Análisis del Script de Generación de Datos Pseudoaleatorios de VIH en Colombia

1. Introducción

El script de Python tiene como finalidad la generación de un conjunto de datos sintéticos que representan la evolución de diversos indicadores epidemiológicos y demográficos asociados al VIH en Colombia. El script utiliza un conjunto base de datos reales para ciertos años y, a partir de estos, extrapola y genera información para un rango de años más amplio, definido por el usuario. Este proceso incorpora tendencias predefinidas, fluctuaciones aleatorias y reglas de consistencia para que los datos generados mantengan una coherencia interna y una plausibilidad epidemiológica. El resultado final incluye visualizaciones de las series temporales y la exportación de los datos a un archivo CSV.

2. Descripción del Código y Funcionamiento

El script se estructura en varias partes principales: definición de datos iniciales, funciones de generación y ajuste, el bucle principal de generación, y finalmente, la visualización y exportación de resultados.

2.1. Librerías Utilizadas

El código hace uso de las siguientes librerías estándar de Python:

- matplotlib.pyplot: Para la creación de gráficos y visualizaciones de los datos.
- random: Para introducir elementos de aleatoriedad en la generación de los valores.
- math: Para funciones matemáticas, como math.exp utilizada en la generación inicial de contagios.
- numpy: Para operaciones numéricas, especialmente en la creación de arrays para la interpolación de las gráficas.
- pandas: Para la manipulación de datos en forma de DataFrame y la exportación a formato CSV.
- scipy.interpolate.make_interp_spline: Para suavizar las líneas en los gráficos mediante interpolación.

2.2. Estructuras de Datos Clave

- **datos_reales (Diccionario):** Almacena los datos históricos disponibles para los años 2017 a 2024. Cada clave es un año y su valor es una lista de métricas correspondientes a ese año. La posición de cada métrica en la lista se corresponde con la definición en la variable columnas. Algunos valores pueden ser None, indicando datos no disponibles.
- **columnas (Lista):** Define los nombres de las 29 variables o indicadores que se

manejan (ej. "Contagios", "%Hombres", "%Sexual", "Inc/100k"). Esta lista es fundamental para interpretar datos_reales y para la estructura de datos_gen.

- **anio_inicio y anio_fin (Enteros):** Definen el rango de años para el cual se generarán los datos. Se solicitan al usuario, con valores por defecto de 1981 y 2030 respectivamente.
- **grupos_complementarios (Diccionario):** Establece relaciones entre variables porcentuales que deben sumar un total específico (usualmente 100%). Por ejemplo, %Mujeres se calcula como 100 - %Hombres. Esto asegura consistencia lógica entre variables relacionadas.
- **tendencias (Diccionario):** Especifica parámetros para la generación de ciertas variables clave. Para cada variable listada, se define una tendencia (cambio anual promedio), un valor min y un valor max. Esto permite simular comportamientos conocidos (ej. aumento de cobertura TAR) y mantener los valores dentro de rangos plausibles.
- **datos_gen (Diccionario):** Es la estructura principal donde se almacenan todos los datos generados. Las claves son los nombres de las columnas, y los valores son diccionarios anidados donde las claves son los años y los valores son los datos generados para esa columna y año.

2.3. Funciones Principales

- obtener_promedio_historico(col):
  Calcula el promedio de los valores existentes en datos_reales para una columna específica. Se utiliza como valor base si no hay datos del año anterior.
- obtener_tendencia_historica(col):
  Intenta calcular la pendiente promedio (tendencia lineal) de los valores en datos_reales para una columna. Aunque está definida, su uso en la generación principal parece ser secundario frente al diccionario tendencias.
- generar_contagios(año):
  Función crucial para generar el número de "Contagios". Su lógica varía según el año:
  - **Años anteriores a datos_reales (antes de 2017):**
    - Si año < 1990: Utiliza un modelo de crecimiento exponencial (50 * math.exp((año - 1981) * 0.2)), simulando la fase inicial de una epidemia.
    - Si 1990 <= año < 2000: Emplea otro modelo de crecimiento exponencial con una tasa menor (500 * math.exp((año - 1990) * 0.15)).
    - Si 2000 <= año < min(datos_reales.keys()): Toma el valor del año anterior, le suma una tendencia creciente (200 + (año - 2000) * 50) y añade una fluctuación aleatoria (±15% del valor anterior).
  - **Años posteriores a datos_reales (después de 2024):**

