femenina				
Años Promedio de Escolaridad Femenina	Estimación promedio de años de estudio para la población femenina entre 20 y 39 años.	Wittgenstein Centre	1950-2020	Datos quinquenales, que se anualizan mediante un *spline* lineal.
Planificación Familiar				
Prevalencia de Métodos Anticonceptivos	Porcentaje de mujeres en edad reproductiva (15-49 años) que utilizan un método anticonceptivo moderno.	División Población (NNUU)	1970-2021	
Necesidades Insatisfechas de Planificación	Proporción de mujeres en edad fértil que no usan anticonceptivos pero desean posponer o limitar la procreación.	División Población (NNUU)	1970-2021	
Edad Media a la Maternidad	Edad promedio de las madres en el momento del nacimiento de sus hijos.	División Población (NNUU)	1950-2021	
Indicadores Económicos y de Desarrollo				
Tasa de Mortalidad Infantil (<5 años)	Indicador de la cantidad de niños que fallecen antes de los cinco años por cada 1,000 nacidos vivos.	División Población (NNUU)	1950-2021	
Proporción de Población Urbana	Porcentaje de la población total que reside en áreas clasificadas como “urbanas”.	Banco Mundial	1960-2021	
PBI per cápita	Producto Bruto Interno per cápita. (US$ a precios actuales)	Banco Mundial	1960-2021	Se suaviza la serie aplicando un promedio móvil de 3 períodos.

Cuadro 1: Covariables dinámicas históricas finales utilizadas en el modelo de predicción de la Tasa Global de Fecundidad.

3.4.2. Preprocesamiento de Datos

El preprocesamiento de los datos es un paso fundamental en cualquier modelo de aprendizaje profundo. Antes de ser utilizadas para el entrenamiento, las series temporales tanto de la TGF como las covariables seleccionadas fueron sometidas a las siguientes transformaciones:

- **Transformación Logarítmica**: Se aplicó una transformación logarítmica a la variable objetivo (TGF). Esta operación permite que el modelo trabaje con la variable en un rango de valores más acotado, lo que empíricamente resultó en una mejora en el rendimiento de las métricas de evaluación y en la convergencia del entrenamiento.

- **Estandarización**: Todas las variables dinámicas (tanto la TGF transformada como las covariables) fueron estandarizadas. Este proceso es obligatorio para el correcto funcionamiento de los algoritmos de optimización basados en gradientes, pues asegura que ninguna variable domine el proceso de aprendizaje debido a su escala. En el contexto de modelos globales de series temporales, el escalado de la variable objetivo (en este caso, la TGF) presenta ciertas distinciones donde se abren distintas alternativas. Por un lado, se puede optar por un escalado individual, donde cada serie se normaliza según sus propios parámetros; este mecanismo es útil en escenarios donde la magnitud de las series a modelar es muy dispar, pero tiene la desventaja de que se pierde la información de las escalas relativas entre series. Por otro lado, se puede realizar un escalado global con un único conjunto de parámetros para todas las series, estrategia que es útil siempre que las magnitudes de las series no difieran en exceso. Por simplicidad, y dado que las evaluaciones preliminares no indicaron una necesidad imperiosa de un escalado individual que complejizaría el modelo, se optó por el escalado global para la TGF. Este enfoque asegura que los valores escalados sean directamente comparables entre países, permitiendo que el modelo interprete la magnitud de la fecundidad de manera consistente. En cuanto a las covariables históricas, la práctica usual es la de utilizar el escalado global, ya que permite la comparación directa de sus magnitudes relativas. Para cada observación x de una variable, el valor estandarizado z se calcula como:

  $$z = \frac{x - \mu}{\sigma}$$

  donde μ y σ son la media y la desviación estándar globales de la variable, dichos valores deben ser obtenidos solamente con los datos con los que se va a entrenar el modelo de forma de evitar "fuga de datos" también conocido como *data leakage*, una vez obtenidos todos los conjuntos posteriores tanto de pruebas o predicciones futuras deben utilizar estos mismos parámetros.

- **Imputación de Datos Faltantes**: Para el manejo de valores ausentes, se adoptó una estrategia doble. En los casos en que una covariable completa era inexistente para un país, se realizó una imputación utilizando la media de la subregión (basada en las definiciones de NNUU) a la que pertenece dicho país. Esta estrategia es consistente con la idea que los procesos de las distintas covariables analizadas no deberían diferir en exceso entre un país y sus países vecinos. Para las covariables que presentaban datos faltantes al inicio del período de estudio (e.g., indicadores

con cobertura desde 1960), se aplicó una técnica de relleno en la cual se introducen ceros en dichos valores faltantes. Si bien esta imputación introduce una aproximación, su impacto en las redes recurrentes es limitado, ya que la pérdida de información es más crítica en los pasos temporales cercanos al inicio de la predicción (la transición del encoder al decoder) que en los pasos iniciales de la secuencia de entrada, donde esta información imputada se va diluyendo entre los estados ocultos.

Para enriquecer el modelo y permitirle discernir entre datos reales e imputados, se incorporó una característica de imputación dinámica. Esta consiste en ingresar un conjunto de datos binarios (*one-hot*) que acompaña a los datos de entrada, indicando explícitamente a la red si un valor en un paso de tiempo específico es original o si ha sido imputado. Al proporcionar esta metainformación directamente al modelo, se le dota de la capacidad de aprender a ponderar la fiabilidad de cada punto de dato, ajustando la influencia de realizar esta imputación.

3.4.3. Estructura de Entrenamiento y Optimización

En los modelos de aprendizaje profundo, que poseen una alta capacidad para ajustarse a los datos, medir el error de predicción sobre el mismo conjunto de datos utilizado para el entrenamiento puede conducir a una evaluación excesivamente optimista y poco fiable del verdadero rendimiento del modelo. Esto se debe a que la red puede memorizar las particularidades y el ruido de los datos de entrenamiento, en lugar de aprender los patrones subyacentes que permiten generalizar a nuevos datos.

Para obtener una estimación insesgada de la capacidad de generalización del modelo, es imprescindible evaluarlo sobre un conjunto de datos que no haya sido utilizado durante ninguna etapa del entrenamiento o de la selección de hiperparámetros. Dado que se trabaja con series temporales, la división de los datos debe respetar estrictamente el orden cronológico para no introducir información del futuro en el entrenamiento. Por este motivo, se realizó una división temporal del conjunto de datos como se ilustra en la Figura 15. Los últimos 15 años de datos disponibles (del período 2007 a 2021) para los 183 países fueron reservados como un conjunto de prueba final (*test set*). El resto de los datos, que abarcan desde 1950 hasta 2006, se utilizaron para el proceso de entrenamiento y validación.

A partir de dicho conjunto de entrenamiento se obtienen los parámetros de estandarización y se generan las ventanas deslizantes utilizadas para el entrenamiento de la red a partir de los datos ya preprocesados. Para los anchos de las ventanas se utiliza una estrategia en la cual se define un largo mínimo y máximo para las ventanas de entradas construidas, de esta forma se generan casos de entrenamientos con variaciones en la cantidad de historia que observan obligando al modelo a no depender siempre de una cantidad fija de historia que funciona también como una forma de *data augmentation*. Aunque el largo sea variable el modelo por construcción toma el tamaño máximo donde las ventanas de largo más pequeños se rellenan con ceros en las primeras iteraciones hasta llegar al tamaño máximo. Para ser computacionalmente eficientes a su vez las librerías de aprendizaje profundo tienen herramientas para que no se procesen esos pasos de relleno.

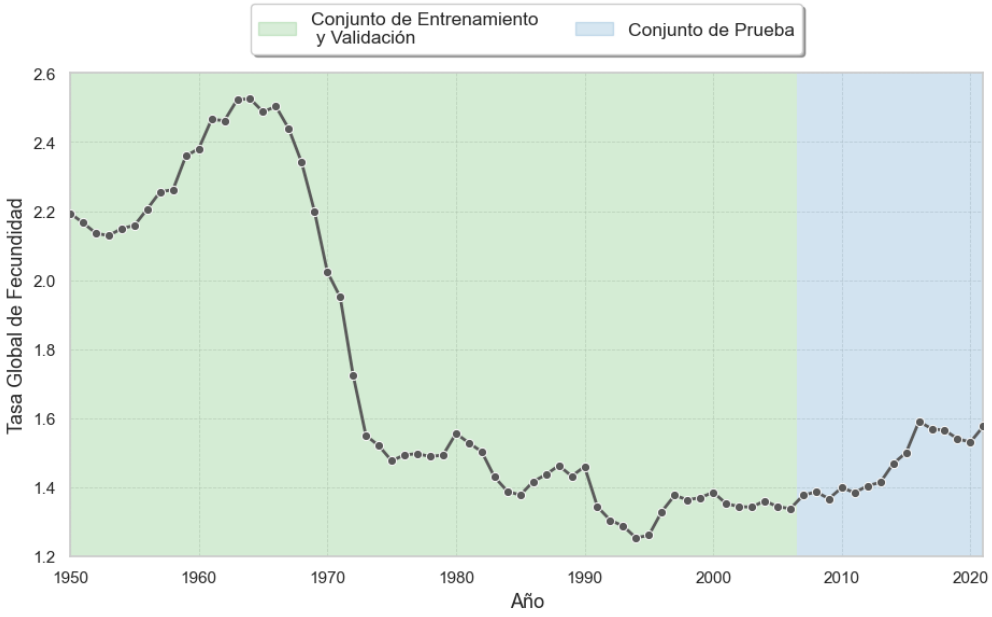

Figura 15: Ilustración de la estrategia de división temporal aplicada de forma consistente a las series de TGF de todos los países. Los últimos 15 años de datos (2007-2021) se reservan como un conjunto de prueba no observado durante el entrenamiento. El conjunto de entrenamiento y validación comprende todo el período histórico previo (1950-2006).

De esta manera se obtienen un conjunto de muestras de entrenamiento que están compuestas por ventanas de datos de entrada con la idea anteriormente descrita y un correspondiente conjunto de ventanas etiqueta que se utilizarán para computar la pérdida. Estas ventanas se agrupan en mini lotes llamados "*batch*" de entrenamiento, estos lotes de forma usual se generan de forma mediante un muestreo aleatorio sin reemplazo, donde en cada iteración el modelo va a ver mezclado ventanas con casos de entrenamiento de distintos países, una vez pasados por el modelo todos las ventanas disponibles (1 época) se vuelve a generar nuevamente un nuevo conjunto de mini lote con muestras aleatorias hasta nuevamente completar todas las ventanas. El entrenamiento del modelo al igual que otros algoritmos de aprendizaje profundo se guía por la minimización de una función de pérdida que cuantifica la discrepancia entre las predicciones y los valores reales, en este caso mediante la predicción de cuantiles y la correspondiente función de pérdida multicuantílica.

El proceso de entrenamiento se puede describir en los pasos siguientes:

1. **Preprocesamiento**. Se realiza todo el preprocesado y transformaciones requeridas en las series históricas de TGF y las covariables.
2. **Construcción de ventanas**. De cada serie se extraen ventanas de entrenamiento formadas por:
    - La secuencia de información histórica de covariables y la TGF.
    - Un identificador categórico (región, país, subregión).

- La *ventana-etiqueta* con los valores futuros de la TGF usados para calcular la pérdida.

3. **Mini lotes**. Las ventanas se agrupan aleatoriamente en mini lotes de tamaño fijo donde se realiza el siguiente proceso iterativo:

   a) **Entrada al codificador**. Cada mini-lote contiene:

      a) Covariables históricas que pasan por el mecanismo de atención por características;
      b) Identificadores geográficos que se convierten en vectores vía *embeddings*.

      Ambos conjuntos se concatenan y forman lo que será la entrada en el codificador.

   b) **Generación del contexto**. El codificador (GRU) procesa la secuencia y produce un vector contexto que resume la información histórica relevante.

   c) **Predicción**. De forma secuencial se generan las predicciones para los 15 pasos del horizonte, emitiendo en cada uno los cinco cuantiles $(0.05, 0.10, 0.50, 0.80, 0.95)$. En cada iteración se decide aleatoriamente si el siguiente paso se alimenta con la mediana pronosticada o con el valor real de la TGF. En la que la probabilidad de *teacher forcing* se va disminuyendo de forma progresiva al ir pasando las iteraciones.

   d) **Cálculo de la pérdida y actualización**. Se calcula la pérdida cuantílica total de ese mini-lote y el optimizador actualiza los parámetros mediante retropropagación.

Este proceso de ajuste se realiza de forma iterativa a lo largo de múltiples épocas, donde una época consiste en una pasada completa del modelo sobre todo el conjunto de ventanas de entrenamiento.

El mecanismo de optimización que permite el ajuste de los parámetros del modelo se basa en el descenso de gradiente estocástico (SGD). El SGD calcula una estimación del gradiente utilizando el subconjunto de mini lotes antes nombrados en lugar de todo el conjunto de datos, lo que lo hace computacionalmente más eficiente. Si bien cada paso es una aproximación "ruidosa", la trayectoria promedio converge eficientemente hacia un mínimo de la función de pérdida. Para este trabajo, se utilizó *Adam (Adaptive Moment Estimation)*, un optimizador que refina la idea del SGD y que se ha convertido en una solución estándar y muy flexible para una amplia gama de problemas de aprendizaje profundo.

Adicionalmente, el comportamiento del optimizador Adam se complementa mediante dos técnicas clave. La primera es el decaimiento de la tasa de aprendizaje. La idea es comenzar el entrenamiento con una tasa de aprendizaje relativamente alta para explorar rápidamente la superficie de la función de pérdida y, a medida que el entrenamiento avanza y el modelo se acerca a una solución óptima, reducir gradualmente esta tasa. Esto permite dar pasos más pequeños y precisos en las etapas finales, evitando "saltar" por encima del mínimo. En este trabajo, se implementó un decaimiento por pasos, que reduce la tasa de aprendizaje en un factor fijo cada cierto número de épocas. La segunda técnica es el decaimiento de los pesos (*weight decay*), una forma de regularización L2. Este método añade un término de penalización a la función de pérdida que es proporcional al cuadrado de la magnitud de los pesos del modelo. Al minimizar la pérdida, el optimizador también se ve incentivado a mantener los pesos pequeños, lo que desincentiva al modelo de aprender relaciones excesivamente complejas y reduce el riesgo de sobreajuste.

Una estrategia adicional que es usualmente utilizada que se aplica en este caso para evitar el sobreajuste y optimizar el tiempo de cómputo, es la parada temprana (*Early Stopping*). Este método consiste en monitorear una métrica de rendimiento, como el error o pérdida en el conjunto de validación, al final de una cierta cantidad de épocas de entrenamiento. El entrenamiento se detiene de forma automática si la métrica de validación no muestra una mejora significativa durante un número predefinido de épocas consecutivas, conocido como "paciencia" (*patience*). Al detener el proceso justo cuando el modelo deja de generalizar, se evita que este comience a memorizar el ruido del conjunto de entrenamiento, capturando así el modelo con el mejor rendimiento en datos no observados.

3.5. Selección de Hiperparámetros y Estrategia de Validación

El rendimiento de una red neuronal es altamente sensible a la elección de sus hiperparámetros: aquellos parámetros de configuración que no se aprenden durante el entrenamiento, sino que definen la arquitectura y el proceso de aprendizaje. La gran cantidad de hiperparámetros en un modelo de este tipo genera un espacio de búsqueda de configuraciones muy extenso sumado a el tiempo requerido en cada entrenamiento, haciendo que una búsqueda manual o exhaustiva (como una búsqueda en grilla o *grid search*) sea computacionalmente inviable.

Por este motivo, se recurrió a métodos de búsqueda inteligente de hiperparámetros. En este trabajo, se utilizó el *framework Optuna*, que emplea técnicas de optimización bayesiana para explorar el espacio de búsqueda de manera eficiente. A diferencia de una búsqueda aleatoria, *Optuna* construye un modelo probabilístico de la función objetivo (usualmente el error de validación en función de los hiperparámetros) y utiliza esta información para decidir qué combinación de hiperparámetros probar a continuación, equilibrando la exploración de nuevas regiones del espacio con la explotación de aquellas que ya han demostrado un buen rendimiento (Akiba et al., 2019).

Adicional a un buen sistema de búsqueda de hiperparámetros se debe obtener una estimación robusta del rendimiento de cada conjunto evaluado, con lo que al uso de *Optuna* se lo combina con técnicas de validación cruzada. Esta técnica esencial en el aprendizaje automático requiere de ciertas distinciones al tratarse de un problema con estructural temporal, donde existen distintas estrategias todas ellas teniendo en cuenta el comportamiento temporal. En este caso se implementó una estrategia de validación cruzada con origen móvil (*rolling forecast origin*). Este procedimiento consiste en entrenar y evaluar el modelo múltiples veces sobre ventanas de datos expansivas. Se comienza con un conjunto de entrenamiento inicial (en este caso un año), se entrena el modelo y se evalúa su rendimiento en el período de validación subsiguiente (los siguientes 15 años). Posteriormente, el conjunto de entrenamiento se expande incorporando varios años de datos, y el proceso se repite. Esta técnica, aunque computacionalmente más costosa, permite evaluar la estabilidad del rendimiento del modelo a lo largo de diferentes períodos históricos, proporcionando una estimación mucho más fiable de su capacidad de generalización.

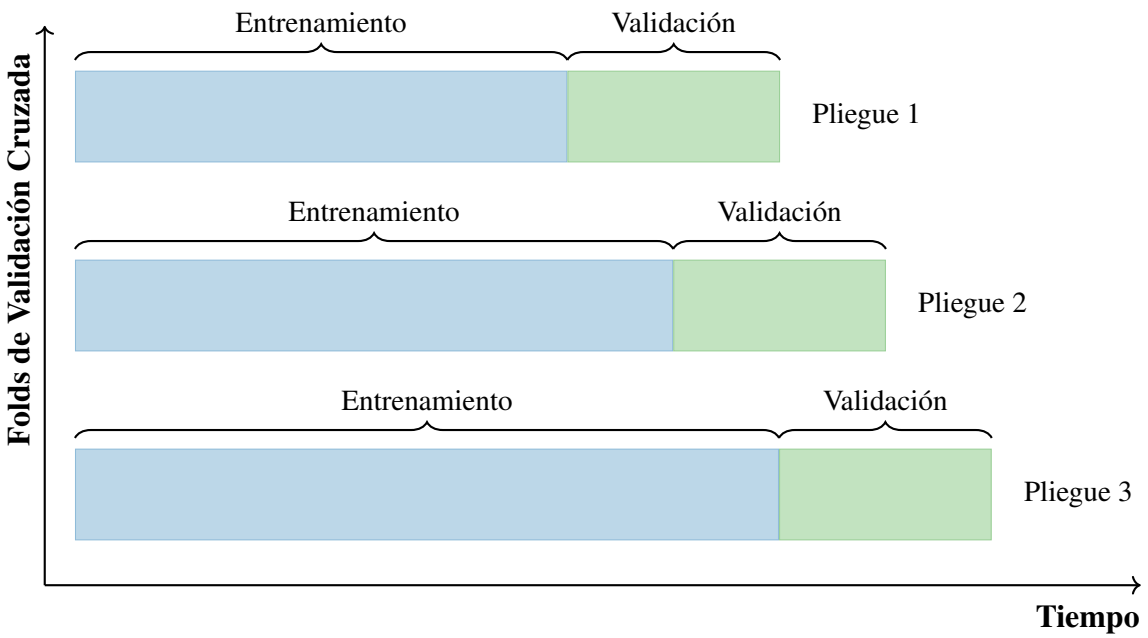

Figura 16: Ilustración de la estrategia de validación cruzada con origen móvil (*rolling forecast origin*). En cada pliegue (*fold*), el modelo se entrena con un conjunto de datos históricos que se expande progresivamente, y se valida en el período de tiempo inmediatamente posterior. .

En redes neuronales recurrentes los hiperparámetros mas sensibles que afectan directamente el rendimiento del modelo suelen ser:

- **Tasa de Aprendizaje**: La tasa de aprendizaje (*learning rate*) controla la magnitud de cada actualización de los pesos; un valor inadecuado suele hacer que el entrenamiento diverja o sea extremadamente lento.
- **Tamaño del Estado Oculto**: Define el largo del vector de estado oculto lo que controla la capacidad de "memoria" de la red. Un valor muy grande ocasiona un aumento en los parámetros y por tanto en un aumento del riesgo de sobreajuste, un valor muy bajo puede limitar la capacidad de la red en captar patrones subajustarse.
- **Longitud de la secuencia/ventana de entrada**: La cantidad de secuencias o pasos temporales que entran al encoder es un hiperparámetro clave, un valor muy corto puede hacer que no se capte correctamente las dependencias de la estructura de los datos, mientras que un valor muy largo corre riesgo de presentar problemas de desvanecimiento de gradiente haciendo que la información mas antigua de la secuencia no tenga ningún aporte en las predicciones.
- **Número de capas**: Dicho valor controla la profundidad de la red, añadiendo más capas de unidades en este caso GRU que capturan patrones de manera jerárquica. En problemas de series de tiempo suele bastar con utilizar una sola capa y el aumento del mismo corre riesgo de sobreajuste.

- **Tamaño del *Batch***: El tamaño del mini lote utilizado suele ser un hiperparámetro clave que afecta la velocidad como la estabilidad del aprendizaje. El utilizar lotes de datos más pequeños y no optimizar con todos los datos de entrenamiento no es solo una ventaja desde el punto de vista de la eficiencia computacional sino que permite introducir un pequeño ruido que ayuda a que el modelo no se sobreajuste.

Los hiperparámetros relacionados a la arquitectura como lo son el tamaño del estado oculto o el número de capas se debe definir tanto para el encoder como el decoder, en este caso se tomó el mismo tamaño oculto para ambas partes de la red, un método que se utiliza usualmente para simplificar.

La búsqueda de hiperparámetros utilizada se ejemplifica en los siguientes pasos:

1. **Definición del Espacio de Búsqueda**. Para cada hiperparámetro clave (tasa de aprendizaje, tamaño del estado oculto, probabilidad de dropout, etc.), se define un rango de valores o una lista de opciones posibles.
2. **Proceso Iterativo de *Optuna* (*Trial*)**. Optuna ejecuta un número predefinido de "pruebas" o *trials*. Donde en cada una se realizan los siguientes pasos:
    a) *Muestreo de Hiperparámetros*: Optuna sugiere una nueva combinación de hiperparámetros a partir del espacio de búsqueda definido.
    b) *Evaluación con Validación Cruzada*: Para evaluar esta combinación, se ejecuta un procedimiento de validación cruzada con origen móvil, como fue mencionado previamente. Por ejemplo, se entrena con datos hasta el año 1985 y se valida en 1986-1995; luego se entrena con datos hasta 1990 y se valida en 1991-2005, y así sucesivamente.
    c) *Obtención del Error*: De cada pliegue (*fold*) de la validación cruzada se obtiene una métrica de error (en este trabajo, se utilizó el Error Porcentual Absoluto Medio (MAPE) como métrica de evaluación). El resultado final para el *trial* actual que evalúa el rendimiento de esa combinación de hiperparámetros es el promedio de esta métrica en todos los pliegues.
3. **Optimización**. *Optuna* utiliza el error promedio del *trial* para actualizar su modelo probabilístico interno. Este modelo le permite decidir qué combinación de hiperparámetros probar a continuación, balanceando la exploración de nuevas regiones del espacio de búsqueda con la explotación de aquellas que ya han demostrado un buen rendimiento.

Una de las ventajas destacables de este tipo de *framework* como se observa en los pasos es la adaptabilidad a cualquier tipo de modelo. Donde el optimizador de *Optuna* solamente requiere de darle un rango de valores posibles para cada hiperparámetro y en cada intento darle un valor numérico que se le indica si debe minimizar o maximizar, con lo que no se debe de cambiar en absoluto el funcionamiento o la arquitectura utilizada del modelo.

4. Análisis de Resultados

Una vez finalizado el proceso de búsqueda y validación de hiperparámetros, se obtiene una configuración del modelo considerada como final. El siguiente paso consiste en evaluar el rendimiento predictivo final de esta arquitectura. Para ello, se realiza una evaluación única y definitiva sobre el conjunto de prueba, que comprende el período de 2007 a 2021 y que, como se estableció previamente, fue estrictamente reservado y no se utilizó en ninguna fase del entrenamiento o la optimización. Esta evaluación final permite obtener una estimación insesgada de la capacidad del modelo para generalizar y pronosticar trayectorias de fecundidad no observadas.

4.1. Comparación del Rendimiento Predictivo

En el desarrollo de modelos de aprendizaje automático, especialmente con arquitecturas complejas como las redes neuronales, es fundamental evaluar su rendimiento en un contexto comparativo. Para ello, se establece un *modelo base*. Este modelo suele ser más simple, computacionalmente menos costoso y, a menudo, representa un estándar tradicional en el campo de estudio.

El propósito del modelo base tiene dos utilidades cruciales en cualquier problema de aprendizaje automático:

- **Justificar la complejidad**: Sirve como un "control de calidad" para determinar si el esfuerzo y la complejidad de un modelo avanzado (como la red *GRU* propuesta) se traducen en una mejora tangible en el rendimiento. Si un modelo complejo no puede superar a uno simple, su utilidad queda en entredicho.
- **Establecer un piso de rendimiento**: Proporciona un umbral mínimo de calidad. Cualquier modelo propuesto debe, como mínimo, igualar el desempeño del modelo base para ser considerado viable.

En el ámbito de las series temporales, existen diversos modelos candidatos que pueden cumplir esta función, desde pronosticadores ingenuos hasta modelos estadísticos tradicionales como los *SARIMA* o las distintas variantes de *suavizamiento exponencial*, caracterizados por su simplicidad interpretativa y su reducido número de parámetros.

Para este trabajo se optó por utilizar el método de suavizamiento exponencial amortiguado de Holt (*Holt's Damped*). La elección de este modelo se fundamenta en su capacidad para capturar la dinámica de la TGF, modelando tanto la tendencia que esta presenta durante la fase de transición, como el posterior desaceleramiento de dicha tendencia cuando la tasa se aproxima a valores bajos, gracias a su parámetro de amortiguación.

Conceptualmente, el método puede ser visto como un promedio móvil donde las observaciones pasadas reciben un peso que decrece exponencialmente. El modelo descompone la serie en un componente

de nivel (ℓ_t) y uno de tendencia (b_t), que se actualizan en cada período de tiempo t mediante las siguientes ecuaciones de suavizado:

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \quad \textbf{(Ecuación de Nivel)}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \quad \textbf{(Ecuación de Tendencia)}$$

Donde y_t es el valor observado de la serie, mientras que α y β son los parámetros de suavizado para el nivel y la tendencia, respectivamente. A partir de la última estimación del nivel y la tendencia, el pronóstico para h pasos hacia adelante se obtiene con la siguiente ecuación:

$$\hat{y}_{t+h|t} = \ell_t + \sum_{i=1}^{h} \phi^i b_t$$

Para la evaluación, se ajusta un modelo independiente serie por serie, es decir, para cada país. En el que los parámetros de suavizamiento (α, β) y el parámetro de amortiguación (ϕ) se estiman mediante máxima verosimilitud.

La comparación del poder predictivo del modelo propuesto, el suavizamiento exponencial y del modelo bayesiano requiere el uso de métricas de error estandarizadas. Dado que se está evaluando un modelo global sobre un conjunto de países con distintas trayectorias y niveles de fecundidad, es fundamental seleccionar métricas relativas que no dependan de la escala de cada serie. Esto asegura que los errores de predicción de un país con una TGF de 4.0 hijos por mujer sean comparables a los de un país con una TGF de 1.5 hijos por mujer. Por este motivo, se seleccionaron las siguientes métricas:

- **MAPE (Error Porcentual Absoluto Medio)**: Esta métrica expresa el error como un porcentaje promedio del valor real observado, siendo una medida relativa fácilmente interpretable. Su fórmula es:

  $$\text{MAPE} = \frac{1}{h} \sum_{i=t+1}^{t+h} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\,\%$$

  donde y_i es el valor real de la TGF y $\hat{y}_i$ el valor pronosticado. Para evaluar las predicciones cuantílicas, se utiliza la mediana (cuantil 0.5) como la estimación puntual.