        - Calcula la media de contagios de los últimos 4 años disponibles en datos_reales.
        - Aplica una variación anual aleatoria (entre -12% y +15%) sobre la media para generar un incremento (o decremento).
        - Este incremento se suma al valor de contagios del año anterior generado (o a la media si es el primer año después de los datos reales).
    - Asegura que el número de contagios sea como mínimo 100.
- ajustar_complementarias(principal, complementarias, año):
  Asegura la consistencia entre las variables definidas en grupos_complementarios.
    - Si hay una única variable complementaria (ej. %Hombres y %Mujeres), calcula una a partir de la otra para que sumen 100%.
    - Si hay múltiples variables complementarias (ej. %Hetero, %Homo, %Bi), y ya se han generado valores para ellas, las ajusta proporcionalmente para que su suma alcance el objetivo (generalmente 100% menos el valor principal, o normaliza su suma si se espera que juntas representen un todo).

2.4. Lógica de Generación de Datos

1. **Inicialización de Contagios:** Se copian los datos de "Contagios" de datos_reales a datos_gen.
2. **Generación de Contagios (Extrapolación):** Para los años dentro del rango anio_inicio a anio_fin que no están en datos_reales, se llama a generar_contagios(año).
3. **Generación del Resto de Variables:** Se itera por cada año y cada columna (excluyendo "Contagios"):
    - **Variables con Tendencia Definida (en tendencias):**
        - Se toma el valor del año anterior como base. Si no existe, se usa el promedio histórico de datos_reales o el punto medio entre min y max de la configuración de tendencias.
        - Se calcula un cambio anual sumando una tendencia_anual (basada en config["tendencia"] con aleatoriedad) y un ruido aleatorio.
        - El nuevo valor se acota entre config["min"] y config["max"].
    - **Variables sin Tendencia Definida (y no directamente complementarias):**
        - Se toma el valor del año anterior como base y se añade un "ruido" gaussiano (desviación estándar del 2% del valor anterior).
        - Si no hay valor previo, se usa el promedio histórico de datos_reales más un ruido similar.
        - Si es un porcentaje, el valor se acota entre 0 y 100.
4. **Ajuste Final de Variables Complementarias:** Se itera nuevamente por los años y se llama a ajustar_complementarias para asegurar la consistencia interna de los

porcentajes.

2.5. Visualización y Exportación

- **Gráficos:** Para cada columna, se genera un gráfico de línea mostrando su evolución. Se diferencian los datos reales de los generados y se intenta suavizar la curva. Se incluye una línea de promedio.
- **Resumen en Consola:** Se imprime una tabla con valores para años clave y un subconjunto de indicadores.
- **Exportación a CSV:** Todos los datos de datos_gen se convierten a un DataFrame de pandas y se guardan en datos_vih.csv.

3. Similitud con la Realidad y Validación Epidemiológica

3.1. Fortalezas en la Aproximación a la Realidad

- **Base en Datos Reales:** El uso de datos_reales (2017-2024) como punto de partida y anclaje es una fortaleza significativa, ya que las extrapolaciones y generaciones parten de cifras observadas.
- **Modelado de Tendencias Observadas:** La inclusión del diccionario tendencias permite simular comportamientos conocidos y esperados para ciertos indicadores (ej. incremento en la cobertura de Terapia Antirretroviral - TAR, disminución de la letalidad). Los límites min y max ayudan a mantener los valores dentro de rangos plausibles.
- **Simulación de Fases Epidémicas:** La función generar_contagios intenta reflejar cualitativamente las distintas fases de una epidemia: un crecimiento exponencial inicial, seguido de un crecimiento más moderado, y luego una fase de fluctuación o estabilización relativa.
- **Consistencia Interna de los Datos:** La función ajustar_complementarias es crucial para mantener la coherencia lógica entre variables que son interdependientes (ej. los porcentajes de hombres y mujeres deben sumar 100%).
- **Introducción de Variabilidad:** El uso de funciones de random (como random.uniform y random.gauss) introduce una variabilidad estocástica, que es una característica inherente de los datos epidemiológicos reales.