- **RMSSE (Raíz del Error Cuadrático Medio Escalado)**: Es otra medida de error independiente de la escala. A diferencia del MAPE, el RMSSE escala el error del pronóstico en función de la variabilidad de la propia serie, utilizando como referencia el error de un pronóstico ingenuo sobre los datos de entrenamiento. Su fórmula es la siguiente:

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{h}\sum_{i=t+1}^{t+h}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=2}^{t}(y_i - y_{i-1})^2}}$$

donde nuevamente, y_i es el valor real y $\hat{y}_i$ es la mediana pronosticada para cada paso del horizonte de predicción.

La evaluación del rendimiento se realiza mediante ambas métricas para obtener una visión completa y equilibrada. Mientras que el MAPE ofrece una ventaja clara en interpretabilidad al expresar el error en términos porcentuales, presenta conocidas limitaciones estadísticas como su asimetría en la penalización de errores. El RMSSE, en cambio, proporciona una medida más robusta que, al ser cuadrática, penaliza más los errores grandes. A su vez permite comparar de mejor forma el rendimiento de los modelos entre trayectorias de fecundidad con mayor o menor variabilidad.

Para realizar una evaluación comparativa del rendimiento, se definieron cuatro modelos distintos. Donde todos los modelos fueron entrenados y evaluados utilizando exactamente los mismos datos y misma partición: el período de 1950 a 2006 para el entrenamiento y el de 2007 a 2021 como conjunto de prueba final. Esto asegura una comparación justa y directa de su capacidad predictiva. Los modelos evaluados son:

1. **Suavizamiento Exponencial (Modelo Base)**: El método de *Holt's Damped*, aplicado de forma independiente a cada país, que sirve como punto de referencia.
2. **GRU S/C**: El modelo de redes neuronales recurrentes con arquitectura encoder-decoder propuesto, pero entrenado únicamente con los datos históricos de la TGF e información geográfica, sin el uso de covariables dinámicas externas, ni mecanismos de atención.
3. **GRU**: La versión completa del modelo propuesto, que incluye las covariables socioeconómicas y el mecanismo de atención sobre las características de entrada.
4. **BayesTFR**: El modelo bayesiano jerárquico utilizado por las Naciones Unidas.

Antes de analizar los resultados, es fundamental reiterar una aclaración sobre la naturaleza de los datos, ya adelantada previamente. Si bien se basan en registros empíricos, los datos de la TGF utilizados en esta comparación son en sí mismos interpolaciones y estimaciones construidas sobre supuestos conceptualmente similares a los que utiliza el modelo BayesTFR. Esto implica que el error de BayesTFR, especialmente en países con menor calidad de datos empíricos, podría ser sobreoptimista. Por lo tanto, su inclusión en este análisis es principalmente comparativa, con el fin de contextualizar el rendimiento del modelo propuesto frente a las alternativas ya establecidas

Los resultados que se presentan a continuación muestran que ambas modalidades del modelo propuesto (con y sin covariables) superan consistentemente al modelo base de Suavizamiento Exponencial, con una mejora media en el MAPE de aproximadamente un 23 % entre el base y el modelo GRU

completo. Asimismo, se observa que la inclusión de covariables mejora en cierta medida las métricas en el conjunto de prueba, confirmando su relevancia para las proyecciones. Finalmente, el hallazgo más destacado es que el modelo GRU logra un nivel de error, tanto en MAPE como en RMSSE, que es directamente comparable al del estándar BayesTFR. Este resultado demuestra que el enfoque de aprendizaje profundo es una alternativa competitiva y viable para la proyección de la TGF a nivel global en un horizonte de 15 años como el planteado en esta prueba.

A continuación, en la Figura 17 y el Cuadro 2 se presenta la distribución detallada y los promedios de los errores de predicción para cada uno de estos modelos en el conjunto de prueba.

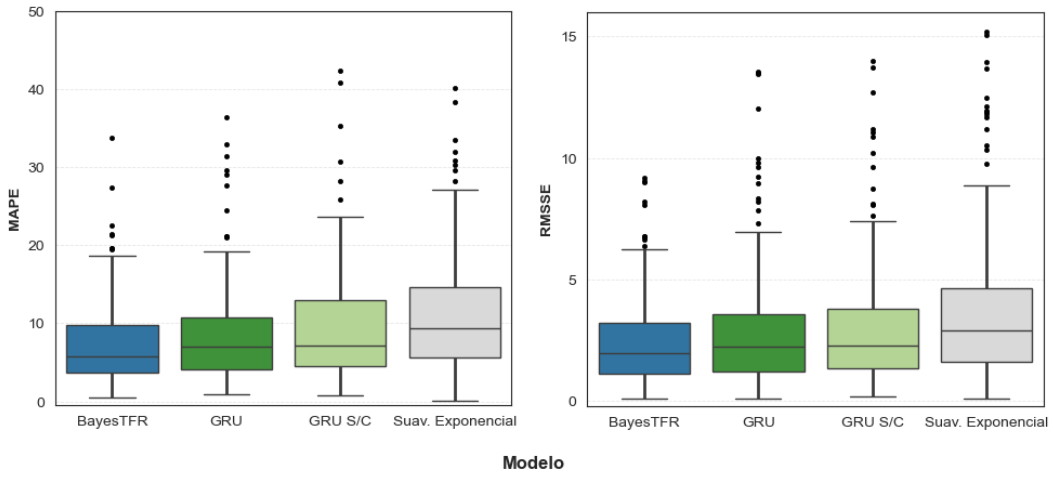

Figura 17: Distribución de los errores de predicción (MAPE y RMSSE) en el conjunto de prueba (2007-2021), en todos los países considerados. Donde cada modelo fue entrenado bajo el mismo conjunto de datos.

Modelo	MAPE	RMSSE
BayesTFR	7.31	2.46
GRU	8.37	2.86
GRU S/C	9.40	3.10
Suav. Exponencial	11.04	3.73

Cuadro 2: Errores medios porcentuales absolutos (MAPE) y raíz del error cuadrático medio escalado (RMSSE) promediados sobre el conjunto de todos los países para cada modelo considerado.

Para profundizar en el análisis, es útil desagregar el rendimiento predictivo por regiones geográficas. Esta visión permite evaluar si la robustez del modelo se mantiene en contextos demográficos diversos, desde regiones con fecundidad elevada y en plena transición hasta aquellas con niveles bajos y post-transicionales. La Figura 18 presenta la distribución del Error Porcentual Absoluto Medio (MAPE) para los cuatro modelos en cada continente.

El análisis regional confirma los hallazgos observados a nivel global. El modelo propuesto (GRU) mejora el rendimiento del modelo base de Suavizamiento Exponencial en todos los continentes, exhibiendo además un comportamiento más estable entre regiones. El modelo base, en cambio, muestra una mayor variabilidad, con un desempeño comparativamente peor en Europa. Al comparar las dos variantes del modelo neuronal, la inclusión de covariables socioeconómicas (modelo GRU) demuestra nuevamente un valor añadido, ya que su distribución de error se sitúa consistentemente por debajo de la del modelo sin covariables (GRU S/C), reforzando la importancia de incluir estos factores en el modelado.

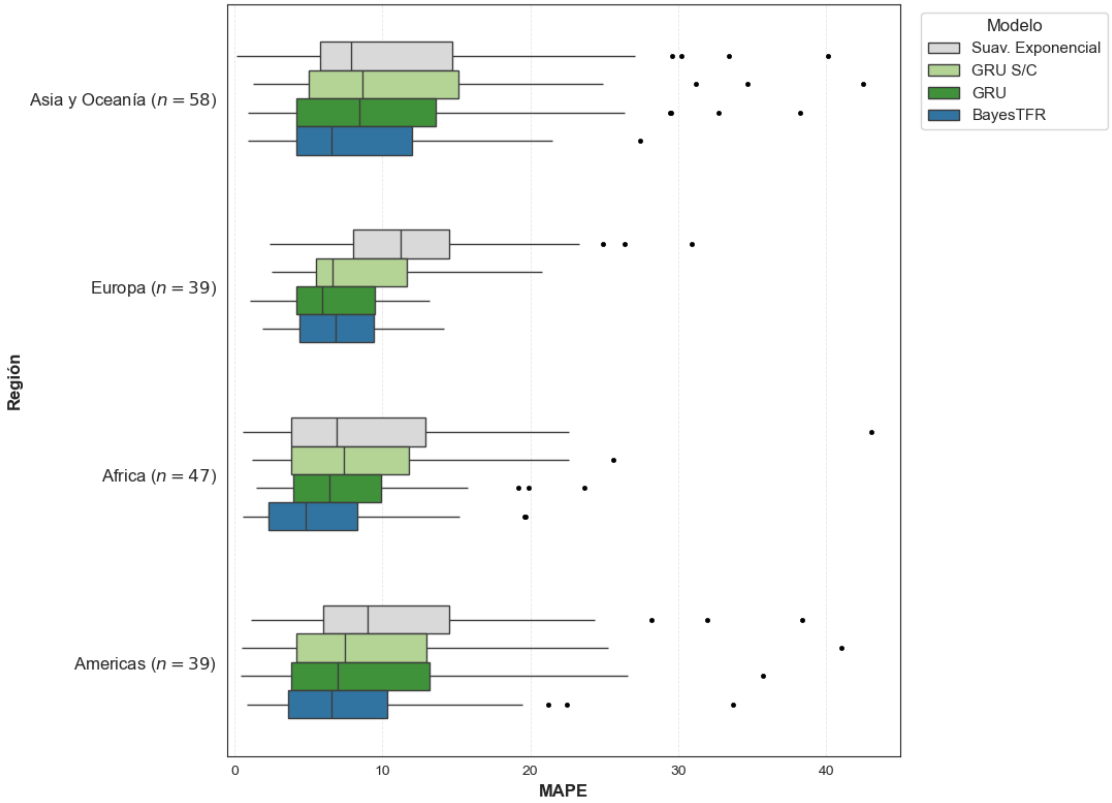

Figura 18: Distribución del error de predicción (MAPE) en el conjunto de prueba (2007-2021), desagregado por continente para los cuatro modelos evaluados.

El punto más relevante surge de la comparación con el modelo de referencia, BayesTFR. Teniendo en cuenta la posible ventaja que los datos de prueba le otorgan a dicho modelo, el modelo GRU con covariables logra acercarse a su precisión en la mayoría de las regiones. El hallazgo más notable se da en Europa, donde el modelo GRU no solo compite, sino que logra la mediana de error más baja de todos los modelos evaluados. Este resultado es particularmente relevante ya que los datos para dichos países suelen ser los de mejor calidad tanto en el caso de las covariables como los propios datos empíricos de la TGF.

4.2. Evaluación Proyecciones del Modelo Final

Una fortaleza conceptual de este enfoque global es su capacidad para modelar trayectorias demográficas muy diversas dentro de una arquitectura unificada, sin necesidad de segmentar el análisis por fases. El modelo aprende de la totalidad de las experiencias históricas para generar pronósticos adaptados al contexto específico de cada país. La Figura 19 ilustra precisamente esta flexibilidad, donde se observan las proyecciones del modelo GRU para cuatro países con realidades demográficas marcadamente distintas. Es notable cómo el modelo captura tanto las dinámicas de países europeos con baja fecundidad como las de naciones africanas en plena transición, demostrando un funcionamiento robusto en ambos escenarios a partir de las relaciones internas extraídas de los datos.

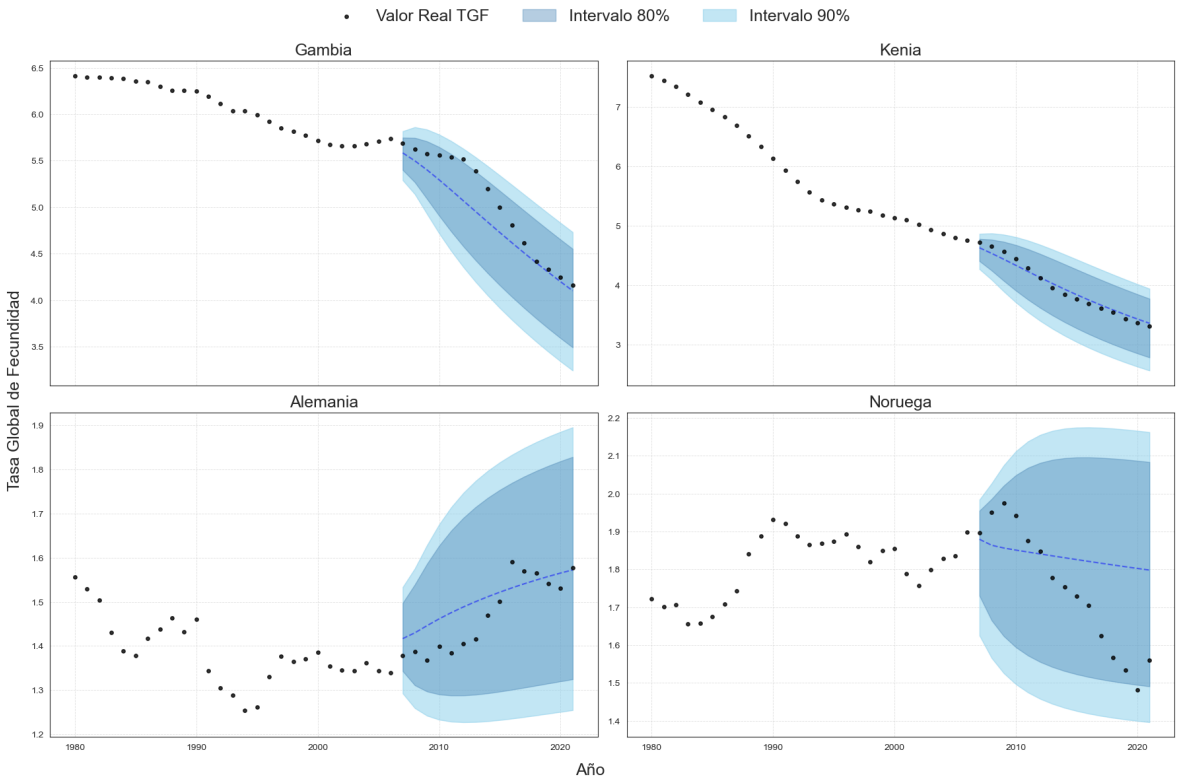

Figura 19: Proyecciones a 15 años en el conjunto de prueba para cuatro países (Alemania, Noruega, Gambia y Kenia) obtenidas mediante el modelo GRU. Donde se observa la flexibilidad del modelo para adaptarse a los distintos contextos demográficos.

Dentro de las distintas dinámicas que presentan las trayectorias de la fecundidad en el mundo, el modelo GRU es capaz de captar muchos de estos patrones en el conjunto de prueba. Por un lado, el modelo logra identificar el cierto repunte que presentaron algunos países de Europa tras haber alcanzado niveles cercanos a un hijo por mujer, como se representa en la Figura 20.

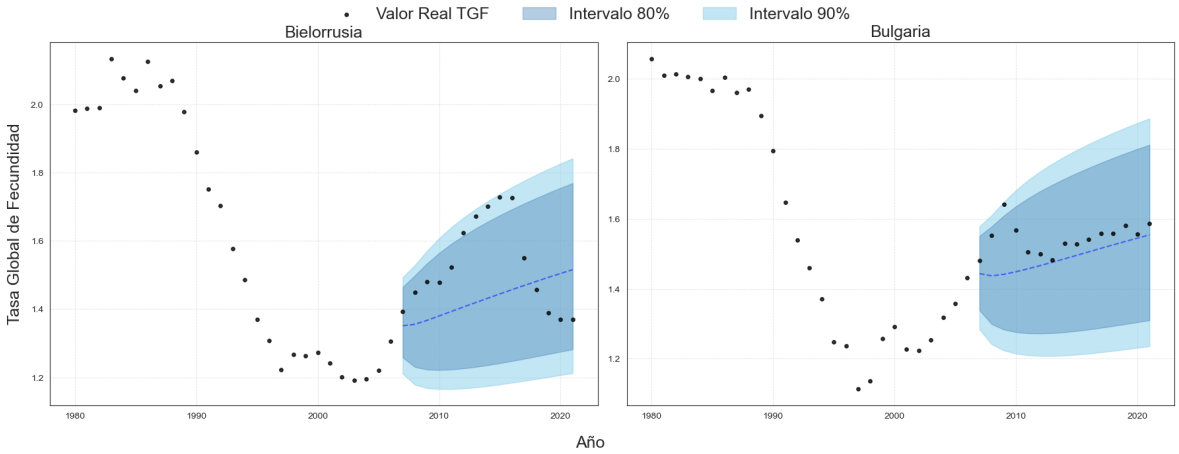

Figura 20: Proyecciones a 15 años en el conjunto de prueba para Bielorrusia y Bulgaria. Estos gráficos ejemplifican la capacidad del modelo para capturar la ligera recuperación de la TGF tras haber alcanzado niveles ultra-bajos que presentaron muchos países en las últimas décadas.

El modelo también demuestra su capacidad para capturar la dinámica de estabilidad que se observa en países una vez que su Tasa Global de Fecundidad (TGF) desciende y se consolida por debajo del nivel de reemplazo. Este comportamiento, en el que se basa la tercera fase del modelo demográfico, es característico de muchos países desarrollados durante la última década. Los casos de Suecia e Irlanda, ejemplificados en la Figura 21, muestran esta estabilidad con un muy leve descenso.

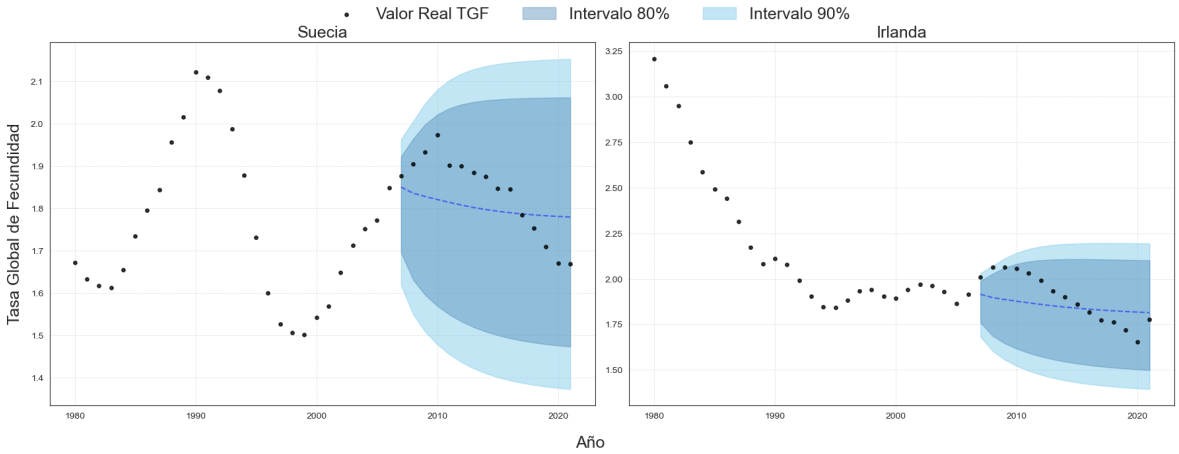

Figura 21: Proyecciones a 15 años en el conjunto de prueba para Suecia e Irlanda. Los gráficos ilustran la capacidad del modelo para capturar la dinámica de estabilidad en países cuya fecundidad ya se encuentra consolidada por debajo del nivel de reemplazo.

En el extremo opuesto del espectro demográfico, el modelo también gestiona con eficacia a los países que se encuentran en plena transición. La acelerada caída de la fecundidad, característica de las naciones del África subsahariana, es correctamente identificada, como se ilustra en la Figura 22 con los casos de Burundi y Burkina Faso.

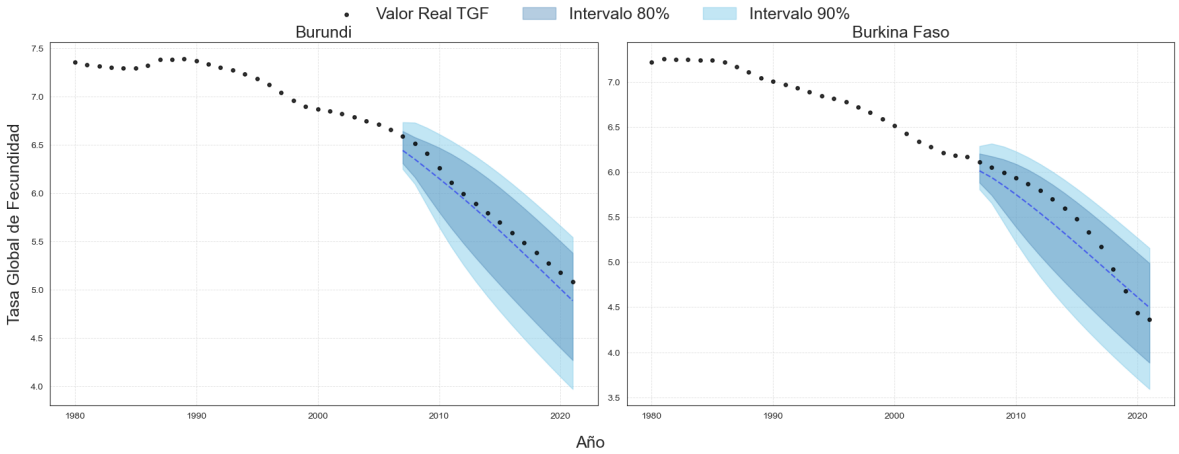

Figura 22: Proyecciones a 15 años en el conjunto de prueba para Burundi y Burkina Faso. Los gráficos muestran cómo el modelo captura la dinámica de descenso acelerado, característica de países en plena transición demográfica.

Finalmente, el comportamiento reciente de países como Argentina y Uruguay remarca la importancia fundamental de un enfoque que pueda cuantificar la incertidumbre en la proyección. Ambas naciones experimentaron una aceleración en su descenso de fecundidad durante la última década, un fenómeno que desafía los modelos puntuales como los utilizados por el Instituto Nacional de Estadística de Uruguay. Como se observa en la Figura 23, aunque la mediana de la proyección resulta más conservadora, la trayectoria real es capturada eficazmente dentro de los intervalos de predicción.

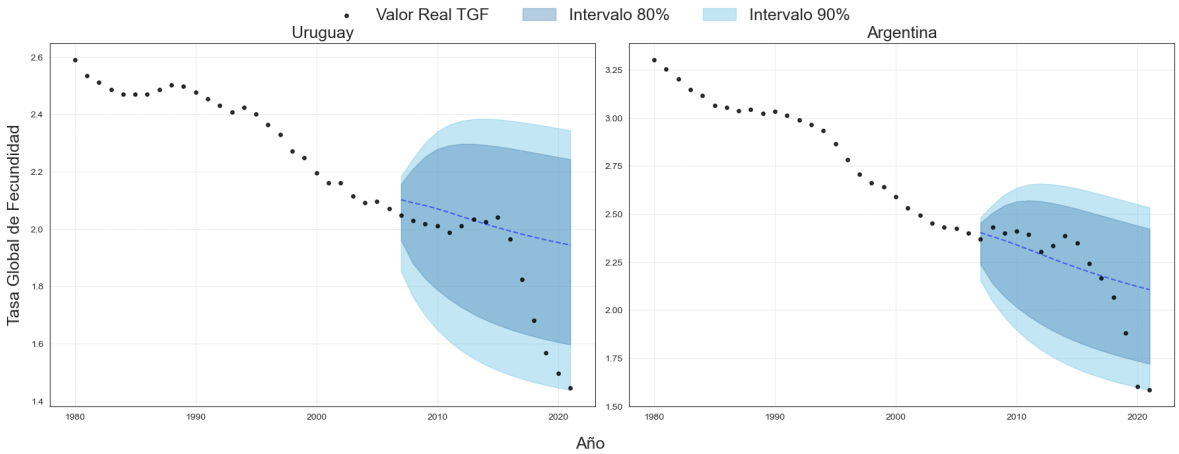

Figura 23: Proyecciones a 15 años en el conjunto de prueba para Uruguay y Argentina. Estos ejemplos destacan la importancia de los intervalos de predicción para capturar descensos abruptos en la TGF, subrayando el valor de un enfoque probabilístico o que contemple la incertidumbre en las proyecciones.

La capacidad del modelo GRU para generar proyecciones plausibles en escenarios demográficos tan dispares, como se observó en la trayectorias anteriores, sugiere una notable flexibilidad. Para verificar si esta cualidad se traduce en un rendimiento consistente a nivel global, se analizó el Error Porcentual Absoluto Medio (MAPE) para cada subregión del mundo, como se ilustra en la Figura 24.

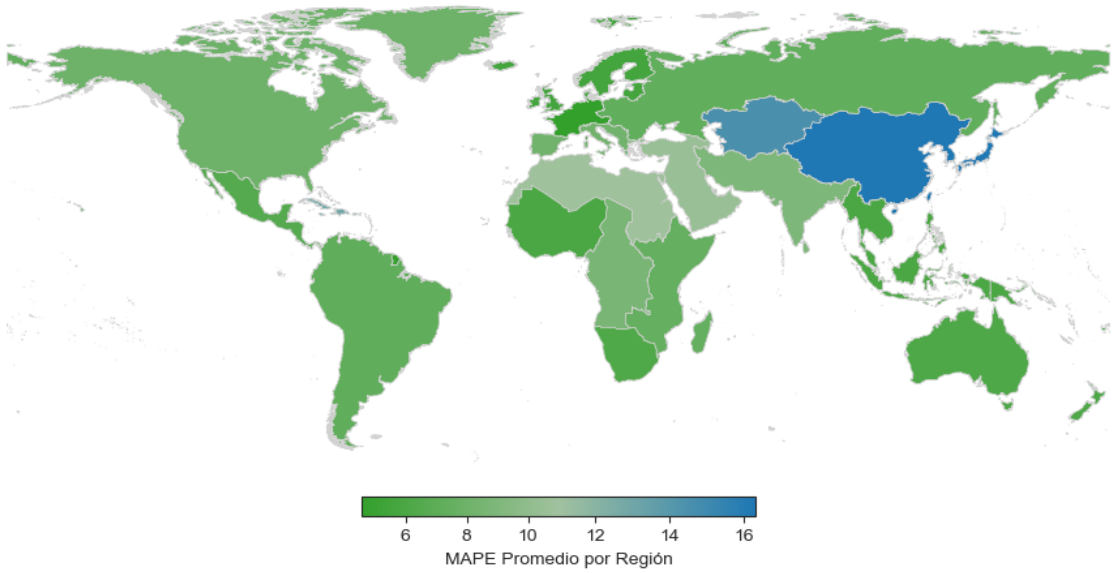

Figura 24: Distribución geográfica del error de predicción (MAPE) del modelo GRU. El color de cada subregión indica el nivel de la media del MAPE para los países que la componen.

El análisis revela que el error de predicción es notablemente uniforme en la mayoría de las regiones, incluyendo las Américas, Europa, África y Asia. Este hallazgo es fundamental, ya que demuestra que el rendimiento del modelo no presenta un sesgo geográfico y que su arquitectura es lo suficientemente robusta para generalizar su aprendizaje a través de múltiples y variados contextos demográficos.

La principal excepción a esta consistencia es el error comparativamente más elevado en la subregión de Asia Central. Dicho error no es una casualidad y está fuertemente vinculado al comportamiento irregular que presentaron varios países de dicha región durante el período de prueba. Países como Kazajistán y Kirguistán experimentaron un aumento sostenido de la fecundidad en esos años, una dinámica que ni el modelo propuesto ni BayesTFR logran capturar con la información disponible. Como se observa en la Figura 25, ambos modelos subestiman esta tendencia creciente, lo que resulta en un desempeño similar.