3.2. Limitaciones y Naturaleza Pseudoaleatoria

- **No es un Modelo Predictivo Mecanicista:** Es fundamental entender que este script **no es un modelo epidemiológico predictivo** en el sentido estricto. No simula los mecanismos subyacentes de transmisión del VIH, el impacto detallado de las intervenciones, ni las dinámicas poblacionales complejas. Es un generador de datos que sigue patrones estadísticos y tendencias observadas o predefinidas.

- **Simplificación de Interdependencias Complejas:** Aunque maneja variables complementarias, las interacciones más complejas entre diferentes métricas (ej. cómo un aumento en %TAR podría influir en la "Incidencia/100k" o en la "Tasa de Mortalidad/100k" más allá de sus tendencias individuales) no están modeladas explícitamente. La generación de muchas variables se basa en su propia tendencia o en el valor del año anterior más un ruido.
- **Dependencia de la Calidad y Cobertura de datos_reales:** La fiabilidad de la extrapolación depende críticamente de la calidad, representatividad y completitud de los datos_reales. Si estos datos tienen sesgos o son incompletos, estos problemas podrían propagarse.
- **Fiabilidad de la Extrapolación a Largo Plazo:** Cuanto más se alejan los años generados de los años con datos reales, mayor es la incertidumbre. Las tendencias pueden cambiar debido a factores no contemplados en el script (ej. nuevas intervenciones, cambios socioeconómicos, emergencias sanitarias).
- **Naturaleza de la Aleatoriedad:** Aunque se introduce aleatoriedad, la forma específica del ruido (gaussiano, uniforme) es una simplificación y podría no capturar completamente la naturaleza de las fluctuaciones o "shocks" que pueden ocurrir en un sistema real (ej. el impacto de la pandemia de COVID-19 en el subregistro o diagnóstico de VIH en 2020-2021, que se refleja parcialmente en la caída de contagios en datos_reales para 2020 y 2021).
- **Valores Constantes o Poco Variables:** Algunas métricas en datos_reales (especialmente hacia el final de la lista de cada año) presentan valores constantes o con muy poca variación. Si estas no tienen una tendencia específica definida en el script, su generación se basará en promedios históricos o en el valor previo más ruido, lo que podría no reflejar cambios futuros esperados.

4. Salida del Script

El script produce tres tipos principales de salida:

1. **Gráficos en Pantalla:** Se muestran múltiples ventanas de matplotlib, cada una presentando la serie temporal de una variable. Esto permite una inspección visual inmediata de los datos generados y su comparación con los datos reales.
2. **Tabla Resumen en Consola:** Ofrece una vista rápida y tabular de un subconjunto de indicadores clave para años seleccionados, facilitando una revisión numérica preliminar.
3. **Archivo datos_vih.csv:** Este es el producto final más tangible y reutilizable. Contiene la totalidad de los datos generados (para todas las variables y todos los años del rango especificado) en un formato estructurado y fácil de importar en otras herramientas de análisis estadístico, hojas de cálculo o bases de datos.

5. Conclusión

El script constituye una herramienta para la generación de datos sintéticos sobre el VIH en Colombia. Logra un equilibrio entre la fidelidad a los datos históricos disponibles y la flexibilidad para generar series temporales extendidas mediante la aplicación de tendencias y variabilidad controlada. Las fortalezas del script incluyen su anclaje en datos reales, el modelado de tendencias específicas, la simulación de fases epidémicas para los contagios y la garantía de consistencia interna entre variables relacionadas.

No obstante, es crucial reconocer sus limitaciones inherentes: se trata de un generador de datos pseudoaleatorios y no de un modelo epidemiológico mecanicista capaz de realizar predicciones precisas o de simular interacciones causales complejas. La validación por una epidemióloga es un paso fundamental que aporta credibilidad a la plausibilidad de los datos generados, pero siempre deben utilizarse teniendo en cuenta las simplificaciones y supuestos del modelo. El archivo datos_vih.csv resultante es un recurso útil para diversos propósitos, como pruebas de software, análisis exploratorios o fines educativos, siempre que se interprete en el contexto de su metodología de generación.