Esta dificultad para modelar trayectorias atípicas que se desvían de los patrones históricos es un problema presentan lógicamente la mayoría de modelos. Un caso de estudio aún más extremo es el de la República de Corea, cuya fecundidad ha seguido un descenso continuo hasta niveles sin precedentes, situándose por debajo de un hijo por mujer. La Figura 26 ilustra las limitaciones de ambos enfoques en este escenario.

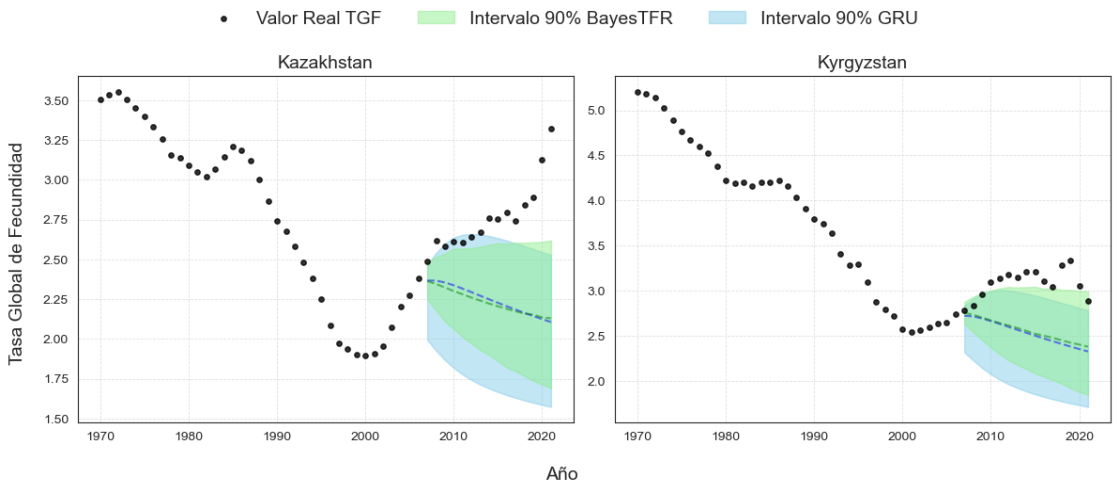

Figura 25: Comparación de las proyecciones de los modelos GRU y BayesTFR para Kazajistán y Kirguistán en el período de prueba. Ambos modelos no son capaces de capturar el comportamiento atípico de la región en dicho periodo.

Al igual que BayesTFR, el modelo GRU no logra anticipar la persistencia de la caída y, en cambio, proyecta un repunte. Notablemente, el modelo propuesto parece ser más reacio a generar predicciones en valores que no ha observado extensivamente durante el entrenamiento (por debajo de 1.0), lo que evidencia una dificultad para la extrapolación en regímenes demográficos completamente nuevos. A su vez marca el hecho que las covariables seleccionadas (educación, planificación familiar, etc.), si bien útiles para describir la transición clásica, no parecen ser suficientes para explicar las fuerzas que impulsan estas dinámicas extremas. En estos casos atípicos, el modelo GRU, a pesar de su capacidad para incluir múltiples variables, sufre de problemas similares a los de un modelo univariado.

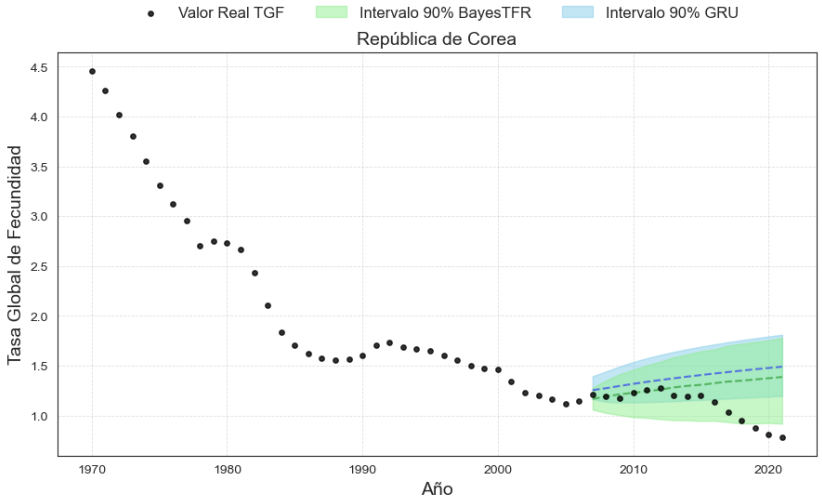

Figura 26: Comparación de las proyecciones para la República de Corea. Ambos modelos fallan en predecir el descenso continuo a niveles ultra-bajos, con un peor rendimiento de GRU en dicho caso.

Si bien las métricas como el MAPE y el RMSSE son fundamentales para evaluar la precisión de la estimación puntual, un pronóstico probabilístico completo requiere además una evaluación de la calidad de la incertidumbre que lo rodea. Con lo cual no basta con que la predicción central sea acertada; como se obtienen en gran parte en este modelo a su vez los intervalos de predicción deben ser fiables e informativos.

En este caso se evalúan dos propiedades clave de los intervalos de predicción: la calibración y la nitidez (*sharpness*). Un modelo ideal debe estar bien calibrado (ser fiable) y, al mismo tiempo, ser lo más nítido posible (ser preciso). Para medir estas dos cualidades, se utilizan las siguientes métricas:

- **Probabilidad de Cobertura del Intervalo de Predicción (PICP)**: Para evaluar la calibración, esta métrica calcula el porcentaje de observaciones reales que efectivamente caen dentro de un intervalo de predicción.

$$\text{PICP} = \frac{1}{h} \sum_{i=t+1}^{t+h} \mathbf{1}_{(L_i \leq y_i \leq U_i)}$$

  donde h es el largo del horizonte de predicción y $\mathbf{1}_{(L_i \leq y_i \leq U_i)}$ es una variable indicadora que toma el valor 1 si la observación real y_i en este caso la TGF se encuentra dentro del intervalo de predicción a evaluar, y 0 en caso contrario.

- **Ancho Medio del Intervalo**: Para evaluar la nitidez, esta métrica calcula la diferencia promedio entre los límites superior e inferior del intervalo.

$$\text{Ancho Medio} = \frac{1}{h} \sum_{i=t+1}^{t+h} (U_i - L_i)$$

  donde h es el horizonte de predicción, y U_i y L_i son los límites superior e inferior del intervalo para cada punto i del horizonte.

En este caso, como se visualiza en la Tabla 3, la cobertura media que presenta el modelo GRU tiene buenos resultados, demostrando una sólida calibración al aproximarse la cobertura empírica a la teórica.

Si se comparan los resultados con los obtenidos mediante BayesTFR, se observa un claro trade-off entre precisión y fiabilidad. Los resultados muestran cómo el modelo GRU presenta intervalos más conservadores, con un ancho medio ligeramente superior (0.89 frente a 0.77 en el nivel del 90 %), pero a cambio obtiene una calibración mejor. Esta superioridad en la calibración es especialmente evidente en el intervalo del 90 %, donde la cobertura del GRU (87.98 %) es mucho más cercana al valor teórico que la de BayesTFR (85.81 %).

Modelo	Métrica	80 %	90 %
GRU	PICP (%)	77.85	87.98
	Ancho Medio	0.64	0.89
BayesTFR	PICP (%)	77.63	85.81
	Ancho Medio	0.60	0.77

Cuadro 3: Comparación de Calibración (PICP) y Nitidez (Ancho Medio) por Modelo y grado de incertidumbre. Cada valor representa la media de la métrica calculada sobre el conjunto de todos los países.

4.3. Interpretabilidad

Una de las críticas más comunes a los algoritmos de aprendizaje profundo es su naturaleza de "caja negra". Dada su alta complejidad y el gran número de parámetros, a menudo resulta difícil comprender cómo el modelo llega a una determinada predicción. En respuesta a este desafío, ha surgido un campo de investigación en pleno desarrollo conocido como aprendizaje automático interpretable, cuyo objetivo es crear técnicas para hacer estos modelos más transparentes, fiables y auditables.

Este trabajo aprovecha una de estas técnicas, que se deriva directamente de la arquitectura del modelo propuesto: el mecanismo de atención sobre las características de entrada. Como se describió, este módulo aprende un conjunto de pesos que ponderan la importancia de cada covariable histórica antes de que esta sea procesada por las capas recurrentes. Si bien su propósito principal es mejorar el rendimiento predictivo, una ventaja fundamental es que estos pesos de atención pueden ser extraídos y analizados, ofreciendo una ventana al razonamiento interno del modelo.

Dado que el modelo es capaz de gestionar trayectorias de fecundidad marcadamente distintas desde países en plena transición hasta aquellos con baja fecundidad, surge una pregunta clave: ¿las covariables históricas ingresadas tienen la misma relevancia en todos los contextos, o el modelo ha aprendido a discriminar su importancia según la realidad demográfica de cada país?

Para explorar esta pregunta, se analiza la distribución de los pesos de atención en el conjunto de prueba para dos grupos con realidades demográficas opuestas: las naciones de Europa, en un contexto de baja fecundidad, y las de África, en su mayoría en plena transición. La Figura 27 muestra la distribución de la importancia (pesos de atención) de cada covariable para los dos grupos. Los resultados obtenidos guardan gran semejanza con lo posiblemente esperable en la teoría demográfica.

En el caso de los países europeos, el modelo atribuye la mayor importancia a factores que, según la literatura demográfica, rigen las fluctuaciones de la fecundidad en contextos de niveles ya bajos. La Edad Media a la Maternidad y los Años de Escolaridad Femenina se destacan como los predictores predominantes. Esto es coherente con la idea de que, en esta fase, el calendario de la fecundidad (efecto tempo) y las decisiones asociadas a niveles educativos avanzados son los factores más determinantes.

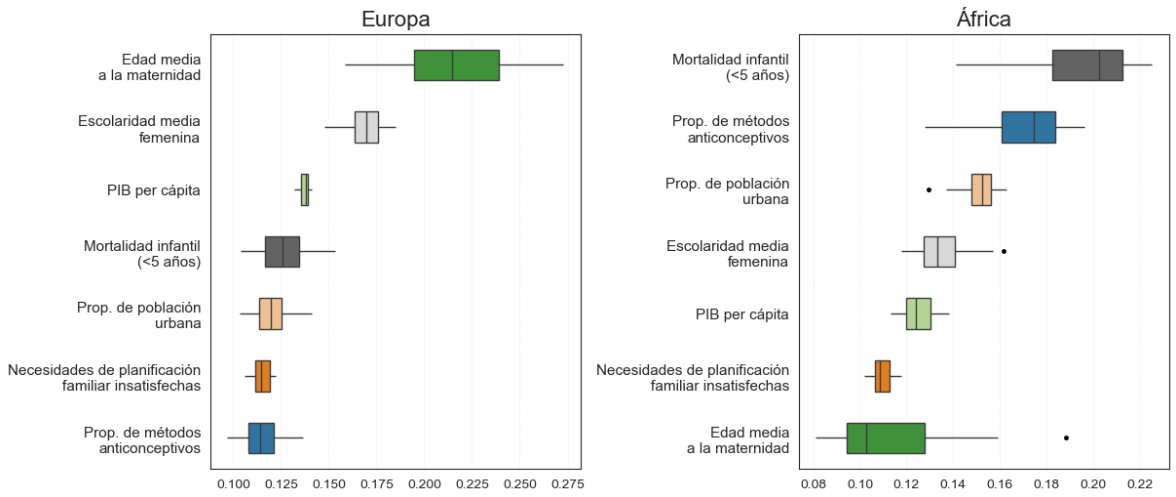

Figura 27: Distribución de los pesos de atención para cada covariable en el conjunto de prueba, desagregado por continentes: Europa y África. Los gráficos muestran cómo el modelo asigna mayor importancia a distintos factores según el contexto demográfico: en Europa, la edad a la maternidad y la educación son clave, mientras que en África lo son la mortalidad infantil y el uso de anticonceptivos.

Por el contrario, para los países africanos el modelo prioriza un pilar de la transición demográfica en países poco desarrollados como lo es la Mortalidad Infantil. A su vez, el uso de Métodos Anticonceptivos Modernos gana una importancia fundamental. Esta ponderación es coherente con la evolución demográfica de la región en las últimas décadas, donde la expansión de la planificación familiar ha sido un factor determinante en el descenso de la fecundidad.

El análisis de importancia de las características, si bien es revelador, no establece relaciones causales. Lo que estos pesos de atención muestran es el poder predictivo que el modelo le asigna a cada covariable para minimizar su error de pronóstico. En otras palabras, se pondera qué tan útil es la trayectoria de una variable para predecir la trayectoria de la TGF en un determinado contexto, lo cual no es necesariamente un vínculo de causa y efecto. Este análisis es valioso para entender la lógica interna del modelo y confirmar que se alinea con la teoría demográfica, más que para extraer inferencias causales directas.

5. Proyecciones Globales de Fecundidad a 2040: Un Análisis Comparativo

El objetivo último de cualquier modelo, más allá de comprobar su ajuste retrospectivo y su viabilidad, es ofrecer proyecciones fiables, en este caso del comportamiento futuro de la población. Esta capacidad resulta especialmente valiosa en el contexto actual ya mencionado, caracterizado por un acelerado cambio demográfico asociado al descenso sostenido de la fecundidad y al envejecimiento poblacional.

Con este fin, en esta sección se analiza qué puede decir el modelo de aprendizaje profundo sobre el comportamiento de la fecundidad en las próximas décadas y en qué difiere de los métodos ya establecidos. Para esto se reentrenan tanto el modelo GRU con Covariables históricas que se destacaba en las pruebas y *BayesTFR* utilizando la información disponible hasta 2021 y se proyectaron hasta 2040.

Al analizar el promedio para el conjunto de los países del mundo, como es de esperarse ambos modelos proyectan una continua caída en la fecundidad, como se observa en la Figura 28 los modelos proyectan a nivel global una tasa promedio aproximada de 2 hijos por mujer para el año 2040, siendo el modelo GRU propuesto levemente menos profundo con una diferencia de aproximadamente 0.05 hijos por mujer para dicho año. La realidad predicha por los mismos muestra un escenario en el cual el promedio de la tasa se encontrará por debajo del nivel de reemplazo para la próxima década. Este escenario que no tiene señales de revertirse en el corto plazo, tiene una implicancia fundamental en el cual el planeta se acerca a un eventual fin de la era de crecimiento demográfico sostenido. Esta nueva realidad demográfica desafía fundamentalmente los modelos económicos y de bienestar social vigentes, que fueron diseñados bajo el paradigma de una pirámide poblacional en expansión. Haciendo que estas proyecciones como las aquí presentadas cobre año a año más valor.

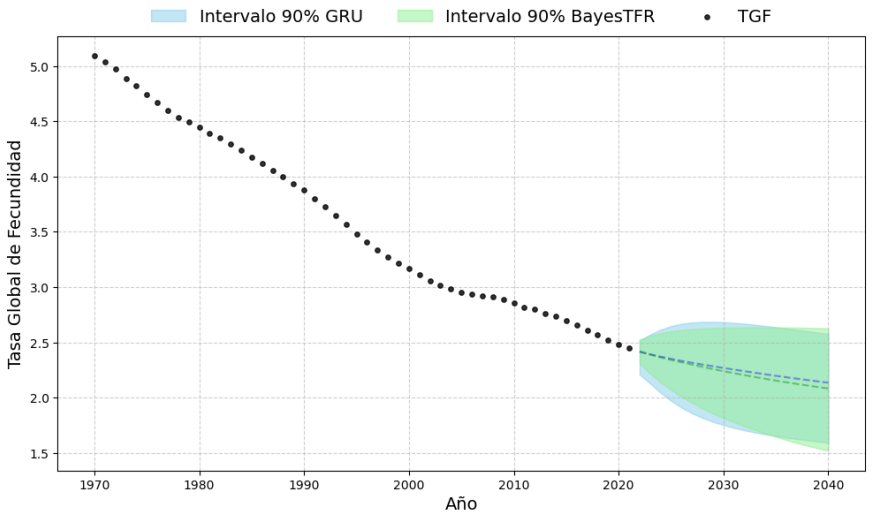

Figura 28: Proyección del promedio simple de la Tasa de Fecundidad por país hasta 2040, mediante los modelos GRU y *BayesTFR*.

Esta caída prácticamente igual a nivel global tiene ciertos comportamientos levemente distintos si se visualiza a nivel desagregado por regiones. Las proyecciones obtenidas mediante el modelo GRU como se muestra en la Figura 29 presenta para países con baja fecundidad un comportamiento levemente superior, presentándose una leve estabilización más optimista que la del modelo bayesiano. Por otro lado los países africanos con altas tasas de fecundidad parecen tener un comportamiento opuesto con una leve tendencia del modelo propuesto a una disminución más importante que la propuesta por BayesTFR sobre todo en la región de África subsahariana.

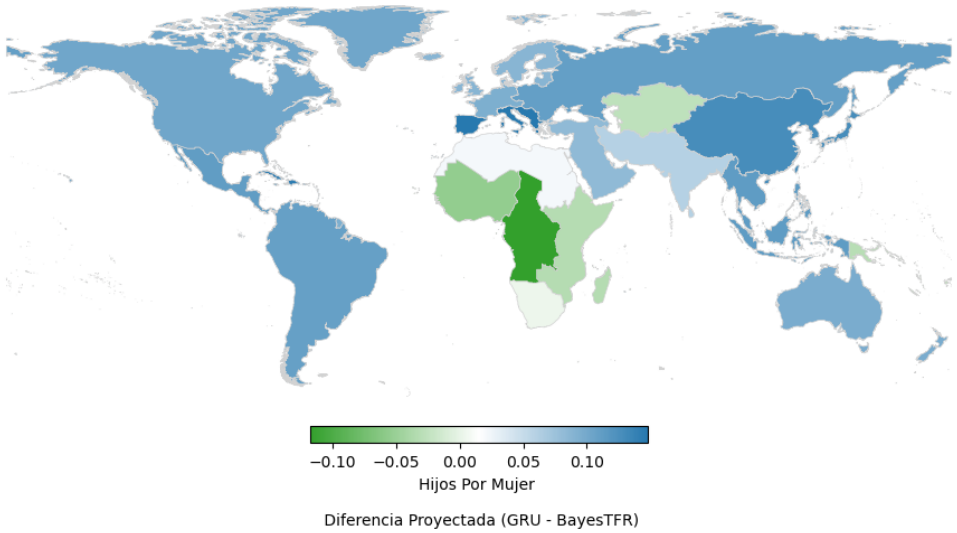

Figura 29: Diferencias en las proyecciones de TGF para el año 2040. El mapa muestra la diferencia (Proyección GRU - Proyección BayesTFR) en el número de hijos por mujer. Los colores azules indican que el modelo GRU proyecta una fecundidad más alta que BayesTFR, mientras que los colores verdes indican que proyecta una fecundidad más baja.

En resumen, el escenario futuro que representan ambos modelos es notablemente consistente, particularmente a nivel agregado. Esta convergencia, lejos de ser una redundancia, puede interpretarse como una de las principales fortalezas y una validación clave del enfoque de aprendizaje profundo aquí propuesto.

Mientras que el modelo bayesiano se fundamenta en un robusto marco teórico y en formas funcionales predefinidas para guiar sus proyecciones, el modelo GRU llega a conclusiones análogas desde un paradigma radicalmente distinto: uno flexible y agnóstico a la teoría, que aprende las dinámicas complejas directamente de los datos históricos y las covariables. El hecho de que dos metodologías tan diferentes una guiada por la teoría explícita y otra por los patrones implícitos en los datos coincidan en la trayectoria futura de la fecundidad, no solo refuerza la plausibilidad del escenario proyectado. Más importante aún, sugiere que el modelo de aprendizaje profundo no es una "caja negra" que genera resultados arbitrarios, sino una herramienta capaz de capturar la señal demográfica subyacente de manera efectiva, consolidándose como una alternativa viable y complementaria en el campo de las proyecciones poblacionales.

6. Conclusiones y Líneas Futuras

El presente trabajo partió de un desafío central en la demografía contemporánea: la necesidad de proyecciones de fecundidad más precisas y flexibles ante la rigidez de los modelos tradicionales. En respuesta a las limitaciones de los enfoques estándar como su dependencia de estructuras predefinidas como el modelo de tres fases y su incapacidad para incorporar covariables socioeconómicas, este trabajo ha demostrado la viabilidad de un paradigma basado en aprendizaje profundo. Los resultados validan que, mediante la aplicación de estrategias como el modelado global para mitigar la escasez de datos, esta metodología se muestra eficaz y se consolida como una herramienta potente y relevante para el campo.

Dentro de las fortalezas de este enfoque, los resultados permitieron confirmar varias ventajas fundamentales. Primero, su rendimiento predictivo es sólido y competitivo, logrando una precisión comparable al estándar de referencia BayesTFR. Segundo, su notable flexibilidad le permite modelar las diversas dinámicas de la fecundidad desde transiciones aceleradas en África hasta estabilizaciones en Europa dentro de un único marco unificado. Esto supera la necesidad de segmentar el análisis en fases, una de las principales rigideces del modelo de Naciones Unidas. Finalmente, el modelo no solo cumple la promesa de integrar determinantes socioeconómicos, sino que el análisis de sus mecanismos de atención revela que aprende a ponderarlos de forma coherente con la teoría demográfica, ofreciendo un valioso grado de interpretabilidad que contrarresta la crítica habitual de ser una "caja negra".

No obstante, el enfoque presenta limitaciones claras que fueron identificadas en el análisis. La principal debilidad reside en la proyección de escenarios sin precedentes históricos, como el descenso a niveles de fecundidad ultra-bajos en la República de Corea o los inesperados repuntes en Asia Central. En estos casos, el modelo puede ser reacio a extrapolar a valores fuera de su rango de entrenamiento y los determinantes clásicos incluidos pierden poder explicativo, haciendo que las covariables no siempre mejoren los resultados. A esto se suma el desafío práctico de obtener un conjunto de covariables de alta calidad y con cobertura global, una condición necesaria para maximizar el potencial del modelo.

Estas conclusiones abren futuras líneas de investigación claras y necesarias para robustecer la metodología.

- *Uso de datos empíricos*: En primer lugar, una extensión valiosa sería evaluar el modelo utilizando datos puramente empíricos de registros vitales, censos y encuestas. Esto permitiría construir un marco de proyección independiente, midiendo su rendimiento sin los sesgos que podrían existir en los datos de referencia actuales de la NNUU, cuya construcción se basa en interpolaciones y ajustes. Este enfoque exige, sin embargo, resolver un desafío práctico: la integración de múltiples fuentes que cubren los mismos períodos con confiabilidad y definiciones heterogéneas. Los registros vitales pueden ser casi completos en ciertos países y fragmentarios en otros; los censos y encuestas difieren en diseño haciendo que la calidad de las mismas difiera entre países.

- *Nuevos modelos y arquitecturas*: En segundo lugar, dado que el aprendizaje automático está en constante evolución, la exploración de nuevas arquitecturas podría ofrecer mejoras en la capacidad predictiva y la interpretabilidad. Si bien las redes recurrentes son efectivas, sería valioso investigar otras familias de algoritmos para contrastar sus fortalezas y debilidades en este problema demográfico.

- *Expansión de covariables*: Por último, para abordar los desafíos observados en escenarios atípicos como los de fecundidad ultra-baja, sería fundamental expandir el conjunto de covariables. Esto implica investigar determinantes no tan clásicos que capturen las nuevas fuerzas sociales y culturales que impulsan estas dinámicas, y evaluar si subconjuntos particulares de variables podrían mejorar el rendimiento predictivo en estas regiones específicas. Adicionalmente, la arquitectura del modelo abre la puerta a un uso analítico más allá del pronóstico, una práctica ya consolidada en otros modelos de proyecciones demográficas. Se podría utilizar el modelo para generar escenarios, proyectando las covariables a futuro bajo distintos supuestos (por ejemplo, un aumento acelerado en la escolaridad o cambios en políticas de planificación familiar) para luego evaluar el impacto simulado de dichas intervenciones sobre la trayectoria de la fecundidad.

En definitiva, este trabajo consolida al aprendizaje profundo no como un mero ejercicio técnico, sino como una alternativa metodológica que responde directamente a las necesidades teóricas y prácticas de la demografía actual. Al demostrar que es posible construir un modelo que aprende las dinámicas complejas directamente de los datos sin renunciar a la interpretabilidad teórica, se ofrece un poderoso complemento que fusiona la flexibilidad de las redes neuronales con el rigor de la ciencia demográfica.

A. Reproducibilidad, Hiperparámetros y Val. Cruzada

Componente/Librería	Versión
Entorno Principal	
Python	3.12.3
Deep Learning	
PyTorch	2.5.1
Manipulación de Datos y Cálculo	
Pandas	2.2.3
Numpy	1.26.4
Scikit-learn	1.6.1
Optimización y Visualización	
Optuna	4.1.0
Matplotlib	3.10.0
Seaborn	0.13.2
Proyecciones Bayesianas	
R	4.4.3
RStudio	2025.05.01.1+513
BayesTFR	7.4.4

Cuadro 4: Componentes de software y librerías principales utilizadas en el proyecto.

Parámetro	Valor
Configuración General	
Modelo Anual	True
AR(1) Fase II	True
Fase II (Transición) - `run.tfr.mcmc`	
Número de Cadenas (MCMC)	3
Número de Iteraciones	65,000
Intervalo de Adelgazamiento (Thin)	1
Fase III (Post-Transición) - `run.tfr3.mcmc`	
Número de Cadenas (MCMC)	3
Número de Iteraciones	50,000
Intervalo de Adelgazamiento (Thin)	10

Cuadro 5: Parámetros de configuración para el modelo BayesTFR.

Hiperparámetro	Valor
Arquitectura y Características	
Arquitectura del Modelo	Enc. GRU + Atención + Dec. GRU
Tamaño Oculto (Encoder/Decoder)	16 / 16
Número de Capas (Encoder/Decoder)	1 / 1
Dimensión del Embedding	6
Longitud de Ventana de Entrada	26 años - 20 años
Transformación del Target	Logarítmica
Escalado del Target/Covariables	Estandar
Cuantiles Evaluados	$[0.05, 0.1, 0.5, 0.90, 0.95]$
Entrenamiento y Regularización	
Tasa de Aprendizaje (lr)	5.5×10^{-4}
Decaimiento de Peso (Weight Decay)	2.4×10^{-4}
Épocas de Entrenamiento	40
Tamaño del Lote (Batch Size)	64
Dropout (Encoder Features)	0.1

Cuadro 6: Hiperparámetros del modelo GRU con Covariables (GRU) *(seed: 876)*.

Hiperparámetro	Valor
Arquitectura y Características	
Arquitectura del Modelo	Enc. GRU + Dec. GRU
Tamaño Oculto (Encoder/Decoder)	8 / 8
Número de Capas (Encoder/Decoder)	1 / 1
Dimensión del Embedding	6
Longitud de Ventana de Entrada	26 años - 20 años
Transformación del Target	Logarítmica
Escalado del Target/Covariables	Estandar
Cuantiles Evaluados	$[0.05, 0.1, 0.5, 0.90, 0.95]$
Entrenamiento y Regularización	
Tasa de Aprendizaje (lr)	5.0×10^{-4}
Decaimiento de Peso (Weight Decay)	2.0×10^{-4}
Épocas de Entrenamiento	35
Tamaño del Lote (Batch Size)	64
Dropout (Encoder Features)	0.05

Cuadro 7: Hiperparámetros del modelo GRU sin Covariables (GRU S/C) *(seed: 8765)*.

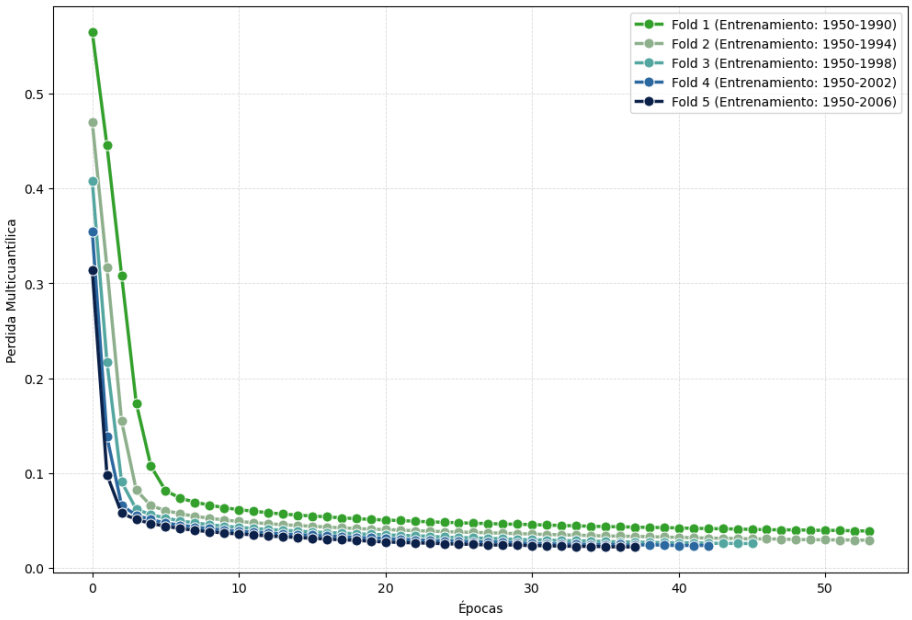

Figura 30: Evolución de la pérdida de entrenamiento (training loss) por época del modelo GRU. Cada línea representa uno de los pliegues de la validación cruzada, mostrando la correcta convergencia del modelo durante el proceso de ajuste a los datos. En cada Fold se realiza *Early Stopping* en base a el error de validación.

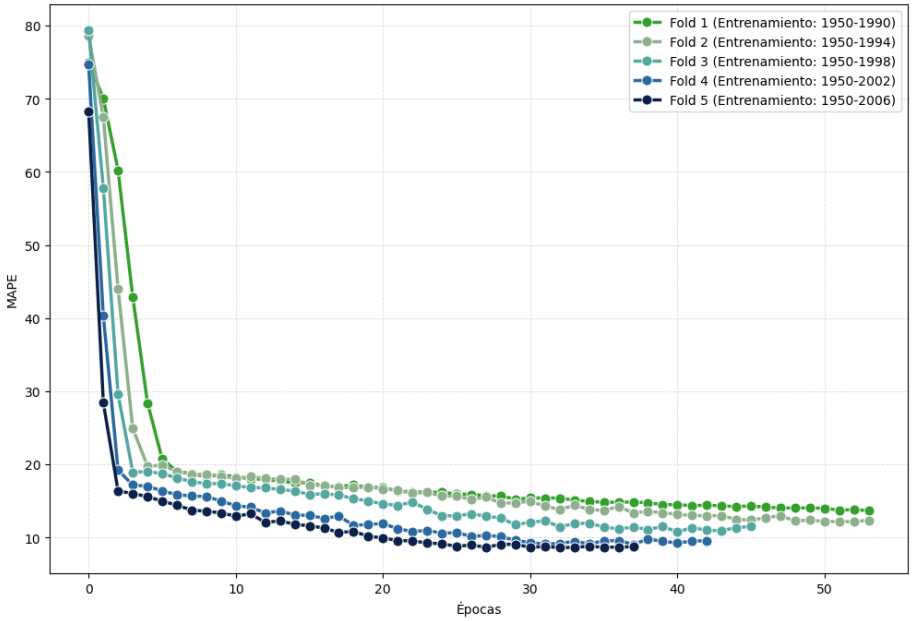

Figura 31: Evolución del error de validación (MAPE) por época para el modelo GRU. Cada línea corresponde a un pliegue de la validación cruzada, ilustrando la capacidad de generalización del modelo en datos no observados durante el entrenamiento. En cada Fold se realiza *Early Stopping* en base a esta métrica.

B. Métricas Por País en el Conjunto de Prueba

		GRU		GRU S/C		Suav. Exponencial		BayesTFR	
id	País	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE
4	Afghanistan	2.86	2.38	1.34	0.38	7.59	6.89	4.81	4.31
8	Albania	8.34	1.57	11.01	1.93	19.28	3.21	6.82	1.21
12	Algeria	19.22	4.51	22.60	5.11	2.68	0.87	19.65	4.62
24	Angola	5.69	5.50	4.91	3.50	4.63	5.03	1.53	1.77
28	Antigua and Barbuda	16.42	2.48	15.47	2.56	7.28	1.27	13.68	2.11
31	Azerbaijan	5.71	1.10	5.99	1.15	15.15	2.79	5.46	1.08
32	Argentina	7.49	3.70	8.93	3.81	9.99	5.13	8.04	3.86
36	Australia	5.26	1.47	6.21	1.73	11.05	3.17	5.48	1.52
40	Austria	4.04	1.08	8.99	1.72	4.22	1.20	2.10	0.60
44	Bahamas	18.63	3.83	19.02	4.21	28.20	5.73	19.48	4.01
48	Bahrain	10.87	2.20	9.01	1.89	33.46	7.28	6.52	1.43
50	Bangladesh	5.04	1.28	3.73	1.07	7.89	2.42	5.08	1.32
51	Armenia	2.55	0.56	3.86	0.89	6.79	1.33	4.89	1.01
52	Barbados	2.35	0.30	2.54	0.48	7.93	0.95	3.62	0.46
56	Belgium	4.66	1.82	6.25	2.37	11.27	4.64	5.25	2.19
64	Bhutan	26.36	4.68	23.67	4.35	13.56	2.40	21.26	3.83
68	Bolivia	2.76	1.37	0.53	0.43	4.86	2.42	5.05	2.51
70	Bosnia and Herzegovina	5.19	0.81	14.22	1.64	26.35	4.57	7.03	1.27
72	Botswana	9.93	3.47	14.20	4.13	9.71	3.77	8.48	2.98
76	Brazil	8.21	1.84	9.94	2.27	11.02	2.83	1.39	0.38
84	Belize	10.96	2.90	6.30	2.06	10.12	3.00	6.95	2.02
90	Solomon Islands	11.15	7.31	13.56	8.07	3.48	2.95	8.01	5.28
96	Brunei Darussalam	2.28	0.46	3.31	0.59	18.17	3.63	4.52	0.97
100	Bulgaria	3.58	1.09	4.32	0.87	15.48	3.75	2.28	0.85
104	Myanmar	1.76	0.68	4.42	1.18	2.92	1.07	1.68	0.67
108	Burundi	2.06	2.22	1.17	1.23	12.21	13.94	5.67	6.36
112	Belarus	9.58	2.61	7.88	2.26	7.82	2.11	9.90	2.75
116	Cambodia	3.68	0.44	5.78	0.61	6.85	0.68	5.76	0.66
120	Cameroon	6.91	6.92	7.40	6.15	3.90	4.01	2.09	2.07
124	Canada	6.47	1.57	7.45	1.71	12.90	2.83	5.47	1.35
132	Cabo Verde	15.81	3.12	14.64	2.86	43.08	8.63	7.24	1.61
144	Sri Lanka	1.15	0.34	1.76	0.35	12.42	3.36	1.96	0.55
152	Chile	12.53	3.86	14.64	4.35	7.91	2.13	9.95	2.95
156	China	10.32	0.45	11.08	0.49	29.58	1.15	9.86	0.46
170	Colombia	14.76	2.91	14.56	2.90	5.72	1.36	10.72	2.18
174	Comoros	7.56	4.52	7.18	4.39	1.65	1.09	5.18	3.02
178	Congo	14.20	13.49	15.89	13.99	2.96	3.10	8.67	8.07
184	Cook Islands	11.74	2.51	5.70	1.60	19.06	4.15	11.74	2.52
188	Costa Rica	10.45	1.93	9.90	2.12	19.19	3.09	6.16	1.29
191	Croatia	4.68	1.20	8.32	1.80	3.20	1.02	3.95	1.13

(sigue en la página siguiente)

(continúa de la página anterior)

		GRU		GRU S/C		Suav. Exponencial		BayesTFR	
id	País	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE
192	Cuba	6.13	0.99	6.46	1.11	31.97	4.62	12.31	1.98
196	Cyprus	19.87	2.26	22.08	2.59	10.30	1.44	14.64	1.73
203	Czechia	7.86	1.90	4.56	1.33	7.17	1.73	9.54	2.19
204	Benin	8.78	8.34	7.37	6.77	6.33	6.60	2.24	2.43
208	Denmark	3.24	1.00	4.07	1.24	9.50	2.85	5.05	1.54
212	Dominica	19.25	2.49	15.69	2.29	38.38	5.00	16.90	2.24
214	Dominican Republic	7.49	1.82	8.77	2.00	15.98	4.00	6.86	1.66
218	Ecuador	8.17	2.60	7.67	2.09	5.05	1.52	8.41	2.71
222	El Salvador	8.89	2.27	5.88	1.91	2.27	0.60	3.85	1.20
226	Equatorial Guinea	4.23	5.08	3.90	3.05	3.51	4.57	0.56	0.64
231	Ethiopia	6.41	4.33	8.99	6.35	3.77	2.60	8.58	5.48
232	Eritrea	3.53	3.38	5.01	4.47	0.56	0.50	1.24	1.21
233	Estonia	3.60	0.99	5.30	1.27	9.33	2.47	3.80	1.04
242	Fiji	1.02	0.38	2.84	0.56	10.59	3.36	4.53	1.41
246	Finland	11.02	3.29	12.26	3.58	16.64	4.74	11.91	3.56
250	France	4.34	1.70	5.98	2.38	9.02	3.34	2.86	1.06
258	French Polynesia	14.52	2.71	11.66	2.63	16.62	3.21	12.23	2.40
262	Djibouti	7.32	2.94	3.16	0.97	10.83	5.11	1.25	0.59
266	Gabon	11.13	7.84	13.67	8.73	2.19	1.53	9.62	6.65
268	Georgia	16.44	6.27	14.36	5.29	14.29	5.44	18.15	6.79
270	Gambia	4.13	3.95	1.85	2.36	16.94	15.06	6.88	6.75
275	Palestine, State of	4.20	2.14	3.87	1.70	6.79	3.70	4.20	2.22
276	Germany	2.42	0.73	6.39	1.12	10.38	3.06	5.76	1.68
288	Ghana	6.48	4.11	7.44	4.93	2.62	1.68	1.67	1.37
296	Kiribati	8.44	4.23	10.92	5.44	14.62	8.36	2.65	1.69
300	Greece	9.35	2.89	14.22	3.91	10.22	3.24	8.68	2.69
308	Grenada	19.56	2.34	15.69	2.13	14.61	1.90	15.31	1.95
312	Guadeloupe	3.10	0.64	3.59	0.61	6.23	1.17	2.74	0.57
316	Guam	13.47	3.01	17.15	3.54	2.77	0.69	9.82	2.27
320	Guatemala	6.68	2.74	3.14	1.77	6.29	2.74	8.31	3.41
324	Guinea	3.46	3.61	2.88	2.53	10.84	12.14	2.71	3.23
328	Guyana	6.65	2.18	9.43	3.05	4.25	1.79	6.54	2.08
332	Haiti	0.65	0.45	4.14	2.11	2.38	1.32	1.29	0.74
340	Honduras	3.51	1.27	0.50	0.31	15.28	6.04	1.05	0.40
348	Hungary	5.14	0.93	10.06	1.24	13.01	2.68	4.55	0.81
352	Iceland	6.48	1.23	7.56	1.36	12.34	2.30	7.18	1.34
356	India	5.96	2.29	4.87	1.90	7.68	3.04	7.65	2.93
360	Indonesia	6.58	2.21	8.02	2.40	7.45	2.92	3.37	1.29
364	Iran, Islamic Republic of	6.00	0.94	5.97	0.80	5.76	0.94	9.10	1.34
368	Iraq	9.34	1.72	10.01	1.75	10.67	2.08	5.09	0.98
372	Ireland	5.26	1.41	6.66	1.68	9.65	2.57	5.06	1.35
376	Israel	17.37	6.69	19.21	7.23	11.02	4.25	10.87	4.26

(sigue en la página siguiente)

(continúa de la página anterior)

		GRU		GRU S/C		Suav. Exponencial		BayesTFR	
id	País	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE
380	Italy	9.86	3.22	16.05	4.23	6.78	2.17	8.59	2.81
384	Côte d'Ivoire	7.80	5.77	6.42	4.63	4.28	4.11	1.83	1.47
388	Jamaica	26.59	3.86	25.26	3.90	11.16	1.64	22.47	3.31
392	Japan	6.21	0.91	14.83	1.56	7.88	0.97	3.59	0.68
398	Kazakhstan	18.89	6.98	21.46	7.42	3.26	1.60	18.63	6.79
400	Jordan	5.31	2.07	7.63	2.58	25.84	8.47	4.25	1.65
404	Kenya	1.69	0.85	1.52	0.79	13.18	6.27	5.44	2.55
408	Korea, Dem. of	1.81	0.11	3.50	0.18	3.76	0.23	1.41	0.09
410	Korea, Rep. of	29.53	2.89	42.49	3.58	17.92	1.94	21.48	2.31
414	Kuwait	11.93	2.35	10.20	2.07	27.05	6.31	10.04	2.07
417	Kyrgyzstan	18.40	6.42	21.87	7.32	3.46	1.28	16.66	5.81
418	Lao	2.58	0.91	1.29	0.40	14.66	5.56	2.71	0.94
422	Lebanon	8.99	3.17	10.19	3.50	15.67	5.62	10.16	3.63
426	Lesotho	6.58	3.20	11.86	4.61	6.16	3.47	6.25	3.01
428	Latvia	5.58	1.44	6.25	1.44	8.00	2.06	6.17	1.62
430	Liberia	1.49	1.27	2.46	1.55	19.67	15.19	8.25	6.24
434	Libya	7.27	1.56	11.13	2.32	15.81	3.44	10.45	2.27
440	Lithuania	7.41	2.04	4.66	1.62	24.92	6.68	9.32	2.58
442	Luxembourg	11.16	3.31	12.78	3.58	11.33	3.28	14.16	4.03
450	Madagascar	5.57	4.49	4.13	3.66	6.08	4.96	2.30	1.88
454	Malawi	5.54	3.33	8.46	5.42	22.65	13.70	15.20	9.17
458	Malaysia	6.02	1.81	6.97	1.98	9.66	2.29	5.96	1.77
470	Malta	12.43	1.99	20.78	2.60	8.07	1.24	9.63	1.56
474	Martinique	3.47	0.68	4.55	0.58	12.98	2.45	3.64	0.70
478	Mauritania	12.02	13.48	11.37	12.69	7.03	8.87	4.77	5.20
480	Mauritius	23.67	1.95	25.59	2.29	3.89	0.35	12.90	1.13
484	Mexico	4.12	1.11	3.73	0.90	1.09	0.33	2.29	0.67
496	Mongolia	29.49	5.75	31.19	5.88	6.65	1.65	27.46	5.40
498	Moldova, Republic of	9.70	2.53	5.95	2.07	10.19	2.67	8.83	2.34
499	Montenegro	5.93	1.43	3.87	1.31	11.49	2.63	4.16	1.10
504	Morocco	4.27	1.10	7.98	1.99	5.93	1.77	4.46	1.14
508	Mozambique	11.98	13.53	12.41	13.71	3.04	4.23	8.17	9.01
512	Oman	11.45	2.76	13.49	3.29	40.11	9.76	17.62	4.19
516	Namibia	13.93	6.28	17.55	7.06	5.81	3.15	11.70	5.31
520	Nauru	22.04	10.00	24.91	11.17	8.99	4.51	17.79	8.21
524	Nepal	9.20	2.67	7.60	2.28	27.07	8.68	3.28	1.02
528	Netherlands	3.96	1.14	5.48	1.49	4.59	1.34	4.47	1.32
531	Curaçao	22.15	3.53	23.00	3.73	14.40	2.68	21.24	3.40
533	Aruba	3.78	0.98	4.64	1.21	9.02	2.40	4.09	0.98
540	New Caledonia	0.89	0.30	2.45	0.41	7.75	2.19	0.93	0.33
548	Vanuatu	12.87	9.81	15.34	10.89	2.13	1.76	6.55	4.85
554	New Zealand	7.23	1.66	8.63	1.93	12.17	2.96	7.39	1.70

(sigue en la página siguiente)

(continúa de la página anterior)

		GRU		GRU S/C		Suav. Exponencial		BayesTFR	
id	País	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE
558	Nicaragua	0.41	0.15	3.67	0.72	1.49	0.44	1.95	0.57
566	Nigeria	5.80	5.43	4.03	3.36	10.74	11.89	3.99	4.78
578	Norway	7.60	2.50	8.92	2.94	18.95	6.20	9.21	3.14
584	Marshall Islands	6.01	2.32	4.23	1.75	8.76	3.17	6.46	2.42
585	Palau	3.10	0.51	3.31	0.44	5.89	1.03	4.59	0.80
586	Pakistan	9.14	5.85	8.58	5.32	7.43	5.25	5.05	3.39
591	Panama	3.86	1.50	7.12	2.24	3.33	1.26	3.69	1.38
598	Papua New Guinea	2.14	1.61	6.34	3.68	0.12	0.10	1.98	1.43
600	Paraguay	2.80	1.11	4.19	1.89	12.26	4.84	3.56	1.43
604	Peru	6.05	1.70	3.70	1.07	6.42	1.87	4.55	1.36
608	Philippines	8.70	3.08	9.42	2.97	30.25	10.34	13.79	5.12
616	Poland	5.69	1.34	12.56	2.13	3.81	0.93	3.78	0.89
620	Portugal	10.61	2.15	17.74	2.97	5.59	1.31	4.38	1.07
624	Guinea-Bissau	2.10	1.69	2.26	1.44	8.88	7.47	3.23	2.98
630	Puerto Rico	35.76	4.94	41.01	5.80	22.69	3.18	33.74	4.69
634	Qatar	14.24	3.34	13.72	3.21	14.57	3.47	10.79	2.58
638	Réunion	2.18	0.44	1.99	0.37	7.26	1.14	1.99	0.41
642	Romania	7.12	0.60	4.39	0.43	23.32	1.85	7.01	0.59
643	Russian Federation	10.32	2.98	7.33	2.27	16.66	4.25	11.49	3.21
646	Rwanda	5.42	3.05	6.68	3.61	7.67	4.25	6.62	3.27
659	Saint Kitts and Nevis	13.91	1.74	12.05	1.74	8.20	1.11	7.56	1.01
660	Anguilla	10.45	1.39	13.65	1.88	24.39	2.78	8.61	1.02
662	Saint Lucia	19.45	2.35	19.31	2.52	12.10	1.59	13.62	1.72
670	S. Vincent and the G.	4.79	0.97	6.04	1.07	8.79	1.72	6.24	1.33
678	Sao Tome and Principe	5.91	4.72	6.02	3.69	11.40	10.54	3.30	3.06
682	Saudi Arabia	3.21	0.97	3.79	0.98	22.64	5.89	3.06	0.91
686	Senegal	9.44	6.15	8.83	5.81	4.81	3.29	5.60	4.29
688	Serbia	1.08	0.58	8.50	2.37	13.85	6.12	4.91	2.05
694	Sierra Leone	3.42	2.61	3.34	3.07	15.14	11.66	3.93	3.12
702	Singapore	38.25	2.45	55.86	3.24	2.68	0.22	20.48	1.42
703	Slovakia	3.44	0.89	5.56	0.83	30.92	7.36	10.33	2.34
704	Viet Nam	7.87	0.75	6.83	0.61	14.82	1.37	9.31	0.88
705	Slovenia	7.74	1.96	2.55	1.10	11.50	2.84	9.67	2.39
710	South Africa	2.04	0.71	7.20	1.14	17.16	5.05	2.45	0.82
716	Zimbabwe	19.90	9.21	21.48	10.21	17.41	8.03	19.58	9.04
724	Spain	13.20	3.08	19.99	3.95	13.45	3.12	11.67	2.77
728	South Sudan	4.78	2.92	4.06	3.10	6.69	4.49	0.65	0.41
729	Sudan	10.94	12.03	11.80	11.06	0.95	1.39	0.73	0.88
740	Suriname	2.41	0.81	4.82	1.53	10.56	3.38	2.13	0.70
748	Eswatini	2.84	1.18	7.44	2.20	1.56	0.65	3.19	1.26
752	Sweden	4.11	1.30	5.64	1.65	12.40	4.07	3.95	1.30
756	Switzerland	2.10	0.87	5.59	1.58	2.41	0.82	1.86	0.70

(sigue en la página siguiente)

(continúa de la página anterior)

		GRU		GRU S/C		Suav. Exponencial		BayesTFR	
id	País	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE	MAPE	RMSSE
760	Syrian Arab Republic	3.62	1.22	6.47	2.17	4.02	1.48	3.63	1.21
762	Tajikistan	13.07	3.70	16.72	4.64	4.11	1.22	14.32	3.97
764	Thailand	13.71	2.11	18.52	2.72	6.51	1.09	10.46	1.74
768	Togo	9.68	8.20	9.40	8.10	12.86	11.95	2.83	2.37
776	Tonga	2.33	1.61	5.10	2.84	5.70	4.01	3.48	2.37
780	Trinidad and Tobago	6.95	1.26	8.05	1.74	1.27	0.28	0.83	0.18
784	United Arab Emirates	32.74	3.88	34.68	4.14	64.95	8.42	14.32	1.77
788	Tunisia	9.99	2.21	12.14	2.47	9.16	2.01	13.47	2.89
792	Türkiye	4.13	1.23	4.93	1.31	6.28	1.98	4.09	1.16
795	Turkmenistan	13.78	4.02	16.98	4.74	3.24	1.08	13.98	4.03
796	Turks and Caicos Islands	15.76	3.68	12.34	3.31	7.72	1.89	9.51	2.34
798	Tuvalu	7.82	4.84	12.60	6.95	1.67	0.95	4.05	2.43
800	Uganda	1.64	1.55	3.65	3.93	13.00	11.80	3.64	3.37
804	Ukraine	10.69	3.01	14.05	3.80	9.70	2.76	10.56	2.88
807	North Macedonia	7.45	1.52	4.54	1.07	15.21	2.87	7.02	1.43
818	Egypt	14.25	5.90	17.02	6.90	6.80	3.05	10.99	4.79
826	United Kingdom	5.45	1.82	6.60	2.14	11.61	4.23	5.27	1.88
834	Tanzania	9.34	9.64	7.57	7.62	3.10	3.16	1.83	1.81
840	United States of America	6.91	1.74	6.45	1.76	22.18	5.30	11.39	2.79
850	Virgin Islands (U.S.)	4.11	0.49	2.38	0.42	14.61	1.71	5.33	0.60
854	Burkina Faso	3.29	2.49	2.06	2.01	13.81	11.18	4.85	4.77
858	Uruguay	9.11	4.65	9.55	4.96	8.57	4.73	7.84	4.16
860	Uzbekistan	11.79	3.57	14.97	3.95	6.01	1.96	13.45	3.84
862	Venezuela of	3.25	1.01	2.53	0.74	4.51	1.41	3.94	1.24
882	Samoa	15.20	8.95	17.07	9.62	4.01	2.56	9.82	5.76
887	Yemen	6.48	3.45	7.13	3.20	6.07	3.25	10.74	5.34
894	Zambia	4.15	3.22	1.33	0.56	14.59	12.47	6.88	5.92

Referencias

Akiba, Takuya et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". En: DOI: 10.48550/arXiv.1907.10902.

Alkema, Leontine et al. (ago. de 2011). "Probabilistic projections of the total fertility rate for all countries." eng. En: *Demography* 48 (3), págs. 815-39. DOI: 10.1007/s13524-011-0040-5.

Bongaarts, John (2020). "Trends in fertility and fertility preferences in sub-Saharan Africa: the roles of education and family planning programs". En: *Genus* 76.1, pág. 32. ISSN: 2035-5556. DOI: 10.1186/s41118-020-00098-z.

Bongaarts, John y John Casterline (2013). "Fertility transition: is sub-Saharan Africa different?" En: *Population and development review* 38.Suppl 1, pág. 153.

Cho, Kyunghyun et al. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. DOI: 10.48550/arXiv.1406.1078. arXiv: 1406.1078 [cs.CL].

Hewamalage, Hansika, Christoph Bergmeir y Kasun Bandara (2022). "Global models for time series forecasting: A Simulation study". En: *Pattern Recognition* 124, pág. 108441. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2021.108441.

Instituto Nacional de Estadística (INE) (2025). *Estimaciones y proyecciones de la población de Uruguay y sus departamentos, 2012–2070*. https://www.gub.uy/instituto-nacional-estadistica/estimaciones_proyecciones.

Koenker, Roger y Kevin F. Hallock (dic. de 2001). "Quantile Regression". En: *Journal of Economic Perspectives* 15.4, págs. 143-156. DOI: 10.1257/jep.15.4.143.

Lim, Bryan et al. (2021). "Temporal fusion transformers for interpretable multi-horizon time series forecasting". En: *International Journal of Forecasting* 37.4, págs. 1748-1764. DOI: 10.48550/arXiv.1912.09363.

Lindholm, M. y L. Palmborg (2022). "Efficient use of data for LSTM mortality forecasting". En: *European Actuarial Journal* 12.2, págs. 749-778. ISSN: 2190-9741. DOI: 10.1007/s13385-022-00307-3.

Liu, Daphne y Adrian Raftery (23 de jul. de 2020). "How Do Education and Family Planning Accelerate Fertility Decline?" En: *Population and Development Review* 46.3, págs. 409-441. DOI: 10.1111/padr.12347.

McDonald, Peter (2000). "Gender equity in theories of fertility transition". En: *Population and development review* 26.3, págs. 427-439.

Miller, John A et al. (2024). "A survey of deep learning and foundation models for time series forecasting". En: *arXiv preprint arXiv:2401.13912*. DOI: 10.48550/arXiv.2401.13912.

Notestein, Frank W (1945). "Population-The long view." En: *Food for the World.*, págs. 36-57.

Oreshkin, Boris N. et al. (2020). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting*. DOI: 10.48550/arXiv.1905.10437. arXiv: 1905.10437 [cs.LG].

Raftery, Adrian E. y Hana Ševčíková (2023). "Probabilistic population forecasting: Short to very long-term". En: *International Journal of Forecasting* 39.1, págs. 73-97. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.09.001.

Salinas, David et al. (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". En: *International journal of forecasting* 36.3, págs. 1181-1191. DOI: 10.48550/arXiv.1704.04110.

Sobotka, Tomáš (2017). *Post-transitional fertility: childbearing postponement and the shift to low and unstable fertility levels*. Inf. téc. Vienna Institute of Demography Working Papers. DOI: 10.1017/S0021932017000323.

United Nations, Department of Economic and Social Affairs, Population Division (2024). *World Population Prospects 2024: Data Files*. https://population.un.org/dataportal/.

Wang, Jun et al. (feb. de 2024). "Time-series forecasting of mortality rates using transformer". En: *Scandinavian Actuarial Journal* 2024.2, págs. 109-123. ISSN: 0346-1238. DOI: 10.1080/03461238.2023.2218859.

Wittgenstein Centre for Demography and Global Human Capital (WIC) (2025). *Wittgenstein Centre Data Explorer v3 (WCDE v3)*. https://dataexplorer.wittgensteincentre.org/wcde-v3.